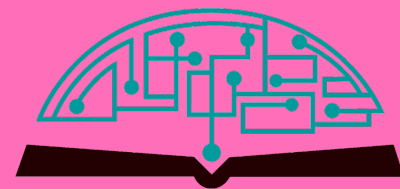


IJHSR

International
Journal of
High School
Research



September 2026 | Volume 8 | Issue 18

ijhsr.terrajournals.org

ISSN (Print) 2642-1046

ISSN (Online) 2642-1054



Let's build a better future together

International Project Fair focused on Sustainability and Environment for Grades 8-12



- Since 2011
- Hosted around 1400 participants in 2025 from 35+ states 70+ countries
- **Disciplines:** STEM, Coding, Robotics, AI, Speech, Entrepreneurship, Arts, Short Film, Music
- Applications start on December 1
- Application Deadline is March 1
- Finalists are announced by March 25
- Event is usually scheduled 2nd week of June
- Monday – Friday, includes a trip to Niagara Falls
- Hosted by large universities at Upstate New York
- Application Fee is \$60/ project
- Participation Fee is \$600/ person, w/ room and board
- Open buffet breakfast, lunch, and dinner
- Trip to Niagara Falls and boat tour is included
- **Instagram and Facebook @Geniusolympiad**
- For more information: **GENIUSOlympiad.org**
- Email: **info@geniusolympiad.org**

GENIUS Olympiad is organized by Terra Science and Education, a N.Y. based 501.c.3 non-profit organization dedicated for project-based learning

Table of Contents

September 2026 | Volume 8 | Issue 18

1	Addressing LLM Context Fatigue via Summarization <i>Aditi Razdan</i>
9	Optogenetics in Epilepsy Research: Current Application and Future Directions <i>Alisha V. Xu</i>
15	The Effects of Mild Traumatic Brain Injuries (mTBIs) on Oculomotor Functions and Effective Treatments <i>André Kurahashi Dantas</i>
22	AI-Assisted Writing, Creativity, and Professional Identity <i>Andrew R. Sion</i>
31	Meaning and Boundaries: How Grammarly and QuillBot Preserve or Interpret Ambiguity in Native English <i>Anya S. Yarlaqadda</i>
37	Evaluating the Effectiveness of US-imposed Sanctions on Venezuela: Fiscal Constraint and Welfare Loss <i>Ariyaana Doss</i>
46	Extrinsic Bias Patterns May Suggest Underlying Intrinsic Bias: A Cross-Domain Analysis <i>Aryan Potukuchi</i>
55	Dual Targeting of EGFR and DRD2 by Plant Agents: As a Potential Anti-Cancer Effect <i>Audrey Hong</i>
61	How High-Skill Immigration Restrictions Eroded Regional Productivity: Evidence from the 2017 BAHA Policy <i>Caroline Y. Su</i>
68	Perceptual Gaps: ASCII Art and Overlapping Audio as CAPTCHA <i>Chong Choon-Hou Rafael</i>
77	Impacts of Nanoplastics on Molecular Mechanisms within Microglia <i>Ella Wilce</i>
85	Minimalism in Korean Aesthetics as a Historically Formed Ideology <i>Hiisan Kim</i>
91	The Quantum Tunneling Effect: From Fundamental Theory to Key Applications <i>Kaito Asakawa</i>
99	Analyzing the Effect of Japanese Barberry on Lyme Disease Using Geographical and Epidemiological Analysis <i>Kavya M. Ballal</i>
105	Hierarchical Communication Structures for Collective Decision Making in Robot Swarms <i>Kevin Wu</i>
113	Exploring the Growth and Limitations of <i>Ex Utero</i> Intrapartum Treatment <i>Kyle N. Alchaar</i>
118	A Stochastic Model of Sabre Fencing <i>Lynn Uemoto</i>
127	Comparing Visual and Working Memory of Teenage and Elderly Populations <i>Prathiksha Raguram</i>
133	Microplastics and Female Reproduction: Comprehensive Review of Hormonal and Fertility Alterations <i>Precious Ojo</i>
140	Exploring Venture Capital Investment Impact on Innovation <i>Safina Kajani</i>
150	Animacy Decoding During Early Visual Processing Using Large-Scale EEG Data <i>Stewart S. Lee, Ioannis Delis</i>
157	U.S. Federal Reserve Monetary Policy: Systematic or Discretionary Based on Real-Time Evidence <i>Zachary Wong</i>

Addressing LLM Context Fatigue via Summarization

Aditi Razdan

Singapore American School, 10 Cuscaden Walk, Singapore, 249693, Singapore; aditirazdan1706@gmail.com

ABSTRACT: This paper addresses context fatigue in large language models, where, after a long conversation, the model begins producing low-quality results that do not follow logically from the prior conversation. To address this issue, we created an in-session summarization tool that condenses prior dialogue into a compact context representation, helping the model recover from context fatigue. Our technique is evaluated through controlled, quantitative experiments that measure retention of injected facts and constraints under long conversational drift. Empirically, our experiment demonstrates that the summarization technique substantially mitigates context fatigue, increasing mean recall by nearly two facts on a five-point scale and restoring retention at 250 turns from near-zero performance (0.0–0.4) to approximately four out of five facts (4.0–4.2). Additionally, this Multi-LLM framework allows users to transfer context and conversations from one model instance to another or across different model families, summarizing conversations between Gemini variants or transferring them to locally hosted models.

KEYWORDS: Computer Science, Artificial Intelligence, Large Language Models, Context Management.

■ Introduction

Imagine working with a large language model (LLM) over the course of an afternoon: refining a research question, debugging code, and iterating on explanations until they are just right. At first, the model behaves like a capable conversational partner, tracking your criteria and prior decisions. Over time, however, earlier details begin to fade, and responses grow less consistent. This degradation often manifests as failures to respect earlier constraints or recall facts that were previously introduced and relied upon.

Today's large language models (LLMs), such as Gemini and ChatGPT, are extremely powerful and have increasingly become conversational partners that assist humans across a wide range of tasks. These tasks include deep research and synthesis (e.g., GPT-based deep research tools), exploring and constructing mathematical arguments (work by Georgiev *et al.* on AI-assisted mathematical discovery), software development and debugging (e.g., Claude and other coding-focused LLMs), providing tax or financial guidance through commercial platforms (e.g., GPT-powered assistance integrated into TurboTax), and even offering forms of emotional or social companionship.^{1–5} In many of these settings, users rely on the model to retain key information, such as goals and constraints, over long conversations.

As tasks become increasingly complex, users naturally interact with LLMs for longer, more involved conversations, relying on the model to remember more key details. A key pain point emerges in these extended interactions, though. LLMs are limited in the amount of information they can handle at once, and so when this limit is exceeded, models may behave unpredictably. As conversations progress, LLMs may forget earlier parts of the session, repeat information, or hallucinate facts, producing responses that are incoherent or misaligned with the user's intent. Prior research has referred to this phenomenon as context fatigue or context drift.^{6,7} Importantly, this failure does not

occur abruptly; it worsens as the interaction continues. Notably, such degradation can occur even when the total conversation length remains within a model's nominal context window.

Unlike systems with explicit recovery mechanisms, an LLM experiencing context drift degrades progressively and cannot reliably recover on its own.^{6,7} Once important context is lost, outputs become increasingly inconsistent with the user's earlier intent.

At present, users typically address context drift by starting a new session with a fresh LLM instance. While starting a new session resets the model's context, it requires the user to repeat themselves and summarize earlier parts of the conversation (similar to movies like *Memento* or *50 First Dates*, in which the character must start from scratch every day). This workaround shifts the burden of recovery entirely onto the user and disrupts conversational flow.

Motivated by this gap, *this paper asks whether structured summarization can be used to recover from this kind of degradation, and presents a system and evaluation framework that tests this directly.* We developed a method to refresh the context. When the model begins giving lower-quality responses, the user can trigger a refresh. In addition, this refresh mechanism can be invoked programmatically after a prolonged interaction to evaluate whether coherence and constraint retention can be restored without user intervention. A different background summarization module condenses the conversation, keeping only the most critical points. This allows the conversation to continue without requiring a full restart.

Our approach is model-agnostic: it is essentially a memory/context transfer mechanism that allows one to port the memory of one LLM to another instance of the same LLM or to an entirely different LLM (e.g., transfer context from Ollama to Gemini). In our implementation, this transfer occurs through a structured summary that serves as a self-contained repre-

sensation of prior dialogue. This allows users to leverage the strengths of multiple LLMs seamlessly.

We hypothesize that our approach works well because, at its core, context drift arises from the model's limited context window and its self-attention mechanism.⁸ Prior work suggests that as the length of a conversation grows, transformer-based models increasingly de-prioritize information from earlier or middle portions of the context, even when that information remains relevant to the task at hand.⁶ Our evaluation framework operationalizes this phenomenon by injecting facts and constraints early in a conversation and measuring their retention after many subsequent turns. Our summarization approach addresses this by condensing the conversation while retaining key facts, the overall logical flow, the user's goals, and any unresolved questions.

We implemented this method in a framework that allows users to select multiple LLMs with interface controls to refresh the current context or transfer it to another model. Beyond the interactive system, we also introduce a controlled evaluation setup that quantitatively measures context retention before and after summarization. Our experimental evaluation shows that summarization is fast and that refreshed contexts produce responses that recover factual and constraint-level information more reliably than uninterrupted long-context interactions. These results suggest that refreshing the context is an effective mechanism for mitigating context fatigue, improving both usability and continuity in extended interactions.

■ Background

Large Language Models (LLMs) are a type of artificial intelligence trained on large datasets to better understand and generate human-like language. They learn patterns from data to perform tasks like writing, translation, and answering questions. When generating a response, an LLM draws on its internal knowledge from training, developer instructions, information it can access online or in a database, its conversation history, and the user's immediate question.

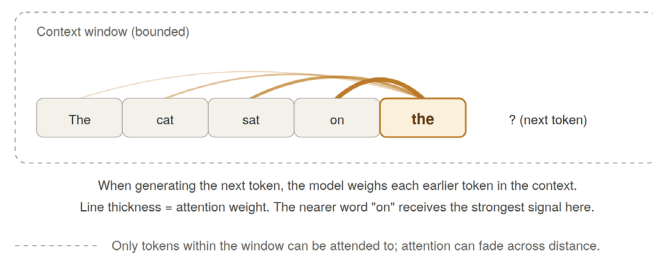


Figure 1: Self-attention lets the model weigh each earlier token in its bounded context window when generating the next one. Nearer tokens receive stronger attention weights, and tokens that fall outside the window cannot be attended to at all, which is the underlying mechanism that produces context fatigue as a conversation grows.

Most modern LLMs are built on the Transformer architecture.⁹ A Transformer processes text one token (a word or part of a word) at a time, and on each step, it uses a mechanism called self-attention to decide which earlier tokens in the conversation are most relevant for predicting the next one. Figure 1 illustrates this relationship: when generating a given word, the model looks back across its input and assigns weights that

reflect how strongly each prior token should influence the output.⁸ Crucially, this lookback is bounded. The model can only attend to tokens that fit within a fixed-size context window, and, as we discuss next, even within that window, its ability to use information effectively can degrade over long inputs.

To focus on context, we define two types of context in LLMs: nominal context and effective context. The nominal context refers to the maximum number of tokens a model can technically process. The effective context, on the other hand, is the portion of that input the model actually uses when generating its response. The cause of this discrepancy is an active area of research. Researchers use terms such as "context fatigue," "context drift," and "Contextual Distraction Vulnerability" (CDV) to describe what happens when a model's conversation history grows beyond what it can effectively handle, often less than the nominal context.¹⁰ This can lead to deterioration in the LLM's performance. This is clearest when the LLM begins to output contradictions, drift from the original topic, and forget earlier details. The model may also lose track of people or entities mentioned earlier in the conversation.^{6,7} These issues have been documented across multiple studies and analyses and are becoming increasingly apparent as users interact with language models over longer and more complex sessions.^{6,7,10,11} Importantly, recent evidence suggests that effective context can degrade even when the total input remains within a model's nominal context limit.

To this day, the precise mechanisms underlying context fatigue remain an open research question. The work of An *et al.* suggests that the primary issue is not memory capacity but rather a lack of training on long-range dependencies.⁷ The authors believe this because, in their experiments, they hide numbers at different positions within long inputs/contexts and find that the model often forgets the ones that appeared earlier in the passage. Testing two smaller models, one trained on longer texts than the other, showed that seeing longer examples helped the model remember better, which led them to conclude that the issue was training rather than memory size.

Conversely, the work of Liu *et al.* identifies a "U-shaped" performance curve. This means that the LLMs are best at recalling information when it's placed at the beginning or end of the context; they perform worst when key details are located in the middle.⁶ The authors argue that when they move the answer to different parts of a long text, the model performs best when the answer is at the start or end and worst when it is in the middle. This consistent U-shaped pattern across tasks and prompts led them to conclude that the model generally struggles with information in the middle of long inputs.

Whatever the root cause may be, the effects of context fatigue are severe and hard to mask: the LLM hallucinates more often and forgets prior conversations. As we show later in this paper, these effects can be observed in failures to retain explicit facts and constraints introduced early and relied upon throughout an extended interaction.

Prior Works:

Developers have attempted several workarounds to prevent context fatigue, and one of the most widely used approaches

is retrieval-augmented generation (RAG).¹² RAG works by storing key information externally, either in a database or on the web, and retrieving it whenever necessary. This enables the model to remain accurate without having to “remember” everything.¹³

Nevertheless, RAG has some limitations. For example, its performance heavily depends on the quality and completeness of its external database; outdated or missing information can lead to inaccuracies.¹²

For instance, a RAG-based assistant retrieving both a valid government report and a misleading blog post can be thrown off by the irrelevant document; a recent study shows that even one “hard distracting” passage reduces answer accuracy significantly compared to giving only the correct passage.¹⁴ In addition, RAG systems are typically designed around explicit document retrieval and are less suited to preserving the evolving goals, constraints, and conversational state that arise organically over a long interactive session.

Another major strategy aims to improve the model's understanding of word positions in long texts. For example, one idea is to modify the model's internal positioning system to leverage its strong understanding of nearby words to handle words that are far apart better.⁷ The goal is to improve the model's use of its full context window without additional data, but this method still requires costly retraining of the entire model and may not fully address the problem of the model's attention weakening over long distances. This is an exciting but still nascent field; hence, no prior work has settled on a final word on how best to address it.

At the application level, commercial LLM products offer partial workarounds to context fatigue, though none target it directly. The most common is simply starting a new conversation, sometimes after the user manually summarizes the prior session, which is a usability workaround that shifts recovery onto the user and disrupts conversational flow. Some systems also maintain lightweight long-term memory; Claude, for example, uses a small preference file (Claude.md) to remember user information across sessions.¹⁵ This form of memory supports personalization across conversations but is not designed to preserve the evolving goals and constraints of an ongoing task-focused session, which is the problem our approach addresses.

■ Problem Statement

In this study, we propose another approach to address context fatigue: a model-agnostic refresh mechanism that compresses the session's key information into a short, structured summary. This approach reduces the need for manual restarts and helps reduce context fatigue. As a benefit, our approach can migrate context from one model to another, allowing a user to use ChatGPT for an initial conversation and then continue seamlessly with Gemini or any other LLM. In our implementation, this migration is achieved through summary-based transfer rather than shared internal state.

The following sections describe our system for refreshing and transferring context and evaluate its impact on performance and accuracy. We also introduce a controlled evaluation

framework that enables testing of this mechanism under a variety of synthetic multi-turn conversational workloads.

■ Methods

Our solution builds on a core architectural idea: using LLMs to generate full-history summaries that act as a compact representation of prior dialogue. This approach allows extended discussions to continue naturally while supporting seamless transitions between models.

The system integrates automated summarization within an intuitive interface that acts as a central control panel for interacting with different LLM providers. By doing so, it helps overcome the limits of traditional context windows and maintains continuity in both within-model and cross-model conversations.

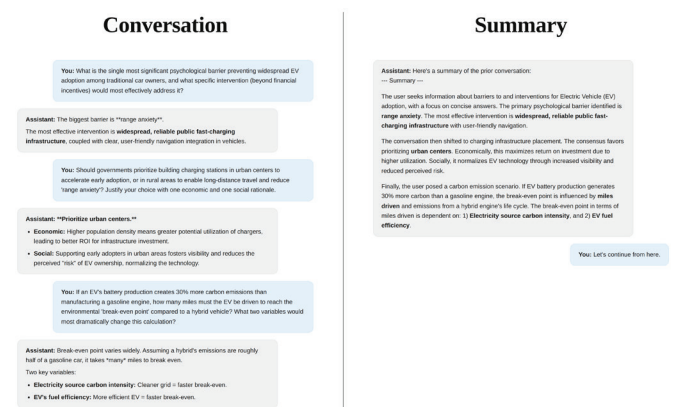


Figure 2: The system condenses a full conversation into a structured summary that preserves key facts and goals. Even in this short example, the summary keeps the core facts, goals, and constraints while dropping conversational filler, and this compression becomes far more pronounced as the underlying conversation lengthens.

Figure 2 illustrates a short exchange and its corresponding summary, showing how the system condenses the main ideas into a concise form that can be reused or transferred across sessions. We used this simple example for illustrative purposes, but we acknowledge that it doesn't fully showcase the strength of our system, since the conversation isn't very long. When the conversation is much longer, the summary's compression is more significant.

At its foundation, the system comprises three core modules: (1) a Context Manager that maintains and tracks both active and summarized conversation states, (2) a Summarization Engine responsible for compressing dialogue into concise and structured digests, and (3) a Multi-LLM Interface that enables model selection and cross-model context transfer by building on the Summarization Engine. Together, these components form an adaptive pipeline that supports continuous dialogue without loss of coherence. While these modules are conceptually distinct, the current implementation emphasizes coordination through message passing rather than persistent shared memory.

The Context Manager serves as the system's persistent memory layer. In practice, it maintains the active conversation state during the running session and synchronizes user preferences, such as the preferred model and feedback style, using a

local SQLite database. One way to think about this database is that it functions similarly to Claude's Claude.md file (previously mentioned in the Prior Works section), except that it works across multiple models. Unlike long-term conversational memory, this database stores only lightweight preference information rather than semantic summaries of dialogue.

When the LLM's responses begin showing signs of context fatigue, the user can trigger the Summarization Engine. Context drift can be identified by the user through qualitative degradation such as repetition, contradictions, or failure to recall recent details. In addition, the system supports summarizing after a prolonged interaction for evaluation, independent of user judgment. The Summarization Engine processes the entire conversation as a single structured prompt. The summarization template explicitly instructs the model to preserve core facts, logical flow, user goals, and unresolved questions. Each summary is generated from the full conversation history rather than from prior summaries, which avoids compounding compression errors across multiple refreshes. This summary is then used as the active context that the model sees on subsequent turns, allowing the conversation to continue without restarting.

Your current active feedback & preferences:

• []

The screenshot shows a chat interface with a light blue header bar containing the text "Your current active feedback & preferences:" followed by a small icon and a list icon. Below this is a chat window with a white background and a light blue border. It contains three messages: a user prompt "You: What molecule is directly responsible for absorbing light energy during photosynthesis?", an assistant response "Assistant: That's a great question! The molecule directly responsible for absorbing light energy during photosynthesis is **chlorophyll**.", and another user prompt "You: Name one specific organelle where photosynthesis occurs.". Below the chat window is a control bar with a text input field "Enter your prompt here:", a "Get Answer" button, and two buttons: "Refresh Context (Same Model)" and "Swap LLM (Toggle Gemini Models)". At the bottom of the control bar is a small blue bar with the text "If you'd like to modify any of your preferences around models or prior feedback, go to settings."

Figure 3: The interface enables controlled context refresh and cross-model transfer, allowing users to recover coherence without restarting the conversation. The image displays a sample conversation between a user and the system. The dedicated "Refresh" and "Transfer" controls let users compress context within a model or move it to another, and visible summary markers show exactly where each context change occurred, keeping users aware of when and how their context is managed.

The interface provides dedicated controls for this process through "Refresh" and "Transfer" functions, allowing users to either compress context within the same model or move summarized conversations to different LLMs with specialized capabilities. This dual-mode structure allows users to preserve conversational continuity while strategically leveraging the strengths of different models. For example, a conversation begun with ChatGPT for creative brainstorming can be transferred to Gemini for concise, factual refinement, without the user having to summarize manually.

Figure 3 shows the cross-LLM chat layout, including the "Refresh" and "Transfer" buttons that manage context compression and model switching.

This layout presents the conversation history with visual distinctions between user prompts, model responses, and transferred contexts. Crucially, it also includes visual markers for summary events, showing where the context was compressed and allowing the user to read the generated summary. The interface prominently features the "Refresh" button for context compression and the "Transfer" button for model switching, with clear visual indicators showing the current active model and a summary of the context state. This design provides users with immediate awareness of context transformations and model transitions.

Before the system summarizes or switches models, it organizes all conversations into a shared message history that stores each exchange as a simple pair of speaker and message, such as `{speaker: "user", message: "Hello"}`. This ensures that every model reads the conversation in a consistent structure. When a user sends a message, the system formats the entire dialogue into a readable transcript that labels each speaker so that different models can follow the same flow of context.

There is no heavy preprocessing or data filtering involved, as the goal is to preserve what was actually said. Once a model responds, the system displays its reply in the chat window with light visual cues that separate user and model turns. When the user chooses to summarize or transfer the chat, that same structured history is passed through the summarization step, and the resulting summary or continuation is added back into the conversation. This design keeps interactions natural while enabling smooth handoffs between models without extra setup or data loss.

The screenshot shows a configuration panel with two columns. The left column has a dropdown menu for "LLM provider" with "Gemini" selected, and a section "Choose which model you'd like to use:" with a dropdown menu for "Gemini" and "Ollama". The right column has a dropdown menu for "LLM provider" with "Gemini" selected, and a dropdown menu for "Gemini" and "Ollama".

Figure 4: The multi-LLM configuration panel supports rapid switching between hosted and local models, demonstrating the portability of summarized conversational state. Because the same structured summary can be loaded into either a hosted Gemini model or a locally run Ollama model, the panel shows that summarized context is portable across model families rather than tied to a single provider.

Figure 4 illustrates how the user can use different LLMs by entering API keys in their terminal and configuring model-specific settings. In the current system, this interface supports both hosted models and locally run models. This centralized configuration enables rapid switching between model providers, in line with the model-agnostic design concept.

The interface is visually divided into the sidebar, the chat window, and the control bar. The sidebar contains preference management options that let users select from multiple models and adjust feedback settings such as tone, format, and verbosity. These preferences are stored locally in SQLite, ensuring user customization persists across sessions without relying on remote servers or external authentication. The chat window displays the conversation as a visually structured dialogue, distinguishing between user prompts, model responses, and summarization events through color-coded formatting. Each summarization event is inserted as a visible summary node that clearly shows when a refresh or transfer has occurred. The control bar includes the "Refresh Context" and "Swap LLM"

buttons, which directly trigger summarization and model-switching events. These controls provide transparency and empower users to decide when and how context management occurs.

Implementation:

The system is implemented using Streamlit, providing a clean, interactive interface for managing multiple LLMs.¹⁶ Users can select which model to interact with, and the interface displays conversation history alongside controls for context management. The current implementation supports both hosted Gemini models and locally run models accessed via an Ollama backend.

Conversation context is stored locally in a structured SQLite database that tracks user settings, such as preferred model and tone, while the active conversation history is maintained in memory during a running session. When the user triggers a context refresh, the Summarization Engine receives the entire conversation as a structured prompt. The summarizer model uses a template to extract core facts, the logical flow, user goals, and unresolved questions.

The system maintains two parallel records of the conversation: a complete history of every message exchanged, and the most recent summary produced by the Summarization Engine. On any given prompt from the user, the model is called with the latest summary followed by all the messages exchanged since that summary was generated. This means that the active context that the model actually sees is bounded.

When the next summarization event is triggered, we use the entire accumulated history (not just the last few messages and summary) to summarize the full conversation.

This design has two characteristics worth noting. First, the summary is limited to 1000 words and focuses only on core facts, logical flow, user goals, and unresolved questions. Summaries, therefore, remain compact even as the underlying conversation grows long, which is what makes them useful as a replacement for the active context. Second, because each summarization passes the full accumulated history to the summarizer in a single call, both the input length and the resulting latency of that inference call grow as the conversation lengthens. This means that on a sufficiently long conversation, the accumulated history could exceed the summarization model's own context window, at which point the Summarization Engine becomes susceptible to the same context fatigue our system is designed to mitigate. Our approach doesn't necessarily eliminate the fundamental constraint of context drift; rather, it extends the duration over which models with bounded context can sustain a coherent conversation. One thing to highlight is that since the summarization model performs a single task (summarizing content), it might be possible that in the future we can have summarization models with very large context windows, since they are trained specifically for this one task rather than being general-purpose models.

Beyond the interactive chat interface, the system includes an evaluation mode that measures context retention under sustained conversational load. Each evaluation run begins by injecting a compact block of five ground-truth facts (a

codename, a numerical budget, a behavioral rule, a forbidden item, and a date) phrased as information the model should remember for later reference. Values are generated from a fixed random seed so that identical configurations produce identical fact sets across runs.

Following this injection, the system replays user turns from one of our synthetic task-oriented datasets while the model under test generates assistant responses live. Because the replayed prompts are unrelated to the injected facts, later recall requires the model to retain the information across many turns rather than rely on repetition. After the conversation reaches a specified length, the model is probed for recall of the injected facts. Depending on the evaluation condition, the system may refresh the context through summarization and, optionally, transfer the conversation to a new model instance before issuing a second probe. This design isolates the effects of interaction length on memory while explicitly testing whether summarization recovers correctness and whether summaries bridge across model instances.

■ Results and Discussion

We evaluate summarization as a mitigation strategy for context fatigue, which we define as the degradation of a model's ability to recall prior constraints and facts over the course of a long conversation. Our central hypothesis is that compressing a prolonged dialogue into a structured summary and re-initializing the model with that summary can recover factual and constraint-level retention while preserving conversational continuity.

We implemented and tested the evaluation pipeline on Gemini 2 (2.5 Flash and 2.0 Flash Lite) and Qwen3-235B, and we report detailed quantitative results for Gemini 2.0 Flash Lite. Gemini was accessed via the hosted API, while Qwen3-235B was executed locally. Evaluating both a cloud-hosted model and a locally deployed model allowed us to measure context drift and refresh across distinct model families under the same protocol. For Gemini, we manually control its context window by supplying only the most recent k tokens (a recency-based truncation), reflecting the bounded sequence processing inherent to transformer-based models.⁸ For Ollama, we can set the context window directly via the `OLLAMA_CONTEXT_LENGTH` environment variable. We set it to approximately 12,000 tokens.

The chosen domains are instruction-heavy and constraint-sensitive, requiring sustained reasoning and careful tracking of prior assumptions. Prompts span definitions, symbolic manipulation, derivations, proofs, applied problem solving, and optimization scenarios, producing extended analytical interaction rather than repetitive exchange and better approximating the structure of real task-focused conversations.

Figure 5 shows the end-to-end evaluation pipeline used to test context retention under sustained conversational load.

Synthetic Dataset Evaluation Runner

Evaluation Mode
☐ Single Run
☒ Batch Grid

Dataset
 Synthetic Calculus (synthetic_calculus_5000.json)

LLM provider
 Ollama

Ollama model name

Batch Evaluation (One-Click Grid)

Datasets to include
 Choose an option

Turn counts (comma-separated)
 50,100,250

Conditions to include
 Choose an option

Max context tokens (approx, 0 = full history)
 0

Output JSON filename
 evaluation_results_1771239763.json
 Incremental JSON filename (optional)
 evaluation_results_1771239763.json

Figure 5: The evaluation framework isolates context drift and quantifies recovery after summarization, enabling controlled measurement of factual retention across conditions. By injecting a fixed set of facts, replaying many unrelated turns, and then probing recall before and after a refresh, the pipeline isolates the effect of conversation length and directly measures how much retention summarization restores.

We evaluate context retention under two experimental conditions. While the system supports summary-based transfer across model instances, this study's quantitative evaluation focuses exclusively on within-model context refresh. In the baseline condition, the conversation proceeds uninterrupted from fact injection through the full sequence of user turns, after which a single probe is issued. Probes consist of five short questions labeled A through E and require responses in a fixed template to minimize parsing ambiguity.

In the refresh condition, the model is first probed after the long conversation, after which the entire interaction is summarized. A new chat session is initialized using this summary as the sole prior context, and the probe is issued again to measure post-refresh retention.

Our primary metric is match recall on the five injected facts. Each probe response is parsed into five labeled answers (A through E), one per injected fact, using the fixed response template described above. Each labeled answer is then compared to its corresponding ground-truth value using a case-insensitive exact string match after whitespace trimming, yielding a per-item score of 0 or 1. The overall recall score for a probe is the sum of these per-item scores, ranging from 0 to 5. We also retain the LLM's responses to identify which types of constraints (e.g., dates, numerical budgets, or stylistic rules) are most vulnerable to drift.

Each cell in Tables 1 and 2 reports the mean and standard deviation of recall across five independent runs of the same configuration, where each run uses a distinct random seed for fact-block generation. We use exact match because the injected facts are short and unambiguous, and because many real-world constraints (dates, numerical limits, prohibited actions) require precise adherence; relaxing the matching criterion would obscure the effects of context fatigue. Higher scores indicate better fact retention.

Table 1: Results of evaluating the baseline and our refresh summarization system on Gemini 2.0 Flash Lite, varying the number of turns in a synthetic calculus long-term conversation dataset we created. Refresh outperforms the baseline at every turn count, and the gap is widest at 250 turns, where baseline recall falls to 0.0 while refresh still retains 4.2 out of 5 facts.

Calculus	Baseline	Refresh
50 Turns	4.2 ± 0.8	4.8 ± 0.4
100 Turns	3.8 ± 1.1	4.4 ± 0.5
250 Turns	0.0	4.2 ± 0.8

Table 2: Results of evaluating the baseline and our refresh summarization system on Gemini 2.0 Flash Lite, varying the number of turns in a synthetic microeconomics conversation dataset we created. The same pattern holds on this harder dataset: baseline recall drops to 0.4 out of 5 at 250 turns while refresh sustains 4.0, though both conditions show slightly lower recall and higher variance than calculus.

Microeconomics	Baseline	Refresh
50 Turns	4.0 ± 0.7	4.6 ± 0.5
100 Turns	3.2 ± 1.3	4.2 ± 0.8
250 Turns	0.4 ± 0.5	4.0 ± 0.9

Several consistent patterns emerge across both datasets. First, the refresh mechanism consistently outperforms the baseline across all turn counts, with the largest gains observed at 250 turns. At this extreme length, baseline recall approaches zero (calculus: 0.0; microeconomics: 0.4 ± 0.5), while refresh maintains robust performance (calculus: 4.2 ± 0.8; microeconomics: 4.0 ± 0.9). This suggests that summarization is most valuable precisely when context fatigue is most severe, namely after very long conversations, when the model has effectively lost access to early information.

Second, the advantage of refresh increases with conversation length. At 50 turns, the improvement is modest (calculus: +0.6 facts; microeconomics: +0.6 facts). At 100 turns, the gap widens (calculus: +0.6 facts; microeconomics: +1.0 facts). By 250 turns, the difference is substantial (calculus: +4.2 facts; microeconomics: +3.6 facts). This trend supports our hypothesis that summarization becomes increasingly beneficial as conversational drift accumulates.

Dataset Differences:

Notably, the microeconomics dataset proved more challenging than calculus across all conditions. Baseline recall at 100 turns was lower for microeconomics (3.2 ± 1.3) than calculus (3.8 ± 1.1), and refresh recall showed higher variance (microeconomics 4.2 ± 0.8 vs. calculus 4.4 ± 0.5 at 100 turns). This may reflect the greater conceptual density of microeconomics conversations, which require tracking interrelated constraints (supply/demand relationships, elasticity concepts) rather than the more linear derivations in calculus. The higher standard deviations in microeconomics (up to ±1.3) also indicate greater run-to-run variability, suggesting that certain types of content are more susceptible to forgetting.

Reliability and Variance:

The standard deviations themselves provide insight into reliability. Baseline conditions show increasing variance as conversations lengthen (calculus: from ±0.8 at 50 turns to effectively 0 at 250 turns; microeconomics: from ±0.7 to ±0.5),

indicating that while most runs fail by 250 turns, the timing of failure varies. Refresh conditions maintain relatively stable variance across turn counts (± 0.4 - 0.9), demonstrating that summarization not only improves average performance but does so consistently.

The recall is intentionally strict. The injected facts are short, unambiguous, and format-controlled, and many real-world constraints, such as dates, numerical limits, or prohibited actions, require precise adherence.

Using looser or fuzzy matching would obscure the effects of context fatigue and weaken causal interpretation.

All evaluations have fixed settings. Fact blocks are generated from a deterministic random seed, ensuring consistent ground-truth values across runs. Conversational inputs are replayed from static datasets in a fixed order, and all evaluation parameters, including provider, model variant, dataset choice, target number of user turns, and evaluation condition, are explicitly specified through the evaluation interface.

Nevertheless, some sources of variation remain. In particular, model sampling behavior and backend differences can affect recall, and decoding parameters are left at provider defaults.

Discussion:

Since we began this study, related ideas have appeared in commercial tools such as Claude Code, which provides two distinct user-facing commands for context management: `/clear`, which wipes the active conversation entirely, and `/compact`, which asks the model to summarize the conversation and then replaces the history with that summary.¹⁷ These features reflect a growing recognition that long-running conversations require explicit context management rather than uninterrupted accumulation. One important distinction remains, however. Our system treats summarized context as a transferable artifact that can be exported across model instances, whereas Claude Code currently operates within a single-model setting and does not support cross-model context transfer.

The quantitative results support the practical value of this approach. At 250 turns, baseline models retain effectively zero facts; they have, for practical purposes, "forgotten" the conversation's starting constraints. Users experiencing this would need to restart or manually recap. Our refresh mechanism restores approximately 80% of facts (4.0-4.2 out of 5), meaning users can continue seamlessly without repetition. This 80% retention rate after extreme conversational length suggests that summarization captures most critical information while discarding irrelevant drift.

Our approach also has clear limitations. Summaries are interpretive compressions of prior dialogue and therefore cannot preserve all information with perfect fidelity. Any detail that is omitted or misrepresented during summarization is effectively lost in subsequent interactions. In addition, our evaluation focuses on short, explicit facts and constraints, which provide a strict and interpretable signal but do not capture more abstract aspects of conversational continuity, such as evolving reasoning strategies or nuanced preferences. The summarization process itself depends on the underlying model's behavior, which introduces additional sources of variability.

The dataset-dependent differences we observed (microeconomics showing lower recall and higher variance than calculus) highlight an important avenue for future work: understanding which conversation types are most susceptible to context fatigue. If certain topics or reasoning styles are particularly vulnerable, summarization strategies might be tuned accordingly; for example, more frequent refreshes for conceptually dense domains.

There are other ways this problem could be addressed. Retrieval-based systems can reintroduce relevant information from external stores, and architectural changes aimed at improving long-context attention may reduce effective context degradation.^{7,12,13} Each of these approaches targets a different aspect of the problem and comes with its own trade-offs. Future work could explore richer automatic summary representations, adaptive mechanisms that trigger refreshes when degradation is detected, and tighter integration between summarization and retrieval. More broadly, treating context as an explicit, portable object rather than an implicit buffer opens new directions for building LLM systems that better support sustained, flexible interaction over time.

Conclusion

In this study, we explored using summarization to address the limited context supported by LLMs. We built a full chat application from scratch to experiment with this idea, created multi-round synthetic datasets on topics in mathematics and economics, and designed an interface that allows us to toggle between different options (LLMs and evaluation conditions). In sum, this effort allowed us to check whether our generic approach worked for both cloud-based and local models. Our results suggest that strategic summarization can help LLMs remember facts introduced earlier in the conversation, even at extreme conversational lengths where baseline retention collapses.

More broadly, this work points to a shift in how we should think about extended interaction with LLMs. As these models take on longer and more consequential tasks (e.g., tutoring students through a semester, supporting writers through successive drafts, assisting researchers across multiple sessions), the limiting factor is clearly continuity. As conversational systems become a default interface to computation and knowledge, the ability to sustain coherent interaction over time will matter at least as much as the ability to generate a good response for any single question or prompt. Treating context as an explicit, portable object rather than an implicit buffer reframes this limitation as something that can be engineered around.

Acknowledgments

The author thanks her mentor, Sebastian Angel, for providing helpful feedback during the writing of this manuscript.

References

1. OpenAI. *Introducing Deep Research*. 2025. <https://openai.com/index/introducing-deep-research> (accessed April 24, 2026).

2. Georgiev, B.; Gómez-Serrano, J.; Tao, T.; Wagner, A. Z. Mathematical Exploration and Discovery at Scale (2025). arXiv:2511.02864. <https://arxiv.org/abs/2511.02864> (accessed April 24, 2026).
3. Anthropic. *Introducing Claude*. 2023. <https://www.anthropic.com/index/introducing-claude> (accessed Feb 14, 2026).
4. Intuit. *Introducing Intuit Assist: The Generative AI-Powered Financial Assistant for Small Businesses and Consumers* (2023). <https://investors.intuit.com/news-events/press-releases/detail/43/introducing-intuit-assist-the-generative-ai-powered-financial-assistant-for-small-businesses-and-consumers> (accessed April 24, 2026).
5. Liu, B.; Sundar, S. S. Designing for Emotional Support with Social Chatbots. *Proc. ACM CHI Conf. Hum. Factors Comput. Syst.* 2018, 1: 1–12. DOI: 10.1145/3173574.3173851.
6. Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *arXiv Preprint*, arXiv:2307.03172 [cs.CL], 2023.
7. An, C.; Zhang, J.; Zhong, M.; Li, L.; Gong, S.; Luo, Y.; Xu, J.; Kong, L. Why Does the Effective Context Length of LLMs Fall Short? *arXiv Preprint*, arXiv:2410.18745 [cs.CL], 2024.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* 2017, 30: 5998–6008.
9. Ferraris, A. F.; Audrito, D.; Di Caro, L.; Poncibò, C. The Architecture of Language: Understanding the Mechanics Behind Large Language Models. *Cambridge Forum AI Law Gov.* 2025, 1(1): e11. DOI: 10.1017/cfl.2024.16.
10. Hong, K.; Troynikov, A.; Huber, J. Context Rot: How Increasing Input Tokens Impacts LLM Performance. Chroma Technical Report, 2025. <https://research.trychroma.com/context-rot> (accessed April 24, 2026).
11. Zhuang, Y.; Yu, X.; Wu, J.; Sun, X.; Wang, Z.; Liu, J.; Su, Y.; Shang, J.; Liu, Z.; Barsoum, E. Self-Taught Agentic Long Context Understanding. *arXiv Preprint*, arXiv:2502.15920 [cs.CL], 2025.
12. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv Preprint*, arXiv:2005.11401 [cs.CL], 2021.
13. Amazon Web Services. *What Is Retrieval-Augmented Generation (RAG)?* 2025. <https://aws.amazon.com/what-is/retrieval-augmented-generation/> (accessed Feb 14, 2026).
14. Amiraz, C.; Cuconasu, F.; Filice, S.; Karnin, Z. The Distracting Effect: Understanding Irrelevant Passages in RAG. *Proc. Annu. Meet. Assoc. Comput. Linguist.* 2025, 63: 18228–18258.
15. Anthropic. *Manage Claude's Memory*. 2025. <https://code.claude.com/docs/en/memory> (accessed Feb 14, 2026).
16. Streamlit. *Streamlit: The Fastest Way to Build and Share Data Apps*. 2024. <https://streamlit.io/> (accessed Feb 14, 2026).
17. Anthropic. *How Claude Code Works* (2026). <https://code.claude.com/docs/en/how-claude-code-works> (accessed April 24, 2026).

■ Authors

Aditi Razdan is a junior at Singapore American School with an interest in artificial intelligence and machine learning systems. She plans to pursue computer science in college, focusing on AI and systems design, particularly at the intersection of technology and social justice.

Optogenetics in Epilepsy Research: Current Application and Future Directions

Alisha V. Xu

Milton Academy, 170 Center St, Milton, MA, 02186, USA; alishavxu@gmail.com

ABSTRACT: Optogenetics represents a transformative approach in neuroscience, enabling light-based control of genetically defined neuronal populations with unparalleled temporal and spatial precision. Within epilepsy research, this technology has opened new avenues for dissecting the mechanisms of seizure generation and developing highly targeted interventions. Epilepsy, a neurological disorder marked by recurrent seizures caused by abnormal neuronal synchronization, remains a major clinical challenge due to the limited specificity of existing treatments. By integrating microbial opsins, viral gene delivery systems, and advanced light delivery tools such as optical fibers and LEDs, researchers can modulate excitatory and inhibitory pathways in real time to restore network balance. Experimental applications across models of temporal lobe, neocortical, and post-stroke epilepsy demonstrate that selective activation or suppression of specific neurons can effectively halt or prevent seizure activity. This review synthesizes current advances in optogenetics and outlines its potential for future closed-loop therapeutic systems capable of detecting and interrupting seizures as they occur. We also discuss existing limitations, including limited light penetration, vector safety, and long-term stability, which currently restrict translation to human use. Despite these challenges, optogenetics remains a powerful platform in understanding and controlling epileptic circuits and a promising foundation for next-generation, personalized treatments for refractory epilepsy.

KEYWORDS: Medical and Health Sciences, Epilepsy, Optogenetics, Channelrhodopsin, Halorhodopsin.

■ Introduction

Epilepsy is one of the most prevalent neurological disorders; it is characterized by recurrent, unprovoked seizures resulting from abnormal synchronization of neuronal activity.¹ Globally, epilepsy affects approximately fifty million individuals and accounts for nearly 2.4 million new cases each year, with a disproportionate burden placed on developing countries where access to diagnosis and treatment may be limited.² These statistics emphasize the urgent need for more effective and accessible therapeutic strategies. Despite advances in pharmacological and surgical treatments, there is still an urgent need for more precise and effective interventions. Conventional antiepileptic drugs act by broadly suppressing neuronal excitability, often causing side effects such as cognitive impairment, fatigue, or mood disturbances; in addition, surgical resection can be risky when seizures originate in brain regions responsible for neurological functions such as language, movement, sensory processing, or memory.^{2,3} These limitations highlight a critical gap in therapeutic specificity and motivate the search for alternative strategies that can target dysfunctional neural circuits with high temporal and spatial precision. Optogenetics, a rapidly advancing technique that integrates genetic and optical methods to control neuronal activity with light, has emerged as a promising tool for both neuroscience research and potential clinical applications.⁴ By introducing light-sensitive opsins, such as *channelrhodopsins* (ChR2) and *halorhodopsins*, into specific neuron populations, optogenetics enables precise activation or inhibition of neural activity with millisecond precision.⁵ This approach offers a distinct advantage

over traditional electrical stimulation and pharmacological interventions, providing reversible and cell-type-specific modulation of brain circuits. In recent years, optogenetics has been successfully applied in experimental models of epilepsy to identify seizure-generating regions, suppress abnormal discharges, and restore network stability.⁶ However, its translation to human therapy remains limited by technical challenges, including restricted light penetration, viral delivery efficiency, and long-term biocompatibility.⁷ This review aims to provide an overview of optogenetic mechanisms, evaluate its advantages over conventional techniques, and examine its experimental applications in epilepsy research, with a focus on its therapeutic potential and current limitations.

■ Optogenetics

History of optogenetics:

Since the discovery of light-sensitive proteins in 1971 by Oesterhelt and Stoerkenius, the field of optogenetics has evolved into a complex technique that uses these proteins, various genetic delivery methods, and light delivery technologies.^{8,9} Specifically, optogenetics is a method that enables control of the activity of genetically modified cells using light. The term “optogenetics” was coined by Karl Deisseroth *et al.* in 2006.⁸ However, its foundational discoveries date back several decades. In 1971, Oesterhelt and Stoerkenius identified bacteriorhodopsin, a light-sensitive proton pump, located in *Halobacterium halobium*, a halophilic bacterium.⁹ This was the first known opsin capable of converting light into cellular activity. In the years that followed, many other microbial opsins were discovered,

including *halorhodopsin* in 1984,¹⁰ and ChR2 in 2002,¹¹ which would later become key components in optogenetic systems. In 2005, researchers successfully expressed ChR2 in mammalian neurons, demonstrating that light could directly trigger electrical activity in excitable cells.⁵ This study opens the door to precise manipulation of neural circuits, thereby advancing neuroscience research.

Various light-sensitive proteins:

The development and application of light-sensitive proteins known as “opsins,” which enable neuronal activation, are a crucial foundation of optogenetics. Among these, ChR2 is a light-gated cation channel originally discovered in green algae.¹² ChR2 can depolarize cells in response to specific wavelengths of light, making it an essential tool for neural excitation.¹³ Moreover, ChR2 is derived from *Chlamydomonas reinhardtii*, which responds to blue light. Initial experiments using ChR2 demonstrated that when expressed in the retinas of mice lacking functional photoreceptors, it could reintroduce light responsiveness by making the remaining neurons sensitive to light stimulation (Figure 1).¹⁴ However, despite its success, ChR2 has several drawbacks, including the need for high-intensity blue light and relatively slow response kinetics, which have driven the development of enhanced opsin variants with improved sensitivity and speed.¹⁵ In contrast to excitatory opsins like ChR2, optogenetics also utilizes inhibitory proteins to suppress neuronal activity. *Halorhodopsin* is a chloride pump that hyperpolarizes neurons by importing chloride ions in response to specific wavelengths of light, which reduces the likelihood of firing such neurons (Figure 1).^{16–19} Archaerhodopsins inhibit neuronal firing by pumping protons out of the cell.

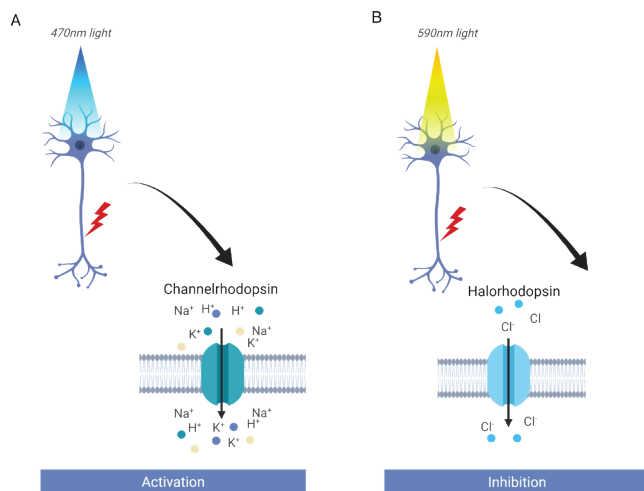


Figure 1: Optical excitation and inhibition of neurons mediated by microbial opsins. Activation of channelrhodopsin channelrhodopsin-2 (ChR2) and *halorhodopsin*. ChR2 is activated by blue light (~470nm), opening a light light-gated cation channel, permitting the influx of positively charged ions such as Na⁺ and K⁺, resulting in membrane depolarization and increased neuronal firing. *Halorhodopsin* is activated by yellow light (590nm), driving chloride ions (Cl⁻) into the neuron, producing membrane hyperpolarization and suppression of action potential generation. Together, excitatory and inhibitory opsins enable wavelength-dependent, bidirectional control of neuronal activity, forming the mechanistic basis of optogenetic modulation of neural circuits (created in BioRender).

Genetic delivery methods:

In optogenetics, efficient gene-delivery techniques are essential for successful optogenetic applications; virus-mediated delivery and transgenic animal models are primary approaches for introducing light-sensitive proteins into target cells. Virus-mediated delivery utilizes viral vectors such as adeno-associated virus (AAV), lentivirus, and retrovirus.²⁰ These viral vectors have been genetically engineered to remove pathogenic components while retaining their capacity to insert transgenes into target cells.²⁰ These vectors efficiently transport genetic material across the cell membrane and into the nucleus without being degraded by lysosomes.²¹ While viral methods are more popular in clinical trials, comprising about 80% of all approved gene therapy protocols, they are not without limitations.²¹ For example, delivery to the central nervous system often requires invasive stereotactic injection, which carries risks including infection, hemorrhage, or tissue damage.²² Additionally, viral vectors may have limited spread within the brain, making it difficult to target large or anatomically complex epileptic networks, and immune responses against viral vectors remain a concern.²² As such, further research is needed to improve upon these challenges.

Light delivery technologies:

Light delivery technologies in optogenetics, primarily optical fibers and light-emitting diodes (LEDs), enable precise control of opsin activation in targeted tissues with minimal invasiveness. Optical fibers are commonly used in deep brain stimulation (DBS) because they can deliver high-intensity light with spatial precision to specific neural populations. Optical fibers are often surgically implanted to deliver light from external sources, such as lasers, directly to targeted brain regions, with careful consideration to minimize heat and light-induced tissue damage.²³ LEDs, on the other hand, provide a more compact and flexible alternative, especially useful for surface or shallow brain regions. LEDs can be integrated into head-mounted or implantable systems, allowing animals to move freely during experiments while still receiving targeted light stimulation.²³ Although optical fibers offer stronger light penetration and precise delivery, they can be more invasive and may disrupt tissue.²³ Recent advances in light delivery focus on improving precision and reducing tissue damage with techniques that offer selective targeting and high efficiency while maintaining cell viability, highlighting progress toward more targeted and less invasive methods.²⁴

Advantages of the optogenetics technique:

The advantages of optogenetics allow researchers to manipulate neuronal activity using temporal precision, bidirectional modulation, and cell-type specificity. Because optogenetics introduces light-sensitive opsins into targeted cells, precise control over specific neuron types can be achieved.⁵ The ability to restrict opsin expression to selected neuronal populations ensures that experimental manipulations affect only intended circuits, reducing the side effects and improving the reliability of results.²⁵ These opsins allow researchers to control the generation of action potentials, rapid electrical signals that

neurons use to communicate, with accuracy, providing the high temporal resolution essential for decoding neural activity patterns underlying behavior.⁵ This temporal precision allows scientists to deliver light pulses in precise sequences to generate specific firing patterns and then observe their effects, a process critical for understanding neural coding.⁵ Moreover, optogenetics offers bidirectional control, enabling both activation and inhibition of neural circuits; this control makes it more feasible to stimulate or suppress brain functions within the same network, even in complex regions such as the cortex.¹⁹ Thus, these features make optogenetics an advantageous tool for studying brain functions and developing treatments with minimized collateral damage.

■ Application of Optogenetics in Epilepsy

Applying optogenetics to temporal lobe epilepsy:

Temporal lobe epilepsy is characterized by focal seizure onset within mesial temporal structures, making it particularly amenable to spatially restricted interventions (Table 1).²⁶ Researchers utilized optogenetic techniques in a mouse model of temporal lobe epilepsy, demonstrating promising results in real-time seizure suppression through targeted intervention. Using a closed-loop detection system, researchers identified spontaneous seizures and responded immediately with light-based activation or inhibition of specific cell populations.²⁷ This strategy was motivated by the focal onset of temporal lobe seizures and the clinical need for seizure-abortive therapies that act precisely at seizure onset, particularly for patients who are not candidates for surgical resection or who are refractory to antiepileptic drugs.²⁷ The first strategy was to directly inhibit excitatory principal neurons using *halorhodopsin*, and the second was to activate GABAergic interneurons expressing ChR2. In this model, opsin expression was restricted to specific neuronal populations using Cre-dependent genetic approaches, ensuring cell-type specificity, while seizure detection relied on depth electrodes and real-time analysis of electrographic seizure signatures.²⁷ Upon seizure detection, light was delivered directly to the ipsilateral hippocampus, resulting in rapid suppression of ongoing seizures. Both approaches successfully stopped ongoing electrographic seizures upon light application, with significant reductions in behavioral seizure frequency as well.²⁷ Although both methods effectively suppressed seizures, they differed in mechanism, suggesting that multiple optogenetic strategies may be capable of restoring excitatory-inhibitory balance in epileptic circuits.²⁷ However, these studies remain limited because researchers were only able to target a small subset of hippocampal neurons, highlighting ongoing challenges in achieving widespread and precise neuronal targeting.²⁷ Despite these limitations, the integration of spontaneous seizure detection and optogenetic intervention in real time demonstrates significant future therapeutic potential.

Applying optogenetics to neocortical epilepsy:

Neocortical epilepsy represents a distinct focal epilepsy subtype in which seizures originate from lateral or extrahippocampal cortical regions and often require advanced diagnostic techniques for accurate localization (Table 1).²⁸ In a well-char-

acterized rodent model of neocortical epilepsy, focal injection of tetanus toxin, which interferes with inhibitory and excitatory signaling between neurons, into the motor cortex produces chronic, frequent seizures with high-frequency epileptiform EEG activity and contralateral motor signs that persist for months and are largely unresponsive to systemic diazepam or phenytoin.²⁹ Focal pharmacology at the seizure focus, however, can acutely suppress activity, as local application of NBQX, a highly selective competitive antagonist of AMPA, reversibly reduced both EEG and behavioral seizures without detectable side effects, underscoring the value of spatially targeted interventions.^{29,30} AMPA receptors are glutamate-gated ion channels crucial in rapid excitatory communication within the brain. Because epileptic seizures are associated with excessive excitatory signaling, selective blockade of these receptors using NBQX decreases neuronal overactivation and could limit seizures.³⁰ Building on this concept, continuous wireless EEG monitoring in the same tetanus toxin model revealed increased high-frequency activity and more frequent epileptiform events, along with heightened intrinsic excitability of pyramidal neurons at the epileptic focus, consistent with a local generator for seizures.³¹ Optogenetic inhibition of a subset of principal neurons was achieved through the delivery of *halorhodopsin* selectively restricted to the epileptogenic zone. Light-mediated suppression of these neurons significantly reduced electroencephalographic seizure activity *in vivo*, demonstrating that inhibiting a relatively small population of hyperexcitable cortical neurons is sufficient to disrupt seizure generation.³¹ These findings demonstrate how optogenetics may offer more targeted seizure control by selectively modulating pathological neural circuits while minimizing the effects on surrounding healthy tissue.^{29–31}

Applying optogenetics to post-stroke epilepsy:

Post-stroke epilepsy arises from structurally damaged cortex and associated networks, with seizure onset often occurring weeks to months after the initial injury (Table 1).³² It can be modeled by photothrombotic infarction, where large rose-bengal-induced cortical lesions extending to the cortical-subcortical interface lead to spontaneous seizures that emerge weeks after injury, with ictal discharges arising from peri-infarct cortex and characteristic motor arrest behaviors.³³ Longitudinal *in vivo* imaging of cortical pyramidal neurons demonstrates that excitatory activity in the peri-injury cortex is markedly reduced in the days immediately following stroke, as evidenced by decreased calcium spike frequency and fewer active neurons, before gradually recovering toward baseline levels, consistent with compensatory homeostatic regulation in surviving neurons.³⁴ Importantly, this transient loss of activity is associated with heightened seizure susceptibility, suggesting that insufficient excitatory drive may contribute to the development of post-stroke hyperexcitability. To test whether restoring activity could counteract this process, optogenetic stimulation was applied to selectively enhance pyramidal neuron activity in the peri-infarct cortex during the subacute post-stroke period. Targeted stimulation significantly reduced seizure susceptibility compared to unstimulated stroke con-

trols, indicating that optogenetic enhancement of cortical activity can prevent the emergence of severe seizures.³⁴ These findings suggest that post-stroke epilepsy may be prevented by timely optogenetic modulation that restores homeostatic balance within peri-infarct networks, highlighting activity enhancement as a novel and preventative therapeutic strategy rather than solely a reactive seizure-suppression approach.

Table 1: Clinical characteristics of focal epilepsy subtypes. Overview of major focal epilepsy subtypes discussed in this review, highlighting differences in seizure origin and clinical presentation.

Epilepsy Type	Primary Seizure Origin	Key Clinical Features
Temporal Lobe Epilepsy (TLE)	Mesial temporal structures (hippocampus, amygdala). ²⁶	Most common focal epilepsy; frequent auras and automatisms; often medically refractory. ²⁶
Neocortical Temporal Lobe Epilepsy (nTLE)	Lateral or extrahippocampal neocortex. ²⁸	Less common than mesial TLE; overlapping symptoms complicate localization. ^{28,31}
Post-Stroke Epilepsy (PSE)	Peri-infarct cortex and connected networks. ³²	Delayed seizure onset following stroke; high recurrence risk after late unprovoked seizures. ³²

■ **Therapeutic Implications and Future Directions (Limitations)**

Potential for clinical translation:

Optogenetics holds significant translational promise as a personalized therapy for epilepsy, offering targeted and reversible seizure control through closed-loop light delivery systems. A major vision for the future is a closed-loop implantable system that can detect abnormal brain activity in real-time and deliver light stimulation to suppress seizures before they spread, offering a more precise alternative to current responsive neurostimulation, which relies on electrical stimulation delivered to relatively broad neural populations, and can inadvertently modulate surrounding circuits, potentially limiting efficacy and contributing to off-target effects.^{7,35} Early studies in rodents have already demonstrated that implantable light-emitting devices can dynamically halt seizures, and encouraging progress has been made in nonhuman primates, where stable opsin expression has been paired with behavioral modulation and MRI-compatible optrodes.^{6,36,37} Computational modeling further supports the feasibility of scaling this approach to the human brain.³⁸ In the long term, optogenetics may even influence drug development by helping identify the neuronal subtypes most critical for seizure generation, thereby guiding new pharmacological strategies. Collectively, these directions highlight the promise of optogenetics as a foundation for future personalized therapies in epilepsy.⁷

Limitations and challenges:

While optogenetics shows immense promise as a therapeutic tool for epilepsy, significant technical, biological, and ethical challenges must be addressed before clinical translation is possible. On the technical side, the large size of the human brain compared to rodents makes it difficult for light to penetrate enough tissue to modulate an epileptic network effectively, and deeper brain foci may be inaccessible without

invasive procedures.^{7,39} Multiple viral injections may be required to reach sufficient neurons, and even then, the limited penetration of visible light risks heating and damaging nearby cells.⁴⁰ Device miniaturization poses another barrier: closed-loop systems would require long-lasting, rechargeable power sources and flexible, durable optrodes capable of maintaining stability in the brain for years.⁷ Biologically, uncertainties remain about the safety and stability of opsin expression over decades in humans, since current methods rely on viral vectors such as AAV or lentivirus; although stable expression has been observed in nonhuman primates for months,^{5,35} the long-term persistence and safety in humans remain unclear.^{41–43} Immunogenicity is also a concern, as opsins are non-mammalian proteins that could elicit antibody-associated immune responses.³⁹ Moreover, functional risks exist, including the possibility that prolonged activation of *halorhodopsin* could shift intracellular chloride gradients, thereby paradoxically rendering GABA signaling excitatory.⁴⁴ Finally, ethical and regulatory considerations pose significant barriers, as invasive brain procedures, such as viral delivery or stem cell transplantation, may face resistance from institutional review boards.⁷ Permanent optical implants raise further concerns about reversibility and patient safety. While comparisons have been made to accepted interventions, including deep brain stimulation (DBS) and vagus nerve stimulation (VNS), optogenetics introduces the added complexity of genetic modification, intensifying regulatory scrutiny.³⁵ Together, these limitations highlight the critical obstacles that must be overcome before optogenetics can move from experimental models to clinical practice.

■ **Conclusion**

Epilepsy, a chronic neurological disorder characterized by recurrent seizures arising from abnormal and excessive neuronal synchronization, remains a major clinical challenge due to the limited specificity of current treatment options.⁴⁵ Optogenetics has revolutionized the study and potential treatment of epilepsy by enabling precise, light-based modulation of specific neuronal populations. Through the integration of microbial opsins, viral delivery systems, and advanced illumination technologies, researchers have achieved unprecedented control over neural circuits, demonstrating that selective activation or inhibition of targeted neurons can suppress seizures in multiple experimental models.⁴ These findings show the promise of optogenetics as a therapeutic approach that could eventually surpass the specificity and adaptability of current interventions such as pharmacological treatments or deep brain stimulation. However, several limitations remain before clinical translation can be realized. Restricted light penetration, concerns about viral vector safety, immune responses, and the long-term stability of opsin expression all pose major challenges for human application.^{41–43} In comparison, while current epilepsy treatments such as antiepileptic drugs and surgical interventions may provide partial seizure control, they lack precision and can fail in drug-resistant cases.⁴⁶ This lack of security in current therapies highlights the need for more targeted approaches. Ethical considerations regarding the invasiveness of optical implants, patient autonomy, long-term invasiveness of implanted optical

devices, and genetic modification further complicate clinical implementation.⁴⁷ Consequently, researchers have also raised concerns regarding the need for stronger legal and regulatory frameworks before optogenetics can be safely translated into clinical practice.⁴⁷ Future research should focus on improving gene delivery methods to ensure stable, safe, and cell-type-specific expression. By addressing these obstacles, optogenetics may ultimately evolve from an experimental technique into a clinically viable therapy.

■ Acknowledgments

I would like to sincerely thank Dr. Daniel Kim, PhD, for his mentorship and guidance, and Indigo Research for their support in the publication process. I am also extremely grateful for the encouragement I have received from my parents.

■ References

- Shorvon, S. D. The Etiologic Classification of Epilepsy. *Epilepsia* **2011**, 52 (6), 1052–1057. <https://doi.org/10.1111/j.1528-1167.2011.03041.x>.
- Alam, S. M. S.; Bhuiyan, M. I. H. Detection of Seizure and Epilepsy Using Higher Order Statistics in the EMD Domain. *IEEE J. Biomed. Health Inform.* **2013**, 17 (2), 312–318. <https://doi.org/10.1109/JBHI.2012.2237409>.
- Paz, J. T.; Davidson, T. J.; Frechette, E. S.; Delord, B.; Parada, I.; Peng, K.; Deisseroth, K.; Huguenard, J. R. Closed-Loop Optogenetic Control of Thalamus as a Tool for Interrupting Seizures after Cortical Injury. *Nat. Neurosci.* **2013**, 16 (1), 64–70. <https://doi.org/10.1038/nn.3269>.
- Deisseroth, K. Optogenetics. *Nat. Methods* **2011**, 8 (1), 26–29. <https://doi.org/10.1038/nmeth.f.324>.
- Boyden, E. S.; Zhang, F.; Bamberg, E.; Nagel, G.; Deisseroth, K. Millisecond-Timescale, Genetically Targeted Optical Control of Neural Activity. *Nat. Neurosci.* **2005**, 8 (9), 1263–1268. <https://doi.org/10.1038/nn1525>.
- Diester, I.; Kaufman, M. T.; Mogri, M.; Pashaie, R.; Goo, W.; Yizhar, O.; Ramakrishnan, C.; Deisseroth, K.; Shenoy, K. V. An Optogenetic Toolbox Designed for Primates. *Nat. Neurosci.* **2011**, 14 (3), 387–397. <https://doi.org/10.1038/nn.2749>.
- Bui, A. D.; Alexander, A.; Soltesz, I. Seizing Control: From Current Treatments to Optogenetic Interventions in Epilepsy. *Neurosci. Rev. J. Bringing Neurobiol. Neurol. Psychiatry* **2017**, 23 (1), 68–81. <https://doi.org/10.1177/1073858415619600>.
- Deisseroth, K.; Feng, G.; Majewska, A. K.; Miesenböck, G.; Ting, A.; Schnitzer, M. J. Next-Generation Optical Technologies for Illuminating Genetically Targeted Brain Circuits. *J. Neurosci. Off. J. Soc. Neurosci.* **2006**, 26 (41), 10380–10386. <https://doi.org/10.1523/JNEUROSCI.3863-06.2006>.
- Oesterhelt, D.; Stoeckenius, W. Rhodopsin-like Protein from the Purple Membrane of Halobacterium Halobium. *Nature. New Biol.* **1971**, 233 (39), 149–152. <https://doi.org/10.1038/newbio233149a0>.
- Sugiyama, Y.; Mukohata, Y. Isolation and Characterization of Halorhodopsin from Halobacterium Halobium. *J. Biochem. (Tokyo)* **1984**, 96 (2), 413–420. <https://doi.org/10.1093/oxfordjournals.jbchem.a134852>.
- Nagel, G.; Ollig, D.; Fuhrmann, M.; Kateriya, S.; Musti, A. M.; Bamberg, E.; Hegemann, P. Channelrhodopsin-1: A Light-Gated Proton Channel in Green Algae. *Science* **2002**, 296 (5577), 2395–2398. <https://doi.org/10.1126/science.1072068>.
- Nagel, G.; Szellas, T.; Kateriya, S.; Adeishvili, N.; Hegemann, P.; Bamberg, E. Channelrhodopsins: Directly Light-Gated Cation Channels. *Biochem. Soc. Trans.* **2005**, 33 (Pt 4), 863–866. <https://doi.org/10.1042/BST0330863>.
- Deisseroth, K.; Hegemann, P. The Form and Function of Channelrhodopsin. *Science* **2017**, 357 (6356), eaan5544. <https://doi.org/10.1126/science.aan5544>.
- Bi, A.; Cui, J.; Ma, Y.-P.; Olshevskaya, E.; Pu, M.; Dizhoor, A. M.; Pan, Z.-H. Ectopic Expression of a Microbial-Type Rhodopsin Restores Visual Responses in Mice with Photoreceptor Degeneration. *Neuron* **2006**, 50 (1), 23–33. <https://doi.org/10.1016/j.neuron.2006.02.026>.
- Poboży, K.; Poboży, T.; Domański, P.; Derczyński, M.; Konarski, W.; Domańska-Poboża, J. Evolution of Light-Sensitive Proteins in Optogenetic Approaches for Vision Restoration: A Comprehensive Review. *Biomedicines* **2025**, 13 (2), 429. <https://doi.org/10.3390/biomedicines13020429>.
- Han, X.; Boyden, E. S. Multiple-Color Optical Activation, Silencing, and Desynchronization of Neural Activity, with Single-Spike Temporal Resolution. *PLoS One* **2007**, 2 (3), e299. <https://doi.org/10.1371/journal.pone.0000299>.
- Lima, S. Q.; Hromádka, T.; Znamenskiy, P.; Zador, A. M. PINP: A New Method of Tagging Neuronal Populations for Identification during *in Vivo* Electrophysiological Recording. *PLoS One* **2009**, 4 (7), e6099. <https://doi.org/10.1371/journal.pone.0006099>.
- Shimano, T.; Fyk-Kolodziej, B.; Mirza, N.; Asako, M.; Tomoda, K.; Bledsoe, S.; Pan, Z. H.; Molitor, S.; Holt, A. G. Assessment of the AAV-Mediated Expression of Channelrhodopsin-2 and Halorhodopsin in Brainstem Neurons Mediating Auditory Signaling. *Brain Res.* **2013**, 1511, 138–152. <https://doi.org/10.1016/j.brainres.2012.10.030>.
- Zhang, F.; Wang, L.-P.; Brauner, M.; Liewald, J. F.; Kay, K.; Watzke, N.; Wood, P. G.; Bamberg, E.; Nagel, G.; Gottschalk, A.; Deisseroth, K. Multimodal Fast Optical Interrogation of Neural Circuitry. *Nature* **2007**, 446 (7136), 633–639. <https://doi.org/10.1038/nature05744>.
- Kamimura, K.; Suda, T.; Zhang, G.; Liu, D. Advances in Gene Delivery Systems. *Pharm. Med.* **2011**, 25 (5), 293–306. <https://doi.org/10.2165/11594020-000000000-00000>.
- Wivel, N. A.; Wilson, J. M. Methods of Gene Delivery. *Hematol. Oncol. Clin. North Am.* **1998**, 12 (3), 483–501. [https://doi.org/10.1016/s0889-8588\(05\)70004-6](https://doi.org/10.1016/s0889-8588(05)70004-6).
- Shen, Y.; Campbell, R. E.; Côté, D. C.; Paquet, M.-E. Challenges for Therapeutic Applications of Opsin-Based Optogenetic Tools in Humans. *Front. Neural Circuits* **2020**, 14, 41. <https://doi.org/10.3389/fncir.2020.00041>.
- Emiliani, V.; Entcheva, E.; Hedrich, R.; Hegemann, P.; Konrad, K. R.; Lüscher, C.; Mahn, M.; Pan, Z.-H.; Sims, R. R.; Vierock, J.; Yizhar, O. Optogenetics for Light Control of Biological Systems. *Nat. Rev. Methods Primer* **2022**, 2, 55. <https://doi.org/10.1038/s43586-022-00136-4>.
- Mohanty, S. K.; Lakshminarayanan, V. Optical Techniques in Optogenetics. *J. Mod. Opt.* **2015**, 62 (12), 949–970. <https://doi.org/10.1080/09500340.2015.1010620>.
- Lin, J. Y.; Lin, M. Z.; Steinbach, P.; Tsien, R. Y. Characterization of Engineered Channelrhodopsin Variants with Improved Properties and Kinetics. *Biophys. J.* **2009**, 96 (5), 1803–1814. <https://doi.org/10.1016/j.bpj.2008.11.034>.
- Blair, R. D. G. Temporal Lobe Epilepsy Semiology. *Epilepsy Res. Treat.* **2012**, 2012, 751510. <https://doi.org/10.1155/2012/751510>.
- Krook-Magnuson, E.; Armstrong, C.; Oijala, M.; Soltesz, I. On-Demand Optogenetic Control of Spontaneous Seizures in

- Temporal Lobe Epilepsy. *Nat. Commun.* **2013**, *4*, 1376. <https://doi.org/10.1038/ncomms2376>.
28. Bercovici, E.; Kumar, B. S.; Mirsattari, S. M. Neocortical Temporal Lobe Epilepsy. *Epilepsy Res. Treat.* **2012**, *2012*, 103160. <https://doi.org/10.1155/2012/103160>.
 29. Nilsen, K. E.; Walker, M. C.; Cock, H. R. Characterization of the Tetanus Toxin Model of Refractory Focal Neocortical Epilepsy in the Rat. *Epilepsia* **2005**, *46* (2), 179–187. <https://doi.org/10.1111/j.0013-9580.2005.26004.x>.
 30. Libbey, J. E.; Hanak, T. J.; Doty, D. J.; Wilcox, K. S.; Fujinami, R. S. NBQX, a Highly Selective Competitive Antagonist of AMPA and KA Ionotropic Glutamate Receptors, Increases Seizures and Mortality Following Picornavirus Infection. *Exp. Neurol.* **2016**, *280*, 89–96. <https://doi.org/10.1016/j.expneurol.2016.04.010>.
 31. Wykes, R. C.; Heeroma, J. H.; Mantoan, L.; Zheng, K.; MacDonald, D. C.; Deisseroth, K.; Hashemi, K. S.; Walker, M. C.; Schorge, S.; Kullmann, D. M. Optogenetic and Potassium Channel Gene Therapy in a Rodent Model of Focal Neocortical Epilepsy. *Sci. Transl. Med.* **2012**, *4* (161), 161ra152. <https://doi.org/10.1126/scitranslmed.3004190>.
 32. Tanaka, T.; Ihara, M. Post-Stroke Epilepsy. *Neurochem. Int.* **2017**, *107*, 219–228. <https://doi.org/10.1016/j.neuint.2017.02.002>.
 33. Kelly, K. M.; Kharlamov, A.; Hentosz, T. M.; Kharlamova, E. A.; Williamson, J. M.; Bertram, E. H.; Kapur, J.; Armstrong, D. M. Photothrombotic Brain Infarction Results in Seizure Activity in Aging Fischer 344 and Sprague Dawley Rats. *Epilepsy Res.* **2001**, *47* (3), 189–203. [https://doi.org/10.1016/s0920-1211\(01\)00294-7](https://doi.org/10.1016/s0920-1211(01)00294-7).
 34. Adhikari, Y.; Ma, C.-G.; Chai, Z.; Jin, X. Preventing Development of Post-Stroke Hyperexcitability by Optogenetic or Pharmacological Stimulation of Cortical Excitatory Activity. *Neurobiol. Dis.* **2023**, *184*, 106233. <https://doi.org/10.1016/j.nbd.2023.106233>.
 35. Tsytsarev, V.; Sopova, J. V.; Leonova, E. I.; Inyushin, M.; Markina, A. A.; Chirinskaite, A. V.; Volnova, A. B. Neurophonic Methods in Approach to *in Vivo* Animal Epileptic Models: Advantages and Limitations. *Epilepsia* **2024**, *65* (3), 600–614. <https://doi.org/10.1111/epi.17870>.
 36. Gerits, A.; Farivar, R.; Rosen, B. R.; Wald, L. L.; Boyden, E. S.; Vanduffel, W. Optogenetically Induced Behavioral and Functional Network Changes in Primates. *Curr. Biol. CB* **2012**, *22* (18), 1722–1726. <https://doi.org/10.1016/j.cub.2012.07.023>.
 37. Han, X.; Qian, X.; Bernstein, J. G.; Zhou, H.-H.; Franzesi, G. T.; Stern, P.; Bronson, R. T.; Graybiel, A. M.; Desimone, R.; Boyden, E. S. Millisecond-Timescale Optical Control of Neural Dynamics in the Nonhuman Primate Brain. *Neuron* **2009**, *62* (2), 191–198. <https://doi.org/10.1016/j.neuron.2009.03.011>.
 38. Selvaraj, P.; Sleight, J. W.; Freeman, W. J.; Kirsch, H. E.; Szeri, A. J. Open Loop Optogenetic Control of Simulated Cortical Epileptiform Activity. *J. Comput. Neurosci.* **2014**, *36* (3), 515–525. <https://doi.org/10.1007/s10827-013-0484-2>.
 39. Wykes, R. C.; Kullmann, D. M.; Pavlov, I.; Magloire, V. Optogenetic Approaches to Treat Epilepsy. *J. Neurosci. Methods* **2016**, *260*, 215–220. <https://doi.org/10.1016/j.jneumeth.2015.06.004>.
 40. Pavlov, I.; Kaila, K.; Kullmann, D. M.; Miles, R. Cortical Inhibition, pH and Cell Excitability in Epilepsy: What Are Optimal Targets for Antiepileptic Interventions? *J. Physiol.* **2013**, *591* (4), 765–774. <https://doi.org/10.1113/jphysiol.2012.237958>.
 41. Bartus, R. T.; Baumann, T. L.; Siffert, J.; Herzog, C. D.; Alterman, R.; Boulis, N.; Turner, D. A.; Stacy, M.; Lang, A. E.; Lozano, A. M.; Olanow, C. W. Safety/Feasibility of Targeting the Substantia Nigra with AAV2-Neurturin in Parkinson Patients. *Neurology* **2013**, *80* (18), 1698–1701. <https://doi.org/10.1212/WNL.0b013e3182904faa>.
 42. Fisher, R. S. Therapeutic Devices for Epilepsy. *Ann. Neurol.* **2012**, *71* (2), 157–168. <https://doi.org/10.1002/ana.22621>.
 43. Palfi, S.; Gurruchaga, J. M.; Ralph, G. S.; Lepetit, H.; Lavis, S.; Buttery, P. C.; Watts, C.; Miskin, J.; Kelleher, M.; Deeley, S.; Iwamuro, H.; Lefaucheur, J. P.; Thiriez, C.; Fenelon, G.; Lucas, C.; Brugières, P.; Gabriel, I.; Abhay, K.; Drouot, X.; Tani, N.; Kas, A.; Ghaleh, B.; Le Corvoisier, P.; Dolphin, P.; Breen, D. P.; Mason, S.; Guzman, N. V.; Mazarakis, N. D.; Radcliffe, P. A.; Harrop, R.; Kingsman, S. M.; Rascol, O.; Naylor, S.; Barker, R. A.; Hantraye, P.; Remy, P.; Cesaro, P.; Mitrophanous, K. A. Long-Term Safety and Tolerability of ProSavin, a Lentiviral Vector-Based Gene Therapy for Parkinson's Disease: A Dose Escalation, Open-Label, Phase 1/2 Trial. *Lancet Lond. Engl.* **2014**, *383* (9923), 1138–1146. [https://doi.org/10.1016/S0140-6736\(13\)61939-X](https://doi.org/10.1016/S0140-6736(13)61939-X).
 44. Raimondo, J. V.; Kay, L.; Ellender, T. J.; Akerman, C. J. Optogenetic Silencing Strategies Differ in Their Effects on Inhibitory Synaptic Transmission. *Nat. Neurosci.* **2012**, *15* (8), 1102–1104. <https://doi.org/10.1038/nn.3143>.
 45. Stafstrom, C. E.; Carmant, L. Seizures and Epilepsy: An Overview for Neuroscientists. *Cold Spring Harb. Perspect. Med.* **2015**, *5* (6), a022426. <https://doi.org/10.1101/cshperspect.a022426>.
 46. Berg, A. T.; Langfitt, J.; Shinnar, S.; Vickrey, B. G.; Sperling, M. R.; Walczak, T.; Bazil, C.; Pacia, S. V.; Spencer, S. S. How Long Does It Take for Partial Epilepsy to Become Intractable? *Neurology* **2003**, *60* (2), 186–190. <https://doi.org/10.1212/01.wnl.0000031792.89992.ec>.
 47. White, M.; Mackay, M.; Whittaker, R. Taking Optogenetics into the Human Brain: Opportunities and Challenges in Clinical Trial Design. *Open Access J. Clin. Trials* **2020**, *2020*, 33–41. <https://doi.org/10.2147/OAJCT.S259702>.

■ Authors

Alisha Xu is a junior at Milton Academy from Massachusetts. She has always shown a love for biology and wishes to continue this passion in the future through studying medicine. Alisha enjoys art, music, and playing various sports.

The Effects of Mild Traumatic Brain Injuries (mTBIs) on Oculomotor Functions and Effective Treatments

André Kurahashi Dantas

American School of Campinas, R. Cajamar, 35 - Jardim Alto da Barra, Campinas, São Paulo, 13090-860, Brazil; 360andre.dantas@gmail.com

ABSTRACT: Traumatic brain injuries (TBIs) occur following external impacts, altering brain functions, such as visual processing and cognitive abilities. Of 1.7 million individuals affected by TBI, 70-80% have mild TBI (mTBI). In 90% of mTBI cases, individuals experience oculomotor dysfunctions. To date, there is no known comprehensive review that evaluates the effects of oculomotor dysfunctions following mild TBIs and the effectiveness of specific treatments to improve these conditions. As such, this review aims to provide a comprehensive overview of how mTBI influences oculomotor functions and visual processing. Mild TBI can affect multiple oculomotor functions, including vergence, accommodation, and version. Oculomotor dysfunctions resulting from mTBI can hinder daily activities, such as reading. Four different studies employed distinct, effective oculomotor training designs: Thiagarajan and colleagues employed effective oculomotor neurorehabilitation for reading; Cade and Turnbull utilized effective chiropractic intervention for oculomotor dysfunctions; Thiagarajan and Ciuffreda implemented effective oculomotor rehabilitation for vergence responsivity; and Smaakjær and colleagues trialed effective vision therapy for binocular vision and ocular motility. Overall, this review contributes to raising medical awareness surrounding oculomotor dysfunctions following mTBI and appropriate treatments for such impairments.

KEYWORDS: Medical and Health Sciences, Mild Traumatic Brain Injury (mTBI), Oculomotor; Traumatic Brain Injury (TBI), Vergence, Version, Accommodation, Effective Treatments.

■ Introduction

Traumatic brain injuries (TBIs) are the result of external impacts (blows/jolts) that directly alter brain functions, such as visual processing and cognitive functions, producing a concussion-like condition in the individual affected.¹ TBIs can occur at any age. However, the severity of the symptoms varies from young to elderly. For example, individuals 75 years and older have higher TBI-related hospitalizations and deaths than younger individuals, who tend to recover faster and more efficiently following TBI.² This widespread injury (TBI) affects nearly 1.7 million individuals in the United States of America annually, of which 70-80% are mild TBIs (mTBI).³ As such, due to this commonality, mTBIs will be the main injury addressed in this paper.

Additionally, of the 1.7 million cases of mTBIs, 90% of individuals experienced oculomotor dysfunctions as a result.³ Oculomotor dysfunctions are common results of mTBIs. Given that oculomotor functions are related to all four brain lobes, in a 2020 clinical trial conducted by Berryman *et al.*, experts revealed how mTBI can produce symptoms that impair an individual's regular oculomotor function.⁴ For example, oculomotor functions impaired following mTBIs include pursuits, saccades, convergence, fixations, and vestibular ocular reflex, which consequently negatively impact the day-to-day routines of affected individuals.⁴ A daily activity examined in a 2014 study by Thiagarajan and Ciuffreda was reading. These researchers conducted a crossover study to evaluate the effects of oculomotor-based vision rehabilitation (OBVR) on visual impairments following mTBIs.⁵ To evaluate the effects of OBVR treatments (which included oculomotor training (OMT) and

sham training (ST)), reading exercises were administered. These readings consisted of a combination of saccadic eye movements, coordination of the individual's vergence eye movements, constant synchrony between ocular accommodation and vergence, along with continuous visual attention elements to process text.⁵ The authors found that after the OMT, 80% of patients improved in the abnormal parameters measured, such as reading rates and amplitudes of vergence and accommodation.⁵ Hence, when mTBIs impact the brain's standard processing of these functions, the individual's ability to read becomes compromised, and OBVR appears to be effective in treating oculomotor dysfunctions.

In addition to reading, there are other daily activities hindered by visual impairments in individuals who have experienced mTBIs. Dysfunctional oculomotor functions (pursuits, saccades, etc.) can hinder an individual's academic performance when required to extensively engage cognitive functions.⁶ This means that one will demonstrate poor effectiveness in computer-related tasks, taking notes in class, and head movements in dynamic sports activities.⁶ Ultimately, it is critical to stress the importance and severity of visual dysfunctions following mTBIs, as they can negatively impact daily activities regularly performed by other individuals.

Although there is a 2017 review of visual impairments following TBIs, the severity of the conditions addressed remains unknown. Additionally, a 2025 review addresses oculomotor dysfunctions following mTBI; however, it lacks evaluations of effective treatments regarding oculomotor dysfunctions. To date, there has yet to be a comprehensive review evaluating the impacts of oculomotor dysfunctions following mild TBIs

and effective treatments to improve such conditions. Taken together, the existing literature reviews lack recent relevance, as one was published eight years ago, and the other does not comprehensively evaluate effective solutions to visual impairments. Therefore, a review incorporating more recent evidence is needed to examine oculomotor dysfunctions following mTBIs and evidence-based treatments.

As such, this literature review aims to provide a comprehensive review of how mild traumatic brain injuries affect oculomotor functions and visual processing, and the methods that have been shown to improve such conditions. A thorough review of the literature may provoke further discussion, leading to an increase in medical awareness of mTBIs' influence on visual impairments. Once medical providers are more aware of this relationship, more efficient treatment protocols for patients with ocular dysfunctions following mTBI can be developed to minimize the harmful effects of visual impairments following mTBIs.

■ Discussion

Oculomotor Dysfunctions Following mTBI:

Following a mTBI, the four brain lobes are damaged, meaning that oculomotor functions become hindered as a result. In general, visual impairments are caused by the hindrance of vergence, accommodation, or version functions, all of which are key elements required to perform many acts done by the eye in an individual's daily activities.³ Due to the complication of such key elements, the eye has further difficulties with fixation, saccades, and pursuit, resulting in the affected individual experiencing difficulty in performing normal vision-related acts.^{7,8} Such hindered acts, as mentioned previously, include athletic activities, note-taking, sleeping, and tracking any moving objects;⁶⁻⁹ all of which ultimately reduce one's quality of life.

Vergence is a key oculomotor function where the eyes perform the function that tracks objects as they move in depth over the range of an individual's binocular view, effectively aiming to maintain a constant focus and singleness of the targeted object.^{3,10} Additionally, the vergence operating system comprises multiple neuronal pathways susceptible to damage from mTBIs,¹⁰ effectively producing an oculomotor dysfunction that hinders one's daily activities. The neuronal pathways, where discharge of neurons occurs during vergence, include the cerebellum, the pons, and areas of the cerebral cortex, including the frontal eye field, middle temporal, and medial superior temporal visual areas, as well as the parietal lobe.³ Hence, when mTBI occurs in an individual, these neural systems are impacted and ultimately impaired, which translates to further negative consequences on the vergence of the mTBI-affected individual. Vergence abnormalities, moreover, can be inferred to be among the most common dysfunctions following TBIs, as seen in the 2007 study by Ciuffreda *et al.*, in which 56.3% of individuals placed in the TBI subgroup were identified as having vergence dysfunctions.¹¹ However, Ciuffreda and colleagues did not specify the type of TBI; therefore, it is unclear the percentage of individuals with mTBIs that would have vergence dysfunctions.

Convergence insufficiency is arguably the most common type of vergence dysfunction, as 42.5% of individuals with vergence abnormalities suffer from convergence insufficiency.^{10,11} More specifically, convergence insufficiency is an oculomotor vergence abnormality where the individual experiences a binocular vision disorder, with a receded near point of convergence and reduced positive fusional vergence at near.¹² This impairment is caused by the nerves surrounding the six eye muscles—superior and inferior rectus muscles, the lateral and medial rectus muscles, and the superior and inferior oblique muscles—¹³ not firing the proper signals for ocular movement.¹⁴ Figure 1 illustrates improper nerve signaling and how convergence insufficiency occurs, using content from Parker E. Ludwig and colleagues' book as a reference for the simple visualization of the eye muscles.¹⁵ As such, this improper functioning of nerves can be caused by mTBIs, and as these neuronal pathways become impaired, they consequently impact the convergence system of the eyes.

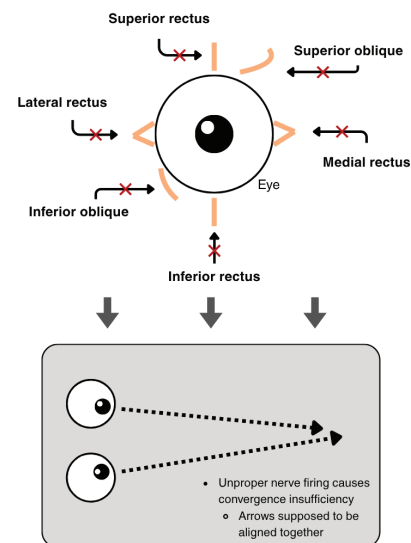


Figure 1: Illustration of where the six ocular muscles are located and their improper signaling (indicated by the red "X"). The grey box below illustrates the effects of convergence insufficiency on ocular function, demonstrating how the individual's ability to fixate at a near object is limited due to drifting outwards from the eyes.¹⁶ Made using Canva.

Accommodation is an understudied oculomotor dysfunction that involves the eyes' ability to change focus between objects at different distances, altering the shape of the eye's lens.¹⁷ Often, individuals who experience mTBI later develop accommodative impairments.¹⁸ Accommodative insufficiency refers to the impaired dynamics of accommodation, which is the reduced capability of sustaining focus on specific targets at short distances;¹⁹ impaired dynamic accommodation includes reduced amplitude, lower peak velocity, and an increased accommodation time constant, potentially causing headaches, double vision, eye strain, and intermittent blur.¹⁸ A 2019 meta-analysis by Natalya Merezhinskaya and colleagues emphasized the occurrence of individuals with accommodative dysfunction following traumatic brain injury (TBI, not to be confused with mTBI), sharing a 42.8% prevalence (95% confidence interval: 31.3 to 54.7) across the 1,271 TBI patients screened primarily in the United States, but also from Asia,

Africa, and Europe.²⁰ It is critical to note that the majority of studies reviewed by Merezhinskaya *et al.* included TBI patients, while the remaining reported on mild TBI.²⁰ This limits the validity of the prevalence of accommodative dysfunction in individuals following mTBI.

Version is another oculomotor function that is impaired following a mTBI. Version, not to be confused with vergence, refers to the synchronous, symmetrical movement of both eyes to interpret a designated element and is composed of saccadic and pursuit movements.^{7,19} More specifically, versional eye movement dysfunction refers to delayed and incorrect saccadic eye movements, along with the impaired optical ability to pursue target elements.²² Interestingly, version and vergence are quite related, as these ocular functions operate using similar neural pathways, operating with precision and synchrony to track the target element.³ Evidence suggests that versional eye movement dysfunction is caused by diffuse axonal injury, which is a downstream effect of mTBI.²² Diffuse axonal injury occurs when acceleration forces induce a rapid stretch of the axon, disrupting microtubules and producing focal breaks.²³ Diffuse axonal injury severely impairs the relay of the occipital lobe to execute visuospatial information in the frontal and parietal lobes, known as the saccadic center.²² Figure 2 visualizes and summarizes the pathways of oculomotor impairments following mTBIs.

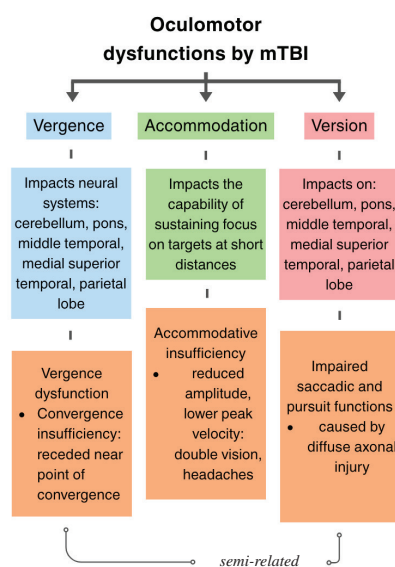


Figure 2: Oculomotor dysfunctions caused by mTBI and their related impacts flowchart. Made using Canva.

Effects of mTBI on Reading:

Following mTBIs, affected individuals can experience several visual impairments.⁵ These visual dysfunctions can lead to many daily functions being disrupted due to mTBIs. One general visual impairment associated with mTBI symptoms is difficulty reading. For an individual to read text, precise coordination of versional eye movements, synchrony with vergence and ocular accommodation, and a constant, advanced attentional state are required. Specifically, the automatic saccadic eye movements during reading involve shifting the eyes left and right, lasting 30–60 milliseconds and having an amplitude of 1–3 degrees.⁵ These are followed by around 250 milliseconds

of fixations for text processing, with occasional regressions to re-read information.⁵ As such, for reading to occur efficiently, all these complex ocular mechanisms must occur synchronously to achieve a truly efficient and functional reading.

These complex oculomotor mechanisms, however, are susceptible to the effects of mTBIs. As such, mTBIs can significantly affect the precise functions that comprise reading. The impairments caused by a mTBI are oculomotor-based, meaning that the eye's motor functioning and capability are hindered due to this injury. For example, since saccadic eye movements are primarily processed in the frontal lobe and the superior colliculus region of the midbrain,¹³ if mTBIs affect these areas, the individual's reading ability will ultimately become difficult due to the dysfunctional saccadic eye movement. Hence, the precision of the saccadic eye movements is severely reduced (due to mTBIs). From this reduction in saccadic precision, affected individuals tend to make inaccurate saccades with varying amplitude degrees, ultimately producing slower reading rates as they experience difficulty fixating words and maintaining focus within the line currently being read.²⁴ Additionally, precision and ease of reading are further affected by vergence changes resulting from mTBIs. Since the precise simultaneous movement of both eyes is impaired, diplopia, or partially overlapping images, can occur intermittently, resulting in difficulty for the individual in organizing and identifying words in the text.^{5,25} For better visualization, Figure 3 illustrates side-by-side the comparison between normal readers and readers affected by visual impairments. Overall, this can limit the learning and comprehension rate of the individual, due to a longer period of time being needed to understand the information given in the text.²⁴

Comparison of reading processes

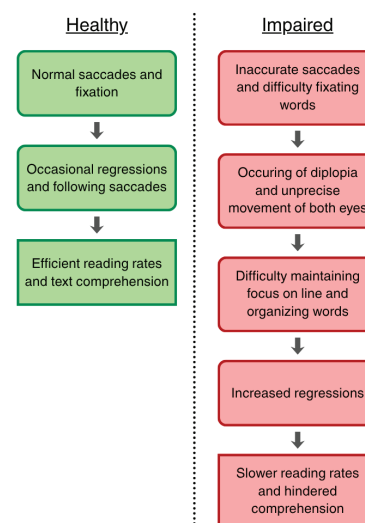


Figure 3: Comparison of occurrences in reading between normal individuals and mTBI-affected individuals. Made using Canva.

Effective Treatments for the Oculomotor Dysfunctions:

In addition to the oculomotor dysfunctions explained previously, numerous researchers and colleagues have undergone multiple trials to evaluate the efficacy of their specific treatment in improving visual impairments and optical abilities.

This section will critically review and analyze four studies that employed various treatments for oculomotor dysfunctions, providing a final overview of the effectiveness of the methodological design.

In *Effects of mTBI on Reading*, Thiagarajan and colleagues conducted experiments in 2014 to identify more efficient methods to improve recovery in patients with impairments in the vergence, version, and accommodative systems. This was a crossover, interventional study that utilized oculomotor training (OMT) and sham training (ST) over 6 weeks (2 sessions per week, each lasting 1 hour).⁵ The OMT group had three training sections, designed to train the version, vergence, and accommodative systems.⁵ The ST group also underwent these three training sections; however, they received a sham version/vergence/accommodation training.⁵

The parameters measured by Thiagarajan and colleagues were actions performed during reading and tests (to evaluate nearvision-related symptoms) that included the Convergence Insufficiency Symptom Survey (CISS) and the Visual Search Attention Test (VSAT). Overall, the visagraph parameters collected post-OMT improved, although some differences were not statistically significant. Specifically, a significant 25% increase in reading rate (wpm) occurred post-OMT, increasing from 142 (baseline) to 177 (post-OMT); the number of fixations per one hundred words decreased significantly, from 164 (baseline) to 135 (post-OMT); and the grade level efficiency increased significantly by two levels, also considered clinically significant, from 4.1 (baseline) to 6.3 (post-OMT).⁵ Evidently, the data collected demonstrate a significant difference between baseline and post-OMT ocular abilities, with visagraph parameters post-OMT all demonstrating positive improvements, although some parameters had p-values that made the data statistically insignificant.⁵

As such, this OMT design to improve oculomotor function in affected individuals can be considered to be efficient. Across all visagraph parameters, the post-OMT data demonstrated improvements in oculomotor functions, with one limitation being the statistical significance of the results, as half of the presented data were not statistically significant.⁵ However, Thiagarajan and colleagues explain this insignificance clearly, such as the possible relatively high inter-subject variability.⁵ In conclusion, the OMT designed by Thiagarajan and colleagues was shown to be effective, with improvements across all visagraph parameters measured by the specialists.

Another study, conducted in 2024 by Alice E. Cade and Philip R. K. Turnbull, in which such experts evaluated the efficacy of chiropractic intervention on oculomotor and attentional visual outcomes in young adults with mTBI.⁷ This unique treatment originated from an observed correlation between convergence insufficiency/saccadic deficits, which altered proprioceptive drives, a function related to spinal joint dysfunction; hence, the inclusion of chiropractic intervention became a possibility.⁷

In this experiment, Cade and Turnbull conducted a single-blind, randomized controlled intervention study comparing a group receiving chiropractic intervention with a control group, where the chiropractic group received care

from four chiropractors instructed to utilize the best practice guidelines and scope of chiropractic practice.⁷ The data collection consisted of recording baseline computerized eye-tracker assessment (CEA) results in 40 young adults experiencing persistent postconcussion syndrome (PPCS) following mTBI.⁷ The assessment in the CEA included egocentric localization, fixation stability, pursuit, saccades, vestibulo-ocular reflex, and the Stroop Test.⁷

As a result, the collected data revealed that the chiropractic group significantly improved dynamic and static gaze stability, improved performance in the Stroop test (split into two parts, the first involved reading, and the second involved matching font colors), and reduced saccadic latency.⁷ Specifically, gaze stability improved following chiropractic intervention, with a previous value of 3.26 and a posterior value of 3.16 (expressed in $\log_{10}$ minarc²), in comparison with the control group, which worsened from a value of 2.94 to 3.12 ($\log_{10}$ minarc²).⁷ The chiropractic group revealed significant performance improvements in the first part of the Stroop test ($X^2[1, N = 40] = 3508.64$); however, the second part's performance remained the same in both the control and chiropractic group.⁷ A significant reduction in saccadic latency was observed in both groups, but the reduction was greater in the chiropractic group.⁷ Prosaccade reduced from 366.9 to 317.0 ms, and antisaccade reduced from 350.3 to 317.0 ms.⁷

In summary, participants in the chiropractic group scored higher than participants in the control group on the Stroop test, showed improved gaze stability in the vestibulo-ocular reflex and fixation stability test, and a decreased vertical error in egocentric localization (p-values indicating statistical significance).⁷ As such, chiropractic intervention proved beneficial and effective for oculomotor dysfunction following mTBI. This is due to observations of differences in data results between the chiropractic group and the control group, where some parameters measured in the control group worsened or did not change, whereas the chiropractic group showed improvements (e.g., Stroop test of selective attention).⁷ The importance of this finding is a new form of treatment for visual impairments, which increases medical awareness of new potential treatments to alleviate oculomotor dysfunctions and provides a better quality of life to affected individuals in different (potentially easier) methods.

Furthermore, in another experiment conducted by Thiagarajan and Kenneth J. Ciuffreda in 2013, these researchers addressed the impact of oculomotor training on vergence responsivity and nearwork-related symptoms in individuals following mTBI (not to be confused with the previous 2014 article focused on reading capabilities).³ The design of this experiment was a crossover, interventional study with single-blind allocation, with a total duration of 15 weeks, with two separate groups: the oculomotor training (OMT) group and the Placebo training group (P training).³ Similar to previous studies mentioned, the OMT group underwent three separate training sessions that included version, vergence, and accommodation training. However, only the results of the vergence training were reported, as it was the primary objective emphasized by Thiagarajan and Ciuffreda. Additionally, to specifically report

on the effects of OMT on vergence responsivity, some of the visual parameters addressed by Thiagarajan and Ciuffreda include the near point of convergence (NPC) break and recovery, positive fusional vergence (PFV) break and recovery, negative fusional vergence (NFV), horizontal near phoria, peak velocity ($^{\circ}/s$), and evaluations from both the CISS and VSAT tests.³

Following OMT, Thiagarajan and Ciuffreda recorded a significant increase in the peak velocity for convergence and divergence. Although peak velocity did not normalize, a significant benefit from OMT was revealed.³ The NPC break and recovery data showed improvement following OMT, as the values recorded decreased significantly for both the NPC break (4.07) and recovery (3.64).³ The PFV values also showed significant improvements, with the PFV break value significantly increasing by 2.80 and the recovery significantly increasing by 4.71.³ The NFV break and recovery also improved significantly, evidenced by the increase in the NFV break (3.40 and normalized) and recovery (2.04).³ Additionally, regarding the CISS and VSAT tests, the CISS resulted in a significant decrease in score (37 ± 4 to 28 ± 3) following OMT, and the VSAT percentiles significantly increased in 80% of the participants (32nd (± 9) to 50th (± 10) percentile).³ It is important to note that not all visual parameters recorded by Thiagarajan and Ciuffreda were mentioned in this paper.

As such, the results reported by Thiagarajan and Ciuffreda indicate the effectiveness of their methods and study design to perform OMT. As indicated by the recorded visual parameters, the OMT designed by Thiagarajan and Ciuffreda improved most laboratory and clinical vergence measures that were abnormal at baseline.³ For instance, the significant decrease/increase in CISS/VSAT scores following OMT indicates that the overall vergence oculomotor system improved and recovered efficiently. Therefore, the OMT design used in this experiment, conducted by Thiagarajan and Ciuffreda, is considered highly efficient for individuals experiencing vergence-related abnormalities following mTBI. Furthermore, a critical difference to note is the duration between this experiment (2013) and the one conducted in 2014: the 2013 experiment spanned 15 weeks, compared to 6 weeks in 2014. This can be associated with the greater number of significant results in the longer experiment (2013) and the greater number of insignificant results in this shorter experiment, thereby suggesting that time duration may be a critically influential variable for researchers when designing efficient OMT.

Additionally, another differentiated 2022 experiment was conducted by Peter Smaakjær and colleagues, aiming to specifically address the positive outcomes of binocular vision and ocular motility training in patients with oculomotor dysfunctions following mTBI.²⁶ Unlike the previous experiments discussed, this investigation was designed as a single-group study without randomization, meaning there were no control groups for comparison. This spanned 10 weeks, with 1-hour training sessions and a minimum of 20 minutes of at-home training each day (except on days when patients received the training session), investigating the effects of their OMT on 28 mTBI patients aged between 25 and 61.²⁶ These patients were previously examined and selected by a neuro-optometrist

and a speech and language pathologist to deduce if individuals fell in the exclusion criteria, giving an unblinded design to the experiment.²⁶ The specific OMT consisted of multiple tests, which were a modified Canadian Occupational Performance Measure (COPM) interview (relevant and important three self-selected activities), the Groffman Visual Tracking Test (GVT) (evaluating pursuit movements), the King-Devick test (K-D test) (testing saccadic eye movement), the Multidimensional Fatigue Inventory (MFI-20) questionnaire (measuring fatigue), and a Danish reading test utilizing a readability index used in Scandinavia (measuring reading rates).²⁶

In the final results collected from the trials, the data acquired and compared indicated overall improvements in the parameters analyzed by the researchers. COPM results at the end of the treatment, for example, indicated significant improvement in performance and satisfaction for the important self-selected chore, with one parameter (not the only one) being COPM2 Performance and Satisfaction, significantly increasing from 3 (2.5 - 4) to 6 (5 - 8) and 2 (1-3) to 7 (2 - 8), respectively.²⁶ More objective improvements were also reflected in data from the GVT and the K-D test, which changed significantly from 28 (20 - 33) to 34 (25-45) and from 83 (63 - 110) to 65 (53 - 85), respectively.²⁶ It is important to note that the increase in the GVT scores is considered an improvement, while a decrease in the K-D test is considered an improvement. The final data parameter presented in this review is the MFI-20 questionnaire, which also indicated significant positive evolution. This parameter significantly decreased from 73 (62 - 78) to 67 (50 - 74), indicating that patients experienced less fatigue (general, physical, and mental) and reduced activity/motivation following Smaakjær and colleagues' OMT design.²⁶

Based on the significant improvement data collected in this investigation, the OMT design used by Smaakjær and colleagues in 2022 can be considered effective and consistent with previous experiments. This is evidenced by changes in baseline test scores administered after the treatment period ended, indicating improvements in ocular functions, such as pursuit and saccades. Nonetheless, the true effectiveness of such OMT remains unknown, as the authors do not specify or guarantee that the patient was consistent and compliant with the required home training, nor do they provide means to verify this. This would be critical to include in this experiment, as patient adherence is not always guaranteed and tends to decline with increasing treatment duration.²⁷ Overall, aside from the limitations identified by the authors themselves, such as potential bias in COPM results and the uncontrolled, nonrandomized design,²⁶ the experiment itself is considered efficient and effective, providing an uncommon experimental design that ultimately produced a positive impact on oculomotor function. This is also supported by the uncommon inclusion of patients at a more advanced age,²⁶ allowing for a broader study of the effects of this vision therapy on that age group.

Overall, in close relation to all studies analyzed in this section, the 2013 study by Thiagarajan and Ciuffreda reported the strongest effect of their OMT design, as most visual parameters normalized and showed significant improvements post-OMT, with the longest trial duration among all studies:

15 weeks. Nonetheless, the study by Smaakjær and colleagues collected data on a broader, more comprehensive set of parameters, including not only oculomotor parameters but also fatigue and personal opinions on self-selected treatments. Furthermore, the 2014 study by Thiagarajan and colleagues can be considered more clinically accessible due to its relatively short 6-week duration and holistic OMT of vergence, accommodation, and version, all measured and tracked using commonly used visual parameters, such as the CISS and VSAT.^{5,28,29} However, the least replicable study can arguably be Smaakjær and colleagues' investigation, which is due to the lack of clarity and detailing of the procedures and specific actions/treatments done in the vision therapy employed, making it difficult for future researchers to access potential gaps in the methodology to possibly improve the OMT design. Finally, the four studies analyzed complement and contribute to one another by addressing features and elements not covered by the others and by revealing, directly and indirectly, critical limitations for future assessment and research. This ultimately highlights the understudied state of oculomotor dysfunctions post-mTBIs, underscoring numerous key areas that remain to be comprehensively investigated by future research.

■ Strengths and Limitations

A major strength of this review is its extensive evaluation of effective treatment options. Of the existing literature, none emphasized the importance of evaluating and reporting the effectiveness of specific oculomotor training in improving visual impairment outcomes in patients. Including this aspect enhances the article's analytical depth by allowing the audience to access not only a recent review of oculomotor dysfunctions but also a recent review of efficient treatments. Another strength of this review is that it revisits oculomotor dysfunctions, drawing on more recent literature in an effort to provide a more robust understanding of this topic, given that more empirical studies on this topic have been published. Thus, this updated review informs scientists about more recent understanding and effective treatments, which can lead to improvements in treatments for individuals with oculomotor dysfunction following mTBI.

A primary limitation of this article is that the effective treatments cannot be generalized across the elderly population, as the clinical trials employed a limited age range from 18 to 35 years old. The severity and impact of mTBI vary across age groups,² so excluding older age groups leaves more specific implications of mTBI unknown.

■ Conclusion

In conclusion, this review aimed to provide an understanding of how mild traumatic brain injuries affect oculomotor functions and visual processing, and to evaluate oculomotor training interventions that have been effective in improving oculomotor dysfunctions. Mild traumatic brain injury (mTBI) can hinder vergence functions when neural systems utilized in vergence are impacted, which ultimately impairs them. Mild TBI can further disrupt the accommodative function of the eyes, resulting in reduced ability to maintain focus on close

targets. Additionally, mTBI causes versional eye movement dysfunction, characterized by impaired pursuit of target elements and incorrect, delayed saccadic movements.

Furthermore, in this review, four clinical trials were chosen for evaluation on their effectiveness. Thiagarajan and colleagues' 2014 study employed effective oculomotor training (OMT), as post-OMT data on visagraph parameters showed improvements. In Cade and Turnbull's 2024 study, they utilized a unique OMT that included chiropractic interventions. This OMT design was also shown to be effective, as parameters in the chiropractic group improved compared with the control group, although some parameters did not improve. The 2014 study analyzed by Thiagarajan and Ciuffreda was also considered effective due to improvements in most laboratory and clinical vergence measures. Finally, the 2022 investigation by Smaakjær and colleagues was shown to be effective, with improvements across all parameters and significance in the majority of the parameters.

Given the limitations previously stated, it would be critical for future literature reviews on this topic to include details on the effects of oculomotor impairments and related treatments in the elderly population. This is crucial to include, as severity varies according to age.² Thus, with broader inclusion of age ranges in this topic, a more holistic and comprehensive understanding of the health consequences of mTBI would be available worldwide.

As such, all studies reviewed in this work indicate the intricate complexity of visual impairments and implications post-mTBIs. Despite extensive literature reviews and numerous clinical trials aimed at identifying new, effective treatments, oculomotor dysfunctions following mTBIs remain an understudied medical topic. This is due to existing literature reviews lacking the scope to encompass what was addressed here and to investigate these implications across a wider age range. Hence, to finalize, this review tackles such gaps (except age ranges), increasing medical awareness of the severity of oculomotor dysfunctions following mTBI and of effective treatments for these dysfunctions.

■ Acknowledgments

I would like to formally thank Dr. Bre Calhoun, my Research Mentor, and the Indigo Research team for their continuous aid and support in the production of this review. They have successfully guided me in the right direction and provided meaningful advice to help me achieve my academic aspirations. I would also like to thank my parents, Fernando Justino Dantas and Alessandra Kurahashi Dantas, for their constant motivation, encouragement, and support throughout every moment of my academic and personal endeavors.

■ References

1. National Institute of Neurological Disorders and Stroke. *Traumatic Brain Injury (TBI)*. <https://www.ninds.nih.gov/health-information/disorders/traumatic-brain-injury-tbi> (accessed 2025-10-30).
2. CDC. *Health Disparities in TBI*. Traumatic Brain Injury & Concussion. <https://www.cdc.gov/traumatic-brain-injury/health-equity/index.html> (accessed 2025-11-05).

3. Thiagarajan, P.; Ciuffreda, K. J. Effect of Oculomotor Rehabilitation on Vergence Responsivity in Mild Traumatic Brain Injury. *J. Rehabil. Res. Dev.* **2014**, *50* (9), 1223–1240. <https://doi.org/10.1682/JRRD.2012.12.0235>.
4. Berryman, A.; Rasavage, K.; Politzer, T.; Gerber, D. Oculomotor Treatment in Traumatic Brain Injury Rehabilitation: A Randomized Controlled Pilot Trial. *Am. J. Occup. Ther. Off. Publ. Am. Occup. Ther. Assoc.* **2020**, *74* (1), 7401185050p1-7401185050p7. <https://doi.org/10.5014/ajot.2020.026880>.
5. Thiagarajan, P.; Ciuffreda, K. J.; Capo-Aponte, J. E.; Ludlam, D. P.; Kapoor, N. Oculomotor Neurorehabilitation for Reading in Mild Traumatic Brain Injury (mTBI): An Integrative Approach. *NeuroRehabilitation* **2014**, *34* (1), 129–146. <https://doi.org/10.3233/NRE-131025>.
6. Kontos, A. P.; Deitrick, J. M.; Collins, M. W.; Mucha, A. Review of Vestibular and Oculomotor Screening and Concussion Rehabilitation. *J. Atbl. Train.* **2017**, *52* (3), 256–261. <https://doi.org/10.4085/1062-6050-51.11.05>.
7. Cade, A. E.; Turnbull, P. R. K. Effect of Chiropractic Intervention on Oculomotor and Attentional Visual Outcomes in Young Adults With Long-Term Mild Traumatic Brain Injury: A Randomized Controlled Trial. *J. Manipulative Physiol. Ther.* **2024**, *47* (1–4), 1–11. <https://doi.org/10.1016/j.jmpt.2024.08.003>.
8. Optometry Center for Vision Therapy. *Oculomotor Dysfunction*. The Optometry Center for Vision Therapy. <https://ocvt.info/oculomotor-dysfunction/> (accessed 2025-12-04).
9. McMahon, P.; Hricik, A.; Yue, J. K.; Puccio, A. M.; Inoue, T.; Ling-sma, H. F.; Beers, S. R.; Gordon, W. A.; Valadka, A. B.; Manley, G. T.; Okonkwo, D. O.; Casey, S. S.; Cooper, S. R.; Dams-O'Connor, K.; Menon, D. K.; Sorani, M. D.; Yuh, E. L.; Mukherjee, P.; Schnyer, D. M.; Vassar, M. J. Symptomatology and Functional Outcome in Mild Traumatic Brain Injury: Results from the Prospective TRACK-TBI Study. *J. Neurotrauma* **2014**, *31* (1), 26–33. <https://doi.org/10.1089/neu.2013.2984>.
10. Thiagarajan, P.; Ciuffreda, K. J.; Ludlam, D. P. Vergence Dysfunction in Mild Traumatic Brain Injury (mTBI): A Review. *Ophthalmic Physiol. Opt.* **2011**, *31* (5), 456–468. <https://doi.org/10.1111/j.1475-1313.2011.00831.x>.
11. Ciuffreda, K. J.; Kapoor, N.; Rutner, D.; Suchoff, I. B.; Han, M. E.; Craig, S. Occurrence of Oculomotor Dysfunctions in Acquired Brain Injury: A Retrospective Analysis. *Optom. St Louis, MO* **2007**, *78* (4), 155–161. <https://doi.org/10.1016/j.optm.2006.11.011>.
12. The Convergence Insufficiency Treatment Trial Study Group. The Convergence Insufficiency Treatment Trial: Design, Methods, and Baseline Data. *Ophthalmic Epidemiol.* **2008**, *15* (1), 24–36. <https://doi.org/10.1080/09286580701772037>.
13. Purves, D.; Augustine, G. J.; Fitzpatrick, D.; Katz, L. C.; LaMantia, A.-S.; McNamara, J. O.; Williams, S. M. Neural Control of Saccadic Eye Movements. In *Neuroscience. 2nd edition*; Sinauer Associates, 2001.
14. American Optometric Association. *Convergence insufficiency*. <https://www.aoa.org/healthy-eyes/eye-and-vision-conditions/convergence-insufficiency> (accessed 2025-12-10).
15. Ludwig, P. E.; Aslam, S.; Czyn, C. N. Anatomy, Head and Neck: Eye Muscles. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2026.
16. Goering, M.; Drennan, K. B.; Moshirfar, M. Convergence Insufficiency. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2026.
17. Colorado Eye Clinic. *How the Eye Focuses: Understanding Accommodation*. <https://coloradoeyeclinic.net/how-the-eye-focuses-explaining-accommodation/> (accessed 2025-12-22).
18. Almutairi, N.; Hayes, J.; Hampson, K.; Liu, C. Accommodation Microfluctuation in Individuals with mTBI and the Potential Effect of Chromatic Filter on This Parameter. *Vision Res.* **2025**, 227. <https://doi.org/10.1016/j.visres.2025.108545>.
19. Enaholo, E. S.; Musa, M. J.; Zeppieri, M. Accommodative Insufficiency. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2025.
20. Merezhinskaya, N.; Mallia, R. K.; Park, D.; Bryden, D. W.; Mathur, K.; Barker, F. M. I. Visual Deficits and Dysfunctions Associated with Traumatic Brain Injury: A Systematic Review and Meta-Analysis. *Optom. Vis. Sci.* **2019**, *96* (8), 542. <https://doi.org/10.1097/OPX.0000000000001407>.
21. Masson, G. S.; Yang, D.-S.; Miles, F. A. Version and Vergence Eye Movements in Humans: Open-Loop Dynamics Determined by Monocular Rather than Binocular Image Speed. *ResearchGate* **2002**. [https://doi.org/10.1016/S0042-6989\(02\)00334-6](https://doi.org/10.1016/S0042-6989(02)00334-6).
22. Almutairi, N. M. Visual Dysfunctions in Mild Traumatic Brain Injury: A Focus on Accommodative System Impairments. *Life* **2025**, *15* (5), 744. <https://doi.org/10.3390/life15050744>.
23. Mesfin, F. B.; Gupta, N.; Hays Shapshak, A.; Margetis, K. Diffuse Axonal Injury. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2025.
24. Facchin, A.; Maffioletti, S.; Maffioletti, M.; Esposito, G.; Bonetti, M.; Girelli, L.; Daini, R. Oculomotor Training Improves Reading and Associated Cognitive Functions in Children with Learning Difficulties: A Pilot Study. *Vision* **2025**, *9* (4), 83. <https://doi.org/10.3390/vision9040083>.
25. Najem, K.; Asuncion, R. M. D.; Margolin, E. Diplopia. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2026.
26. Smaakjær, P.; Wachner, L. G.; Rasmussen, R. S. Vision Therapy Improves Binocular Visual Dysfunction in Patients with Mild Traumatic Brain Injury. *Neurol. Res.* **2022**, *44* (5), 439–445. <https://doi.org/10.1080/01616412.2021.2000825>.
27. Asensio-Jurado, L.; Argilés, M.; Vinuela-Navarro, V.; Quedo-Junyent, L. Analysis of Patient Adherence to Emerging Treatment Tools for Improving Visual Functions in Amblyopia: A Systematic Review and Meta-Analysis. *J. Optom.* **2026**, *19* (2), 100591. <https://doi.org/10.1016/j.optom.2025.100591>.
28. Esteves, M. M.; MacNeil, B.; Taylor, G.; Glazebrook, C. M.; Johnson, M. G.; Passmore, S. R. Chronic Neck Pain Influence on Oculomotor Performance During Near Point Convergence and Fitts's Tasks: A Cross-Sectional Study. *J. Manipulative Physiol. Ther.* **2025**, *48* (1–5), 119–128. <https://doi.org/10.1016/j.jmpt.2025.08.005>.
29. Biscardi, M.; Grossinger, Z.; Colantonio, A.; Bayley, M.; Molayeva, T. Efficacy of Restitutive Interventions for Oculomotor Deficits in Adults with Mild Traumatic Brain Injury: A Systematic Review and Meta-Analysis. *Brain Inj.* **2024**, *38* (7), 499–513. <https://doi.org/10.1080/02699052.2024.2320163>.

■ Authors

André Kurahashi Dantas is a Brazilian high school student born in Campinas, São Paulo. With an aspiring goal to pursue medicine internationally, André has been actively participating in various extracurricular opportunities to strengthen his academic profile, such as olympiads, internships, research, and excelling in resilience, reflection, and analytical skills.

AI-Assisted Writing, Creativity, and Professional Identity

Andrew R. Sion

Poly Prep Country Day School, 1 Poly Prep Dr, Brooklyn, New York, 11228, USA; sionandrew27@gmail.com

ABSTRACT: Artificial Intelligence (AI) is rapidly changing writing-focused industries, yet its role in fields like academia and publishing remains debated. While AI supports writing in ways that improve authorship, creativity, efficiency, and the purpose of writing itself, it also raises important ethical concerns. Although AI tools offer strong advantages, they also complicate questions about originality and responsibility. As AI becomes a quiet co-author in many tasks, the purpose of writing is shifting: with more of the writing process automated, the human role centers increasingly on insight, authenticity, and critical thinking. This paper uses a qualitative review of current literature from multiple writing industries to examine these developments. It discusses the ethical and professional issues that emerge from AI adoption and considers how these changes may develop over the next decade. Overall, the paper argues that AI is creating a long-term hybrid authorship model, one that increases productivity but can sustain trust and creativity only if supported by transparency, human accountability, and fairness safeguards.

KEYWORDS: Artificial Intelligence, Authorship Ethics, AI-Assisted Writing, Authorship and Originality, Human-AI Collaboration.

■ Introduction

Throughout history, technology has steadily made writing easier and more accessible. From the printing press to the typewriter to digital word processors, each innovation reduced the time and effort needed to produce, edit, and share written work. Artificial Intelligence (AI) represents the newest step in this progression, offering tools that not only help with grammar and formatting but can also generate complete drafts of essays, reports, and stories. Yet unlike earlier tools, this shift forces a deeper reconsideration of what writing itself means, bringing major gains in efficiency and creativity while also raising broad ethical and professional concerns.^{1,2}

This paper argues that although AI greatly increases efficiency across writing-intensive fields, it also introduces ethical issues that are both urgent and far-reaching. Two problems stand out: the gradual weakening of human authorship and the limits of AI-detection systems. AI-detection systems are software tools, such as Turnitin, Copyleaks, and Originality.ai, that attempt to identify whether a piece of writing was produced by a human or generated with an AI model. The first issue challenges ideas about originality and responsibility, while the second shows a widening gap between fast-advancing generative models and the less reliable tools meant to detect them. Studies reveal that most detectors struggle to identify AI-generated material, especially from advanced models like GPT-4, which allows unacknowledged AI-written text to pass as original.³ Together, these challenges complicate the benefits of faster writing by putting the credibility of scholarship, journalism, and publishing at risk.^{1,2}

Recent research confirms that AI tools are already entering professional writing practices. One study shows that AI chatbots are used to outline arguments, refine wording, and create early drafts in scientific communication.⁴ Although the study focuses on medical writing, it reflects a trend that extends

across many writing-centered fields.⁴ Similar uses appear in most academic and professional environments, where AI lowers the barriers to producing polished text and speeds up the drafting process.^{1,5}

However, these changes are not only technical. Khogali and Mekid further show that AI's broader automation abilities intensify concerns about authorship and originality by reducing human involvement and increasing fears of declining agency and job loss.⁶ They argue that as AI takes over tasks long associated with human writers, such as crafting narratives, organizing arguments, or refining language, the human role shifts from creator to supervisor, weakening the sense of ownership tied to authorship.⁶ As AI takes over tasks such as crafting narratives, organizing arguments, and refining language, efficiency becomes inseparable from larger questions about ownership, responsibility, and communicative power.

The stakes of this transformation are significant because AI-assisted writing forces institutions to reconsider responsibility, authenticity, and human judgment in an era when machines can generate fluent text at scale. Addressing these challenges requires documenting how AI is reshaping writing today and anticipating the ethical issues that will expand as these technologies advance. Yet current academic literature still lacks broad frameworks that connect these writing changes across different industries. This study helps fill that gap by offering a more integrated view of how AI is redefining authorship, creativity, efficiency, and the purpose of writing across sectors.

■ Methods

This study used a qualitative literature review to explore how artificial intelligence (AI) is affecting writing-focused industries, with attention to both efficiency benefits and ethical implications. All sources were found through systematic

searches on Google Scholar. Searches used multiple keyword combinations to reduce the risk of missing relevant studies, including terms such as “AI,” “Writing,” “AI writing industry,” “AI and creativity,” “AI writing and ethics,” “AI and authorship,” “AI in academia,” “artificial intelligence and writing,” “generative AI and writing,” “machine learning writing and ethics,” “AI-generated writing,” “AI detection systems,” and “AI-assisted authorship.” To broaden the search, the term “AI” was also substituted with “artificial intelligence” across relevant searches to capture studies that used the full phrase rather than the abbreviation. Boolean operators were used when necessary to refine results and ensure relevance.

The selection process centered on peer-reviewed journal articles to maintain scholarly credibility. Most of the chosen sources were published within the last four years (2021–2025) to reflect current debates and developments, although earlier foundational works were included when they offered important historical context. Sources were included if they directly addressed at least one of the paper’s central themes: AI-assisted writing, authorship and originality, AI detection, creativity, professional identity, academic writing, publishing, or writing-related ethical concerns. Sources were excluded if they focused only on technical model architecture without discussing writing practices, ethics, authorship, or professional communication. Review and survey articles were also considered when they helped situate this paper within broader debates in the field.

Each source was evaluated for credibility using several criteria:

1. Reviewing author backgrounds, including their professional affiliations, expertise in AI, ethics, or writing studies, and prior scholarly work.
2. Evaluating the journals in which the articles were published, ensuring they were reputable, peer-reviewed, and relevant to technology, communication, ethics, education, publishing, or creative studies.
3. Comparing claims across multiple sources to confirm accuracy, consistency, and the avoidance of unsupported arguments.
4. Assessing the strength and relevance of the evidence by distinguishing between empirical studies, policy reviews, theoretical arguments, and opinion-based discussions. Greater weight was given to peer-reviewed empirical studies, systematic reviews, and journal policy analyses, while opinion papers were used mainly to frame debates or identify emerging concerns.

The collected literature was coded and analyzed to identify themes that appeared across writing-related industries. Coding was conducted through repeated reading of each source and note-taking on recurring claims, findings, and concerns. Initial codes included efficiency, productivity, authorship, originality, disclosure, plagiarism, AI detection, creativity, professional agency, bias, and equity. These codes were then grouped into broader themes. Three major themes guided the analysis: (1) efficiency improvements supported by AI, (2) ethical concerns surrounding authorship and originality, and (3) broader impacts on professional identity and agency. A

fourth cross-cutting area, AI detection and disclosure, was also tracked because several sources identified detection limits and transparency policies as central to debates about trustworthy AI-assisted writing. These themes were refined through repeated review and cross-comparison to build a comprehensive understanding of the topic. The purpose of this approach was to support the broader conceptual goals of the study by connecting evidence across academia, creative writing, publishing, professional communication, and related writing-intensive fields. These themes are summarized in Table 1.

■ Key Findings from the Literature

Table 1: Key Findings From the Literature on AI-assisted writing.

Theme	Main finding	Evidence base	Why it matters
Efficiency gains	AI reduces repetitive writing labor and speeds up drafting, summarizing, organizing, and reporting tasks.	Workplace adoption studies, professional writing research, medical reporting studies, and scientific writing examples.	AI is useful when it moves writers faster toward analysis, interpretation, and revision rather than replacing judgment.
Authorship, originality, and ethics	AI complicates responsibility, originality, plagiarism, disclosure, and ownership because it can shape text without being accountable for it.	Journal policy reviews, academic authorship studies, communication research, and ethical analyses.	Human authorship remains central because responsibility and accountability cannot be transferred to AI.
AI detection and disclosure	Detection systems are inconsistent, especially with advanced models, edited AI text, different genres, and non-native English writing.	Studies on AI-detection tools, false positives/negatives, model differences, and bias against non-native English writers.	Institutions should not rely only on detection software; disclosure and human review are more reliable safeguards.
Professional identity and agency	AI shifts writers from direct creators toward supervisors, editors, curators, and ethical decision-makers.	Research on creative writing, publishing, workplace automation, and professional communication.	The central issue is not just productivity, but whether human judgment, creativity, and ownership remain meaningful.

This table summarizes the main themes found across the reviewed literature: efficiency gains, authorship and ethics, AI detection, and professional identity. Overall, the findings show that AI improves productivity, but its benefits depend on preserving human judgment, transparency, and accountability.

Efficiency gains in AI-Assisted Writing:

AI-assisted writing clearly offers significant efficiency gains, with findings across different fields showing similar patterns. In academic research, large language models like ChatGPT can automate early research tasks such as scanning scholarly papers, summarizing long articles, and generating lists of research questions. These capabilities reduce the burden of early-stage writing and research tasks, allowing human writers to move more quickly toward analysis, interpretation, and revision. As shown in Table 1, efficiency gains are one of the major patterns identified across the reviewed literature.

Workplace studies show similar trends.⁷ In a survey of 300 employees from Korean companies, Kim, Blázquez, and Oh found that many workers had already begun incorporating generative AI tools into their daily routines.⁷ About 20.7% of employees reported using generative AI at work, and nearly half of these users relied on the free version of ChatGPT as their main tool.⁷ Participants also reported high levels of perceived performance efficiency (mean = 5.60) and competitive

advantage (mean = 5.31) when using AI for tasks like drafting, summarizing, and organizing content.⁷ Experimental evidence supports these self-reported workplace benefits: Noy and Zhang found that ChatGPT reduced the average time needed to complete mid-level professional writing tasks by 40% while improving output quality by 18%.⁸

The study also found that ease of use and social encouragement were the strongest predictors of AI adoption, showing that workers are more likely to use AI when it is simple to integrate and when colleagues support its use.⁷ Together, these findings illustrate how generative AI is reshaping workplace writing practices by speeding up early stages of drafting and allowing workers to shift more quickly toward deeper revision, analysis, and decision-making.⁷ This shift matters because it shows that AI's main value is not simply faster text production, but a redistribution of human labor away from routine drafting and toward judgment-based tasks. In this sense, efficiency becomes meaningful only when it strengthens, rather than replaces, the writer's role in analysis and decision-making.

The same pattern appears in high-stakes fields, where the consequences of AI-assisted writing are not only practical but also ethical. Medical examples are especially useful because they show how efficiency gains must be evaluated alongside accuracy, responsibility, and human oversight. Medical and clinical fields provide clear examples of similar benefits. In radiology, an AI-based chest X-ray triage system reduced the average turnaround time for critical results from 11.2 days to just 2.7 days.⁹ Another study found that AI-generated draft reports lowered average reporting time by 24%, from 573 seconds to 435 seconds, with no increase in significant errors.¹⁰ These results show that AI-driven drafting and triaging can improve efficiency not only in academic or business settings but also in high-stakes medical environments.

Scientific publishing and clinical research show related advantages. AI tools are already used to generate early manuscript drafts, organize literature, suggest structural improvements, and provide proofreading support. In critical care medicine, ChatGPT has demonstrated the ability to assist at each stage of manuscript development, helping streamline the writing process while leaving final interpretation to human experts.¹¹ These examples show that AI is most useful when it reduces repetitive writing labor without replacing expert judgment.

Authorship, Originality, and AI Ethics:

Although AI offers many efficiency benefits, the literature also highlights serious ethical concerns, especially regarding large language models such as ChatGPT in academic authorship. A policy review of 300 top journals found that while about 96.6% allowed AI tools to improve writing quality, 98.9% prohibited listing AI as a co-author.¹² This rule is based on the view that AI cannot take responsibility, give consent, or ensure the integrity of published research. Instead, most journals recommend acknowledging AI contributions in the methods or acknowledgments section, while a smaller number created specific AI disclosure categories.¹ These findings show that although publishers recognize AI's usefulness, they main-

tain that originality and accountability must remain human responsibilities.

Concerns about originality extend far beyond authorship rules. Cardon *et al.* found that communication instructors worry that AI-assisted writing weakens authenticity and credibility by producing polished text that may hide shallow reasoning or recycled ideas.¹³ Instructors noted that AI-generated writing can limit creativity and critical thinking, making it harder to evaluate whether a student truly understands the material.¹³ These risks increase because AI often produces generic, surface-level writing that may contain subtle mistakes requiring human oversight.¹³ Cardon *et al.* also emphasize that users cannot always explain how AI outputs are produced, raising long-term concerns about transparency, accuracy, and unintentional plagiarism.¹³

Detection challenges worsen the situation. In a 2023 study, Walters evaluated 16 AI-detection tools using essays written by ChatGPT-3.5, GPT-4, and human students, finding that while some tools, such as Copyleaks, TurnItIn, and Originality.ai, performed reasonably well, most struggled to identify text generated by GPT-4.³ This growing gap between AI writing capabilities and detection accuracy means institutions cannot rely on detectors alone, increasing the need for human oversight and transparent disclosure.

More recent studies complicate this issue by showing that detection accuracy depends on the tool, the model being tested, and the type of writing being evaluated. Weber-Wulff *et al.* found that many detection systems were inconsistent when distinguishing between human-written and AI-generated text, especially when AI outputs were edited or paraphrased.¹⁴ Elkhatat, Elsaid, and Almeer similarly found that several detectors performed better on GPT-3.5 text than GPT-4 text, suggesting that detection tools may become less reliable as generative models improve.¹⁵ Dalalah and Dalalah add that accuracy can vary by section of an academic paper, with literature review writing more likely than abstracts to be falsely flagged as AI-generated.¹⁶ Liang *et al.* extend these concerns by showing that GPT detectors can be biased against non-native English writers, with several tools falsely classifying non-native English writing as AI-generated more often than native English writing.¹⁷ These findings show that the problem is not only that detectors are imperfect, but that their reliability changes across models, writing genres, and revision methods. This instability changes the ethical problem from one of simple enforcement to one of governance. If detection tools cannot consistently identify AI-generated writing, then institutions must focus less on catching AI use after the fact and more on building transparent norms for disclosure, review, and accountability.

Lund *et al.* expand these concerns by analyzing issues related to authorship, copyright, and plagiarism. They explain that when AI generates text, ownership becomes difficult to define, and some scholars even suggest that developers or AI models might deserve credit in certain cases.² However, because AI training datasets are massive and untraceable, outputs may unintentionally reproduce copyrighted material, creating plagiarism risks and placing legal and ethical responsibility on

human authors.² Lund *et al.* further warn that AI systems can reinforce biases found in their training data, threatening fairness and the accuracy of scientific communication.²

Because simple bans on AI authorship do not fully explain how credit should work in hybrid writing, the literature increasingly points toward more flexible frameworks. These frameworks try to separate the practical use of AI from the ethical question of who deserves credit and who remains responsible. Some scholars propose more detailed authorship frameworks. Jenkins and Lin argue that authorship should depend on two factors: continuity (how much AI-generated content remains in the final work) and creditworthiness (whether the contribution would normally deserve authorship if made by a human).¹⁸ This approach is more nuanced than universal bans, but Walters's findings still show that detection tools are unreliable, making human judgment essential.³

Responsibility for errors is another major ethical concern because AI-generated text can contain inaccuracies, bias, or unsupported claims even when it appears polished. Policies reviewed by Lund and Naheem emphasize that human authors remain fully accountable for all content in a manuscript, including AI-generated text.¹ Dergaa *et al.* also stress that AI should support, not replace, human creativity and critical judgment.¹² A final concern involves fairness: researchers with greater access to advanced AI tools may gain advantages over those with limited resources, widening global inequalities in academic publishing.^{10,19}

Together, these findings show that although AI is increasingly integrated into academic writing, originality, responsibility, and authorship continue to depend on human judgment.

AI's Impact on Professional Identity and Agency:

AI-driven writing tools influence more than writing tasks; they are also reshaping the structure of professional identity and agency. In educational and workplace contexts, many worry that reliance on automated writing reduces critical thinking and weakens the authenticity of communication.¹³ Rather than simply replacing human writers, AI shifts part of the writing process toward evaluating, revising, and managing machine-generated text. This makes professional identity a question of agency: how much control, ownership, and judgment remain with the human writer?

One major aspect of this shift involves redefining creativity. Runco describes AI-generated content as a form of "artificial creativity," writing that it may appear original but lacks the intentionality and depth found in human creativity.²⁰ From this perspective, the writer's identity becomes associated with evaluating and shaping machine-generated material rather than generating ideas independently, raising questions about reduced human agency.²⁰

AI also alters human-technology relationships in creative settings. O'Toole and Horvát argue that AI disrupts traditional notions of authorship, expression, and intention, suggesting that AI-generated work may weaken the expressive authenticity central to human creativity.²¹ They also question whether AI should be seen as genuinely creative or simply as a tool that reshapes creative practices.²¹ These changes extend across

multiple industries. Dwivedi *et al.* show that although generative AI can significantly improve productivity in sectors like banking, education, and publishing, it also disrupts established practices and introduces ethical concerns related to privacy, bias, misuse, and misinformation.²² As a result, professionals may shift from producing content to overseeing AI output, managing risks, and ensuring ethical use.²²

Publishing policies reflect similar transformations. While authorship remains defined as a human responsibility, AI is increasingly becoming a standard collaborator through disclosure requirements.² Thus, professional agency becomes less about whether one uses AI and more about how its use is acknowledged and responsibly managed.²

Across writing industries, automation produces a tension between empowerment and displacement. On one hand, professionals benefit from being freed from repetitive tasks, creating more time for higher-level thinking, creativity, and decision-making. On the other hand, as AI becomes a default tool for efficiency, human contributions may be reduced to oversight rather than creation.²³ This change represents a deeper structural realignment of professional expectations.

A broader cultural perspective adds to this analysis. The dominant "creative industries" model often views creativity as an economic resource, emphasizing productivity and output over identity and labor. AI strengthens this trend by disconnecting creativity from human agency and linking it more closely to efficiency.²⁰ The concept of "everyday creativity" offers a counterpoint by highlighting individual and accessible forms of creativity, but without focusing on equity issues, this approach can overlook the structural challenges faced by those most at risk of being affected by automation.²³

Overall, these findings show that AI's structural impacts are defined by a tension between empowerment and displacement. While AI can free writers from repetitive tasks and create more time for higher-level thinking, it can also reduce visible human contribution when professionals become primarily responsible for supervising or refining automated output. The central concern is therefore not only whether AI improves writing efficiency, but whether creative ownership, fairness, and human judgment remain meaningful as writing becomes increasingly automated. This concern links the findings section to the paper's broader implications: AI's impact on writing cannot be measured only by productivity gains, because those gains also alter how writers understand their authority, creativity, and professional value. Table 2 organizes these broader implications across academic, creative, ethical, professional, and future research contexts.

■ Implications for Writing, Authorship, and Ethics

This table shows that AI's effects extend beyond faster drafting to larger questions about originality, disclosure, fairness, and professional trust. The main implication is that AI can be useful in writing-intensive fields only if institutions create clear standards for responsible use.

Table 2: Implications for AI-Assisted Writing Across Writing Fields.

Area of implication	What the review shows	Recommended response
Academic writing	AI can support scholarly drafting and research tasks, but it can also weaken originality, obscure reasoning, and create detection problems.	Academic institutions should require transparent disclosure, human verification, and writing practices that make author reasoning visible.
Creative writing	AI can expand access and support idea generation, but it may also reduce originality and push writing toward stylistic sameness.	Creative fields should treat AI as a support tool while preserving human voice, intentionality, and cultural diversity.
Authorship and ethics	AI cannot accept responsibility, even when it contributes to wording, structure, or argument development.	Authorship standards should keep accountability with human writers and clarify when and how AI assistance must be disclosed.
Professional acceptance	AI adoption is likely to grow, but trust will depend on governance, fairness, and transparency.	Writing professions should develop clear norms for disclosure, responsible use, and equitable access to advanced tools.
Future research	Existing literature documents major changes, but more empirical research is needed across writing-intensive fields.	Future studies should use interviews, surveys, workplace observation, and experiments to examine how AI changes writing practice over time.

AI’s Impact on Academic Writing:

AI’s influence on academic literature shows both acceleration and disruption. The reviewed studies suggest that AI’s value in scholarship depends on whether it supports, rather than substitutes for, the intellectual labor of synthesis and interpretation. Large language models can speed up early research tasks, but these gains matter only if human authors remain responsible for evaluating evidence, developing arguments, and preserving scholarly credibility. Kasneci *et al.* similarly argue that large language models can support educational writing by generating summaries, outlines, feedback, and personalized learning materials, but they emphasize that effective use requires critical thinking, fact-checking, awareness of bias, and continuous human oversight.²⁴ In academic writing, efficiency is therefore not automatically a benefit; it becomes valuable when it strengthens rather than weakens human judgment. Table 2 summarizes these implications by showing how AI-assisted writing affects academic, creative, ethical, professional, and research contexts.

AI-detection tools remain unreliable. This creates a policy problem rather than only a technical problem. Some perform moderately well, but many have difficulty identifying text produced by advanced models such as GPT-4.²² Because institutions cannot rely on detection software alone, the evidence points toward stronger disclosure norms, human review, and writing practices that make the author’s reasoning visible.

Authorship norms are also under pressure. Most high-ranking journals continue to prohibit listing AI as a co-author, emphasizing that responsibility, originality, and accountability must remain human obligations.^{2,12} At the same time, many journals now require authors to disclose AI use. Together, these policies suggest an emerging model of hybrid authorship: AI may contribute to language, structure, or drafting, but responsibility remains attached to the human author.

AI is also reshaping what it means to be an academic writer. Although AI outputs may appear creative, they lack the inten-

tionality and personal insight that define human scholarship.¹⁸ Scholars may increasingly shift from designing arguments to refining and curating machine-generated text. This transition aligns with trends in other fields, where professionals move from hands-on content creation to oversight, critical evaluation, and ethical decision-making.^{21,22}

In sum, AI speeds up scholarly work but also unsettles foundational concepts such as originality, authenticity, and professional identity. The challenge for academia is not simply deciding whether to permit AI, but developing frameworks that maintain human responsibility while recognizing the evolving nature of hybrid writing.

AI’s Impact on Creative Writing:

AI’s effects on creative literature echo those in academic writing but are, in some ways, even more pronounced. AI tools can generate drafts, imitate writing styles, and suggest new narrative paths.¹ Fang *et al.*’s systematic review of AI technologies used for story writing supports this point, showing that AI has already been applied to several parts of the creative writing process, including story generation, narrative support, and writing assistance.²⁵ Gero *et al.* add that creative writers’ willingness to use AI depends not only on the tool’s technical abilities, but also on the kind of support writers want, how they perceive computer-based assistance compared with peers or mentors, and the values they attach to the writing process.²⁶ These capabilities lower entry barriers and broaden participation in creative writing, allowing people who may lack certain skills or experience to produce high-quality text. However, the data-driven nature of AI means that much of its output reflects existing stylistic patterns, which may reinforce dominant cultural norms.⁶ This risks reducing originality and contributing to a more uniform literary landscape.

AI also challenges traditional understandings of creativity. Doshi and Hauser provide experimental evidence for this tension, finding that access to generative AI ideas made short stories seem more creative, better written, and more enjoyable, especially for less creative writers.²⁷ However, they also found that AI-assisted stories became more similar to one another, suggesting that AI may improve individual creative output while reducing the collective diversity of creative work.²⁷ Historically, creativity in literature is associated with original ideas and a unique authorial voice. AI complicates this model because it can generate fluent text by recombining existing stylistic patterns at high speed. While this can expand creative possibilities, it also risks narrowing originality if creative value becomes measured mainly by productivity.

Authenticity becomes another concern. Machine-generated writing may mimic human creativity but lacks personal experience and intentional decision-making.²⁰ This blurs the distinction between authentic and artificial expression and may reduce the perceived value of a writer’s individual perspective. As a result, AI-assisted creative writing raises a difficult question: whether the author remains the originator of meaning or becomes primarily the curator of machine-generated possibilities.

Similar structural shifts appear in academic writing: creative authors may be valued less for producing original content and more for their ability to revise, select, and guide machine-generated work.²² This change raises equity concerns, as access to powerful AI tools varies across regions and communities.^{1,2,12} Unequal access may widen gaps in who can publish, succeed, or profit in literary fields.

These concerns become more significant when creative writing is placed within the larger creative economy. AI not only affects individual writers' style or originality; it also changes professional workflows, expectations of expertise, and the cultural value attached to human creativity. Recent review work supports this broader shift in creative fields. Tsao *et al.*'s scoping review of 57 papers on generative AI in creative practices and industries finds that AI is changing creative labor across visual art and design, writing and literature, performing arts, and spatial design.²⁸ The study shows that creative professionals increasingly use AI for early-stage idea generation, curation, evaluation, and workflow support, while still relying on human judgment for final creative decisions.²⁸ It also identifies tensions around authenticity, expertise, professional identity, and the risk that AI may push creative work toward cliché or cultural homogenization.²⁸ However, because Tsao *et al.*'s framework focuses mainly on creative practice, the present study extends this discussion across writing-intensive sectors by connecting creative authorship to academic writing, publishing norms, professional communication, and AI-detection challenges.

Ultimately, AI forces a redefinition of authorship in creative literature. The key question becomes whether authorship is defined by generating new ideas or by skillfully managing AI-generated possibilities. The answer will shape whether AI enriches creative practice or contributes to homogenization and the commodification of creative expression.

Ethical Impacts on Authorship and Creativity:

The ethical questions surrounding AI's role in writing are among the most debated. The central problem is that AI can shape a text without being able to accept responsibility for it. The widespread refusal to treat AI as a co-author reflects the understanding that machines cannot take responsibility or make ethical decisions.^{2,12} When AI influences argumentation, phrasing, or structure, human writers must remain responsible not only for the final wording but also for the judgment, accuracy, and originality behind that wording.

The limits of detection technologies intensify these ethical dilemmas. Tools designed to identify AI-generated text often fail to reliably detect content produced by advanced models, especially GPT-4.²² If institutions cannot consistently detect AI-generated text, then academic integrity cannot depend on software alone. The stronger ethical response is to make AI use transparent and to evaluate whether the human author has provided genuine reasoning, interpretation, and accountability. Kreps, McCain, and Brundage extend this concern beyond academic writing by showing that AI-generated news can appear highly credible to readers. In three experiments on synthetic news and foreign-policy misinformation, they found that participants were often unable to distinguish between

human-written and AI-generated text, suggesting that weak detection and low transparency can also undermine public trust in journalism and democratic communication.²⁹

Hosseini, Resnik, and Holmes support this transparency-based approach by arguing that professional and scholarly communities should focus on disclosure rather than outright prohibition.³⁰ They explain that bans on AI writing tools are difficult to enforce and may encourage hidden use, while clear disclosure helps preserve trust, accountability, and attribution.³⁰ Their proposed guidelines include identifying how AI was used, citing the tool when appropriate, and keeping responsibility with the human author because AI cannot accept moral or legal accountability for the final text.³⁰ This argument shows that the ethical challenge is not simply whether writers should use AI, but how openly and responsibly they use it. This makes transparency one of the most practical solutions to AI-assisted writing because it accepts that AI will continue to be used while still protecting the human role in authorship. Disclosure allows writers to benefit from AI's efficiency without pretending that the work was created entirely without machine assistance. At the same time, it reinforces the idea that human judgment remains central, since the writer must still evaluate, revise, and take responsibility for the final product.

The opacity of AI training data adds another challenge. Because large language models are trained on massive and untraceable datasets, they may unintentionally reproduce copyrighted material or replicate bias.² This introduces risks of unintentional plagiarism, distorted attribution, and unfair representation. Some scholars suggest that instead of relying on rigid prohibitions, authorship should be guided by the amount of AI-generated content that remains in the final work and whether such contributions would normally deserve credit if created by a human.¹⁸ This framework attempts to balance ethical responsibility with the practical realities of hybrid authorship.

Authenticity is another central ethical concern. AI-generated text may appear original, but it lacks intentionality, emotional depth, and lived experience.²⁰ This blurring between artificial and genuine creativity can undermine the integrity of scholarly and literary expression.²¹ The responsibility falls on human authors to preserve originality and ensure that their work reflects genuine thought and interpretation.

Doyal *et al.* extend this concern to medical writing by explaining that ChatGPT and other AI tools can assist with drafting, editing, and organizing manuscripts, but they also create ethical risks when outputs contain inaccuracies, fabricated references, or unsupported claims.³¹ These risks are especially significant in medical and scientific writing, where errors can affect research quality, clinical understanding, and public trust.³¹ This shows that AI-generated writing cannot be accepted solely because it appears polished. Instead, human authors must remain responsible for verifying evidence, correcting errors, and ensuring that final interpretations are accurate and ethically sound.

Finally, ethical debates must include questions of equity. Access to powerful AI tools is uneven, giving advantages to well-resourced scholars and writers while leaving others at a

disadvantage.^{12,22} Those who can use AI effectively may gain a significant edge in speed, productivity, and visibility, while under-resourced communities risk exclusion from evolving writing landscapes.² In this way, AI reshapes authorship not only through creative and technical influences but also through the redistribution of communicative power.

AI Acceptance in Writing Professions:

Acceptance of AI in writing-focused professions seems likely, but the level and manner of acceptance will depend on governance and ethical standards rather than technological capabilities alone. This means that professional acceptance should be understood as a trust problem, not only an adoption problem. Writers may use AI because it is efficient, but institutions and audiences will only accept it if its use is transparent, accountable, and fair. Adoption rates across business settings already show increasing normalization: 20.7% of employees report using generative AI at work, and nearly half rely on ChatGPT for tasks such as drafting and organizing written content.⁷ Al-Bukhrani *et al.*'s study of academic researchers supports this pattern, finding that favorable attitudes and professional norms strongly shape intentions to adopt AI writing tools when users see clear benefits.³² Khalifa and Albadawy similarly argue that AI can support academic writing and research workflows, but that responsible acceptance depends on training, ethical transparency, and preserving the balance between AI assistance and human insight.³³ However, these adoption patterns show only that AI is becoming useful, not that it will automatically be trusted. Professional acceptance will depend on clear standards for disclosure, authorship, and responsible use.

Journal policies that permit AI assistance but prohibit AI co-authorship represent one form of compromise.^{1,12} Emerging authorship frameworks that evaluate continuity and creditworthiness suggest that future practices may become more flexible while preserving essential human accountability.¹⁸ Whether these approaches can maintain public and professional trust remains uncertain. If AI use is perceived as hidden or deceptive, confidence in journalism, academia, and publishing may decline. Conversely, if transparency and oversight become standard practice, AI can be integrated without undermining the value of human authorship.

Acceptance will also depend on how writers and scholars adapt to changing expectations of professional identity. As writing becomes more about curation, oversight, and interpretation rather than direct creation, some professionals may welcome the shift toward higher-level thinking. In contrast, others may perceive it as a loss of autonomy and authenticity.^{20,22} These changes may shape whether AI is seen as empowering or diminishing.

Equity remains a major factor. Uneven access to advanced AI systems may divide literary and scholarly communities, giving some writers significant advantages. In contrast, others fall behind.^{2,12,22} Acceptance will depend on addressing these disparities and ensuring fair access to AI tools. Cultural attitudes toward creativity, responsibility, and fairness will also shape how AI is received within writing professions.

Limitations and Future Research:

This review has several limitations that should be considered when interpreting its findings. First, because the study is a qualitative literature review, it synthesizes findings from existing studies rather than collecting original empirical data. Although the review includes empirical evidence from surveys, workplace research, AI-detection evaluations, and medical reporting studies, it does not conduct its own interviews, experiments, surveys, or statistical analysis. Second, although the review draws primarily on recent peer-reviewed scholarship, the search process relied mainly on Google Scholar. This may have limited the range of sources included, especially relevant industry reports, platform policies, or studies indexed in other academic databases. Third, the paper was completed within a four-month research period, which limited the amount of literature that could be reviewed in depth. This constraint is especially important because generative AI tools, detection systems, and institutional policies are developing quickly, meaning that some conclusions about authorship, disclosure, and AI detection may shift as new models and standards emerge.

Future research should build on these limitations by using longer-term and more systematic methods. Hadi *et al.*'s comprehensive survey of large language models reinforces this need by showing that future work must address not only expanding applications of LLMs, but also their technical limitations, reliability challenges, and unresolved risks across different professional domains.³⁴ Fütterer *et al.*'s study of global reactions to ChatGPT in education further supports this point by showing that public responses to AI tools are mixed, ranging from concerns about cheating and misinformation to interest in their potential as learning partners, which suggests that future research should also examine how institutional and public attitudes shape AI adoption in writing-related contexts.³⁵ Studies could use original interviews, surveys, workplace observation, or controlled experiments to examine how writers actually experience AI-assisted writing across academic, creative, medical, and professional contexts. Future work should also compare disclosure policies across institutions and test how effective they are in practice. Finally, researchers should continue studying whether AI changes professional identity, creativity, and access differently across writing-intensive fields, especially as advanced tools become more widely available.

■ Conclusion

AI is transforming writing across academia, creative industries, and professional communication by offering major gains in efficiency while introducing serious ethical and cultural challenges. Evidence across the reviewed literature shows that although AI reduces repetitive work and helps produce polished text, it also disrupts established ideas of authorship, originality, and creativity. The most important finding is not simply that AI makes writing faster, but that it changes what human contribution means within the writing process.

These changes bring both opportunities and risks. On one hand, AI can expand access to creative expression and free

writers to focus on deeper analysis, interpretation, and innovation. On the other hand, it can weaken authenticity, increase inequality, reproduce biases found in training data, and make it harder for institutions to determine when writing is genuinely human-authored. Because these issues appear across multiple writing-intensive fields, they cannot be understood through isolated discussions of academic writing, creative writing, publishing, or professional communication alone. This study addresses that gap by connecting evidence from these areas and showing that AI is redefining authorship, creativity, efficiency, and the purpose of writing across sectors.

Looking ahead, further research is needed to document how AI-assisted writing continues to develop and to anticipate the ethical challenges that will expand as generative models become more advanced. Future studies should examine the long-term effects of AI on writer identity, professional standards, detection and disclosure practices, and unequal access to AI tools. For AI to be accepted responsibly, institutions should strengthen disclosure policies, clarify authorship guidelines, and preserve human accountability as the foundation of trust in collaborative human-AI writing. By addressing these issues together, scholars and institutions can help ensure that AI strengthens rather than replaces the human dimensions of writing.

■ References

- Dergaa, I., Chamari, K., Zmijewski, P. and Saad, H.B., 2023. From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of sport*, 40(2), pp.615-622
- Lund, B.D., Wang, T., Mannuru, N.R., Nie, B., Shimray, S. and Wang, Z., 2023. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5), pp.570-581.
- Walters, W.H., 2023. The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, 7(1), p.20220158.
- Kitamura, F.C., 2023. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology*, 307(2), p.e230171.
- Anantrasirichai, N. and Bull, D., 2022. Artificial intelligence in the creative industries: a review. *Artificial Intelligence Review*, 55(1), pp.589-656.
- Khogali, H.O. and Mekid, S., 2023. The blended future of automation and AI: Examining some long-term societal and ethical impact features. *Technology in Society*, 73, p.102232.
- Kim, Y., Blazquez, V., & Oh, T. (2024). Determinants of generative AI system adoption and usage behavior in Korean companies: Applying the UTAUT model. *Behavioral Sciences*, 14(11), 1035.
- Noy, S. and Zhang, W., 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), pp.187-192.
- Najjar, R. (2023). Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17), 2760.
- Shen, Y., Heacock, L., Elias, J., Hentel, K.D., Reig, B., Shih, G., and Moy, L., 2023. ChatGPT and other large language models are double-edged swords. *Radiology*, 307(2), p.e230163.
- Salvagno, M., Taccone, F.S. and Gerli, A.G., 2023. Can artificial intelligence help for scientific writing?. *Critical care*, 27(1), p.75.
- Lund, B.D. and Naheem, K.T., 2024. Can ChatGPT be an author? A study of artificial intelligence authorship policies in top academic journals. *Learned Publishing*, 37(1), pp.13-21.
- Cardon, P., Fleischmann, C., Aritz, J., Logemann, M. and Heide-wald, J., 2023. The challenges and opportunities of AI-assisted writing: Developing AI literacy for the AI age. *Business and Professional Communication Quarterly*, 86(3), pp.257-295.
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P. and Waddington, L., 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1), pp.1-39.
- Elkhatat, A.M., Elsaid, K. and Almeer, S., 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1), p.17.
- Dalalah, D. and Dalalah, O.M., 2023. The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education*, 21(2), p.100822.
- Liang, W., Yuksekogonul, M., Mao, Y., Wu, E. and Zou, J., 2023. GPT detectors are biased against non-native English writers. *Patterns*, 4(7).
- Jenkins, R. and Lin, P., 2023. AI-Assisted authorship: how to assign credit in synthetic scholarship. Available at SSRN: 4342909.
- Lee, H.K., 2022. Rethinking creativity: Creative industries, AI and everyday creativity. *Media, Culture & Society*, 44(3), pp.601-612.
- Runco, M.A., 2023. AI can only produce artificial creativity. *Journal of Creativity*, 33(3), p.100063.
- O'Toole, K. and Horvát, E.Á., 2024. Extending human creativity with AI. *Journal of Creativity*, 34(2), p.100080.
- Dwivedi, Yogesh K., Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah et al. "Opinion Paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy." *International journal of information management* 71 (2023): 102642.
- Ooi, K.B., Tan, G.W.H., Al-Emran, M., Al-Sharafi, M.A., Capatina, A., Chakraborty, A., Dwivedi, Y.K., Huang, T.L., Kar, A.K., Lee, V.H., and Loh, X.M., 2025. The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems*, 65(1), pp.76-107.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E. and Krusche, S., 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, p.102274.
- Fang, X., Ng, D. T. K., Leung, J. K. L., & Chu, S. K. W. (2023). A systematic review of artificial intelligence technologies used for story writing. *Education and Information Technologies*, 28(11), 14361-14397.
- Gero, K.I., Long, T., and Chilton, L.B., 2023, April. Social dynamics of AI support in creative writing. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1-15).
- Doshi, A.R. and Hauser, O.P., 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28), p.eadn5290.
- Tsao, J., Liang, C.X., Nogues, C. and Wong, A., 2025. Perceptions and integration of generative artificial intelligence in creative practices and industries: a scoping review and conceptual model. *AI & SOCIETY*, pp.1-20.
- Kreps, S., McCain, R.M. and Brundage, M., 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1), pp.104-117.

30. Hosseini, M., Resnik, D.B. and Holmes, K., 2023. The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics*, 19(4), pp.449-465.
31. Doyal, A.S., Sender, D., Nanda, M. and Serrano, R.A., 2023. ChatGPT and artificial intelligence in medical writing: concerns and ethical considerations. *Cureus*, 15(8).
32. Al-Bukhrani, M.A., Alrefae, Y.M.H. and Tawfik, M., 2025. Adoption of AI writing tools among academic researchers: A Theory of Reasoned Action approach. *PloS one*, 20(1), p.e0313837.
33. Khalifa, M. and Albadawy, M., 2024. Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer methods and programs in biomedicine update*, 5, p.100145.
34. Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., and Mirjalili, S., 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.
35. Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: global reactions to AI innovations. *Scientific reports*, 13(1), 15310.

■ Author

Andrew R. Sion is a student at Poly Prep Country Day School in Brooklyn, New York, with academic interests in artificial intelligence, ethics, and the social impacts of emerging technologies. His work focuses on how AI is reshaping creativity, authorship, and professional identity across academic and creative fields. He is particularly interested in the ethical challenges surrounding human-AI collaboration and the future of writing in an automated age.

Meaning and Boundaries: How Grammarly and QuillBot Preserve or Interpret Ambiguity in Native English

Anya S. Yarlagadda

Sharon High School, 181 Pond St, Sharon, MA, 02067, USA; anyasaisha@gmail.com

ABSTRACT: Many people use Grammarly and QuillBot, including their paraphrasing tools. They are excellent at their job, but they still face difficulties when they are presented with ambiguous phrases and sentences. This study is crucial because ambiguity in communication is harmful, and people tend to rely on AI too much, so identifying patterns that could misconstrue a person's writing will be beneficial for understanding between English speakers. While Grammarly tends to preserve pronoun ambiguity when it is very unclear which potential meaning is correct, QuillBot assigns the ambiguous pronoun to the subject of the sentence, even if there is no suggestion that it would be more likely. The researcher created ambiguous sentences to represent native English that focused on four types of ambiguity, and they were tested using the paraphrasing tools for Grammarly and QuillBot. The results found that both applications showed a pattern of selecting a disambiguation if one interpretation was found more often than another one. With these findings, it can encourage future research in specific types of ambiguity and disambiguation in AI paraphrasing, helping to locate and mitigate bias in communication.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, Grammarly, QuillBot, Ambiguity.

■ Introduction

AI grammar and writing tools like Grammarly and QuillBot are used by many people from various professions and backgrounds because of their skills in enhancing written communication. They can improve the quality of writing and help users save time on small issues.¹ While these tools are found to be very advanced, with capabilities of decreasing ambiguity and vagueness, they are not always as accurate with these variables compared to simple grammar checking.² While there is current research being done in the field of computational linguistics, specifically focusing on AI grammar tools such as Grammarly and QuillBot, there are areas that have not been thoroughly studied or discussed yet.

One of the main areas of study is investigating the impact of AI grammar and writing tools on students in academic settings who do not speak English as a first language. A study of undergraduate students with English as a foreign language found that Grammarly gave students useful feedback on their work, and QuillBot was a great tool to practice paraphrasing.³ A different study involving undergraduate students found that the English learners preferred QuillBot over Grammarly because its paraphrasing tool was more useful for them than Grammarly's grammar correction tool, even though both were considered helpful.⁴ While these insights are important to add to the current body of knowledge, they do not look at native English speakers, who are one of the main groups that use these tools.

Many studies on Grammarly and QuillBot and their functions do not explicitly focus on ambiguity, and if mentioned, it is often as a small part of a larger observation. A paper that discusses the drawbacks of QuillBot and paraphrasing tools in general identifies an instance of an ambiguous prepositional phrase, but it is not the focus of the study because the re-

search looks at many different factors.⁵ A study interested in how Grammarly interacts with academic literature in English found that the tool could indicate instances of ambiguity, but the paper does not test paraphrasing or go into more depth about ambiguity.⁶ This creates a gap in research that this study aims to fill.

While the study conducted does not have a large enough dataset to make definitive statements, it is an important pilot for further research, as it has a different focus from many existing studies. Most literature that considers AI grammar tools like Grammarly or QuillBot investigates their impact on writers in an academic setting or non-native English learners.^{3,7} However, this study focuses on native and conversational English by constructing sentences with an informal tone and common words. There are very few academic papers that directly compare Grammarly and QuillBot for native English speakers, and the most common articles are blog reviews that look more at personal and overall experience than specific variables.⁸ Meanwhile, this study was performed with high standards to test specific, isolated variables, and with a detailed explanation of the testing process to ensure reproducibility. In addition, many studies do not focus on ambiguity or ambiguity preservation when conducting research on paraphrasing tools, and most existing work is only theoretical.⁹ This study specifically measures ambiguity preservation and the meaning preserved when applicable, adding a different angle of study in an established field.

Ambiguity is known to confuse the people receiving information. This can lead to misunderstandings that might be small or very large. Therefore, it hinders effective communication and makes the spread of ideas more difficult.¹⁰ Users of AI tools are likely to become overreliant and accept suggestions without thinking for themselves, which damages growth in

critical thinking, individual style, creativity, and writing skills.¹¹ This combination can be dangerous for people trying to understand each other in a time when AI usage is only increasing. People may not verify what an AI tool wrote, so if they had an ambiguous sentence that was disambiguated in a different way than what the original user had in mind, it could cause communication gaps and inefficiencies.

With this study as the starting point for future research, people could be more aware and educated about the tendencies and biases of AI grammar tools when paraphrasing, encouraging critical thinking when using them. Identifying the patterns that AI models are trained on can help in creating clear distinctions between what the tool thinks a person is trying to say and what they actually mean. It can also help anyone developing their own model by demonstrating things to take into account that can improve technology and innovation in the field of AI grammar tools.

■ Methods

The objective of this study is to determine trends of Grammarly and QuillBot when preserving or removing certain types of ambiguity from ambiguous sentences with a conversational tone. This paper describes a small-scale in-field study using online tools. This includes Google Sheets to keep track of the data, the Grammarly website to test sentences, and the QuillBot website for the same purpose as Grammarly.

The researcher formed forty sentences to test four different types of ambiguity with ten sentences each: lexical ambiguity, structural ambiguity, pronoun ambiguity, and punctuation ambiguity. The sentences were created to have a conversational tone to mimic the style of native English. The sentences were copied one at a time into the free versions of Grammarly's paraphrasing tool and QuillBot's paraphrasing tool that do not require an account, and the result was recorded verbatim. Grammarly did not present customization for its free, no sign-up required tool, but QuillBot was configured to standard mode using American English with fewer synonym changes. This was to try to decrease anything that could affect the results of the sentences besides the training of the models. After the results were recorded, the data were analyzed by first considering whether the original ambiguity of the sentence was preserved. If ambiguity was not preserved, the meaning that the model tended toward was recorded based on the two disambiguation options that were considered before testing. Punctuation ambiguity was tested using two different forms of each of the ten sentences, only changing the punctuation, meaning twenty sentences were tested for punctuation ambiguity in total. However, each sentence pair was considered one data point in graphs, with specific differences being noted later in the results section.

This study was conducted using online tools only, so no physical tools or research instruments were utilized. Ethical considerations that should be observed include following the terms of use for both Grammarly and QuillBot by not automating the collection of data or testing in bulk. All sentences used for testing were created by the researcher for this study and do not contain private information. Testing was performed

on the same day to ensure no disparities between potential updates, and the specific version of each tool was documented. Results were evaluated from a linguistic standpoint, and the results were not attributed to the inherent intelligence of an AI application; instead, the focus was on model behavior. Grammarly and QuillBot are third-party services, so it is possible for output to differ from the results of this study in the case of future updates.

Common AI Grammar Tools:

AI tools for writing were not always at the high level that they are today. While they initially only checked for basic grammar and spelling mistakes, they have improved greatly. AI tools use machine learning and natural language processing (NLP) to evaluate complex tone, style, and varying situations.¹² QuillBot is very good at paraphrasing, a skill that can even help native speakers who need to improve. It helps the user understand what they need to do, like changing sentence structure and using synonyms while still keeping the intent of the original passage.¹³ Grammarly is most helpful in error correction. It helps the user avoid mistakes in grammar and punctuation while also looking more widely at sentence structure.¹⁴

Table 1: Comparison of common AI grammar tools, specifically Grammarly and QuillBot, highlights QuillBot's specialized focus on paraphrasing capabilities.

Feature	Grammarly	QuillBot
Grammar Checking	Strong	Good
Paraphrasing	Okay	Strong
Plagiarism Detection	Good	Good
AI Generation	Good	Okay
Integration	Strong	Good

Grammarly and QuillBot offer many similar functions, but they each have different strengths. Comparative research finds that Grammarly specializes in grammar accuracy and mechanics, while QuillBot focuses on the readability of sentences.¹⁵ Along with the baseline feature comparison in Table 1 that QuillBot is slightly less accurate than Grammarly in grammar checking and that Grammarly does not have a paraphrasing model as high-quality as QuillBot, there are other tools that the platforms possess, even if they are not as well-known. Both tools have mostly error-free plagiarism detection tools. However, Grammarly is able to perform better when generating written content using AI with a very specific prompt, and it has better integration across many other services.

It is also important to consider that there are some features that the applications do not share, especially those specific to their target audience. Grammarly can provide specific suggestions based on the tone or style that someone is aiming for, including style guides and a brand tone for businesses that want to keep communication consistent. QuillBot can summarize long text to highlight the main points of a message without having to read it all. It can also be used in over 45 languages,¹⁶ making it accessible to non-English or multilingual speakers in different places around the world.

Different Types of Ambiguity:

Linguistic ambiguity can be found very often in the English language, and it arises whenever a phrase or a sentence can be understood in more than one way. While there are wide varieties, such as pragmatic ambiguity, which focuses on the context surrounding a statement, or vocal ambiguity, which relies on analyzing the way in which a language is spoken verbally,¹⁷ this study only focuses on certain types of ambiguity. They are all concerning written language and do not need to be extrapolated outside of the context of an individual sentence. The following types are lexical ambiguity, structural ambiguity, pronoun ambiguity, and punctuation ambiguity.

Lexical ambiguity is a type of ambiguity when a single word has more than one meaning. While ambiguity can make communication more difficult, it is hard to eliminate from languages because lexical ambiguity is found very frequently, and it is hard to avoid. Examples of lexical ambiguity are also known as homonyms. As there is normally one definition of the word that is used more often, it is called the dominant meaning, and the other, less-used meanings are called subordinate meanings. To understand which definition a person is referring to in a lexically ambiguous phrase or sentence, people consider both the context and how often each meaning is used. An example of this is the sentence, “The bat flew through the air.”¹⁸ If the sentence were used in a conversation about sports, it could be interpreted one way, but if there was not as much context, someone might find it more logical for the animal to be flying instead of the object.

Structural ambiguity occurs not in individual words, but because there are multiple ways to think of the sentence based on how it is worded. For example, the sentence, “I saw the man with the telescope,” can be viewed in multiple ways.¹⁹ One person might think it means that the subject of the sentence saw a man while they themselves were using a telescope, but someone else might think the man was using the telescope and the subject saw that.

Pronoun ambiguity is caused when there is difficulty understanding the antecedent of a pronoun, or the noun that the pronoun is referring to. People often try to interpret pronoun ambiguity by thinking about which noun acting as the pronoun would lead to a more logical cause-and-effect relationship. With the statement, “Albert invited Ron to dinner. He spent hours cleaning the house,” there are two ways an individual could view it. They would either think Albert cleaned the house or Ron did. However, it is more likely that Albert would have to clean the house to prepare for Ron’s arrival as a guest, so that it would be the more common disambiguation.²⁰

Punctuation ambiguity arises when punctuation or the lack thereof can alter or shift the meaning of a sentence. For example, the sentences “John, said Mary, was the nicest boy at the party” and “John said Mary was the nicest boy at the party” can be compared. Only changing the usage of commas, the sentence that originally meant Mary was talking ended up with John being the speaker. This is because commas affect the way people group words to form meaning, also known as syntactic parsing.²¹

Table 2: Summary of lexical, structural, pronoun, and punctuation ambiguity with an explanation of why it causes confusion and an example sentence, highlighting individual differences.

Type of Ambiguity	Reason for Ambiguity	Example Sentence
Lexical	Word with multiple meanings	The bat flew through the air.
Structural	Structure of the sentence	I saw the man with the telescope.
Pronoun	Antecedent, or defining noun, for the pronoun is unclear	Albert invited Ron to dinner. He spent hours cleaning the house.
Punctuation	Punctuation changes the grouping of words, therefore changing the meaning	John, said Mary, was the nicest boy at the party vs. John said Mary was the nicest boy at the party

To summarize the different types of ambiguity discussed in Table 2, all types of ambiguity studied in this research are focused on written English without requiring any outside context. Lexical ambiguity occurs when a single word can have multiple meanings.¹⁸ Structural ambiguity is found when the structure of a sentence does not indicate a single interpretation.¹⁹ Pronoun ambiguity arises when a pronoun in a sentence could refer to one of multiple nouns mentioned before.²⁰ Punctuation ambiguity occurs when the presence or lack of punctuation changes the way words are grouped.²¹

■ Results and Discussion

Ambiguity Preservation:

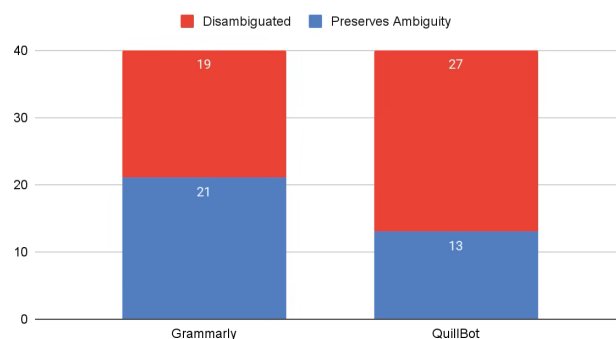


Figure 1: Sentences and sentence pairs vs. tool disambiguation overall. The Y-axis represents the total number of tested sentences, with the stacked bars displaying the relative proportion of instances in which ambiguity was preserved or resolved. This reveals Grammarly’s higher probability of preserving ambiguity in ambiguous sentences. These results signify that Grammarly may be more likely to leave meaning intact, while QuillBot may have a higher chance of defaulting to a single interpretation, possibly changing a user’s intended message.

To evaluate these data points, it is important to acknowledge that Grammarly primarily focuses on grammatical correctness as a mechanics-focused grammar checker, while QuillBot prioritizes readability through paraphrasing suggestions using NLP.¹⁵ Given the exploratory nature of this study, the data set of sentences tested is too small to evaluate apparent trends and percentage differences as generalizations or conclusive findings. Overall, as shown in Figure 1, Grammarly had a higher rate of ambiguity preservation at 52.5%, while the ambiguity preservation rate for QuillBot was only 32.5%. This means that almost half of the time for Grammarly and more than two-thirds of the time for QuillBot, the tool resolved the sentence toward one potential meaning of a sentence over the other. One possible explanation is that QuillBot tended to change the sentence structure instead of only substituting

words for synonyms, more so than Grammarly.¹⁵ This would increase the chance of removing ambiguity and resolving the ambiguity toward a single disambiguation.

The following subsections begin with an analysis of the results by looking at one type of ambiguity from a general perspective, and then investigating specific cases. The order presented is the same as the order tested: lexical ambiguity, structural ambiguity, pronoun ambiguity, and then punctuation ambiguity.

Lexical Ambiguity:

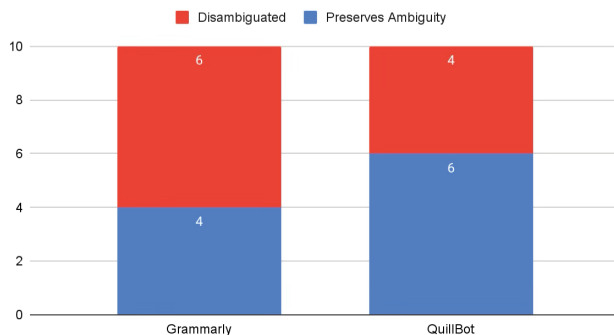


Figure 2: Sentences vs. tool disambiguation in lexical ambiguity. Similar to Figure 1, the Y-axis reflects the total sentence count tested, broken down by the proportion of ambiguous and disambiguation sentences. There was a slightly higher chance of QuillBot preserving ambiguity for a sentence compared to Grammarly. This indicates that, in contrast to the overall findings, QuillBot may be more likely to preserve the intent of a user's message, while Grammarly might select a certain meaning over another at a higher frequency.

According to Figure 2, the rate of ambiguity preservation for Grammarly concerning lexical ambiguity was 40%, while it was 60% for QuillBot. This means the difference between the two is not very large, but it does contradict the overall trend of Grammarly preserving the ambiguity of more sentences. Because lexical ambiguity is dependent on specific words being ambiguous, more words being substituted would likely lead to disambiguation. This occurred with Grammarly inputting a few more unique words than QuillBot, which is interesting because it does not align with current research that finds that QuillBot has a higher lexical diversity when paraphrasing.²² Out of the ten sentences tested, only two of them had ambiguity preserved by both tools, but when both tools disambiguated them, the selection of interpretation was the same, and this happened twice. For example, with the input sentence "I found a pen near the farm," Grammarly paraphrased it as "I discovered a pen close to the farm," while QuillBot wrote, "I discovered a pen close to the property." The word pen is ambiguous because it could be understood as a writing utensil, but it would also be thought of as a holding space for animals, especially when near a farm. However, neither application specified what kind of pen the sentence meant. An example of this not happening for either tool is with the sentence "He offered her a date." A date can be a social appointment, often romantic, or it can be a chewy fruit coming from palm trees. Grammarly very explicitly had a tendency to the first disambiguation option with the output, "He asked her out on a date," while QuillBot wrote, "He proposed a date to her," a more subtle output but still only referring to the first possibility. This implies that both models

believe the first meaning is more common than the second, to the extent that it is acceptable to assume what the person inputting the text meant.

Structural Ambiguity:

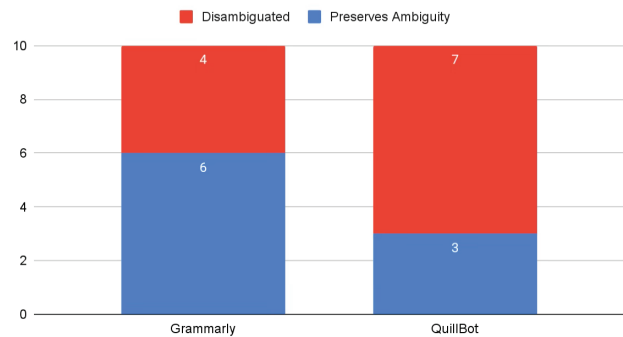


Figure 3: Sentences vs. tool disambiguation in structural ambiguity. Grammarly was more likely to preserve ambiguity in ambiguous sentences.

Figure 3 demonstrates that the rate of ambiguity preservation for Grammarly concerning structural ambiguity was 60%, while it was 30% for QuillBot. This follows the overall trend of Grammarly preserving ambiguity in more sentences than QuillBot. There was only one instance where the paraphrased sentence remained ambiguous for both tools, and just like the results for lexical ambiguity, the two cases with disambiguation for both tools yielded the same result. The only two sentences that both tools did not find ambiguous were "She plays the piano with a red bow" and "The fish is ready to eat." Compared to more confusing sentences like "I invited friends for a party on Friday," the disambiguated sentences both have one more likely meaning. Having a party on Friday is just as likely as sending invitations on that day of the week. Still, a woman wearing a red bow is more commonly seen than a piano with a red bow, and people are more likely to talk about a fish as food than a creature waiting to be fed due to the wording. The one sentence that remained ambiguous was "The parents said their children are stubborn and loud." Both tools only replaced some of the original words with synonyms, but neither tried to restructure the sentence.

Pronoun Ambiguity:

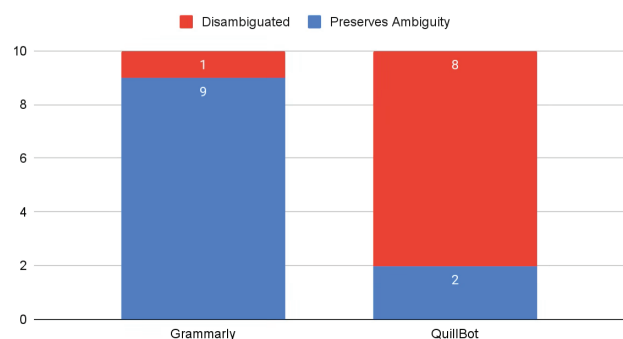


Figure 4: Sentences vs. tool disambiguation in pronoun ambiguity. Grammarly had a much higher chance of preserving ambiguity in sentences compared to QuillBot.

The rate of ambiguity preservation for Grammarly concerning pronoun ambiguity was 90%, while it was 20% for QuillBot, as highlighted in Figure 4. This is a very large difference and suggests that there might be a larger pattern that will persist even with a larger dataset. This could be due to QuillBot's algorithmic tendency to change the sentence structure, forcing it to pick one option.¹⁵ Another important thing to note is that out of the ten sentences, eight used specific names for people, one referred to two objects, while the last sentence involved animals, and it was the only sentence with pronoun ambiguity that Grammarly disambiguated. The sentence was "The puppy chased the squirrel, and it ran up a tree," while Grammarly paraphrased it as "The puppy pursued the squirrel, which scampered up a tree." When there are names, there is no way to predict which meaning is more likely, but in this example, many people would assume the squirrel ran up the tree, which could be why Grammarly also had a tendency toward that meaning.

QuillBot only preserved ambiguity for two sentences, and both of them share something in common. They were the only sentences to start with two names joined by the conjunction "and," which was then followed by an unclear pronoun. This could be because both names were the subjects in their sentence instead of having one subject and one object. Looking at the seven other sentences besides the dog and squirrel example, QuillBot consistently followed a pattern of selecting the subject of the sentence or the person performing the verb to be the one representing the ambiguous pronoun. This formula would not work for the two sentences, as it remains ambiguous because there is no reason to prefer one person over the other.

Punctuation Ambiguity:

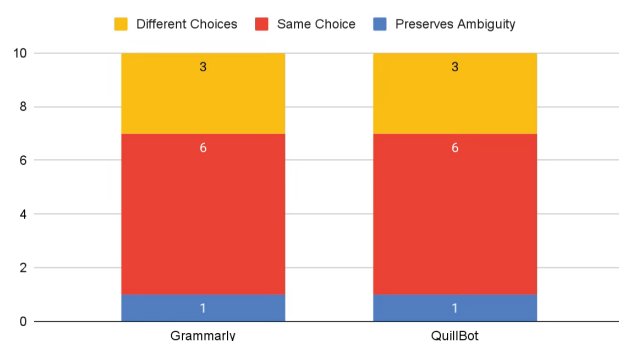


Figure 5: Sentence pairs vs. tool disambiguation in punctuation ambiguity. Grammarly and QuillBot had equal rates of ambiguity preservation and disambiguation.

The rate of ambiguity preservation for both Grammarly and QuillBot concerning punctuation ambiguity in Figure 5 was 10%. This is likely because the nature and testing of punctuation ambiguity is slightly different from other types of ambiguity. Because the interpretation could rely on the presence or absence of punctuation without changing the words, the analysis is a comparison of the two forms of the sentence, as well as Grammarly's and QuillBot's responses to them. For both tools, the sentences that remained ambiguous were from the same sentence pair, meaning all of the other sentence pairs can be evaluated based on their disambiguation.

Figure 5 demonstrates that Grammarly had three of the remaining nine sentences change their disambiguation when the punctuation was altered. Meanwhile, QuillBot also changed the disambiguation three times, but there was only one sentence that they both interpreted one way, and then changed once the punctuation changed. There were four sentences that both tools agreed on completely, defaulting to the same disambiguation option before and after the punctuation change, and selecting the same way the other platform did.

QuillBot was more likely to paraphrase both forms of the sentence the same way, with this occurring three times, while it only happened once for Grammarly. This suggests that even when using the paraphrasing tool that is QuillBot's strength, Grammarly might be more sensitive and aware of punctuation, and it has more variety when entering a similar sentence multiple times.

Synthesis:

Across multiple testing categories, Grammarly has a tendency towards preserving ambiguity, while QuillBot systematically operates as an ambiguity resolver. While the data for structural ambiguity, pronoun ambiguity, and overall performance confirm this, lexical ambiguity and punctuation ambiguity are exceptions. This pattern, especially found in pronoun ambiguity, where Grammarly preserved ambiguity in 90% of sentences tested compared to 20% for QuillBot, can be understood by each tool's main function. Grammarly is designed to focus on grammar and mechanics, and the tool's algorithm prioritizes structural and syntactic correctness, while QuillBot uses NLP to heavily paraphrase, forcing textual transformation.¹⁵ Grammarly has a tendency not to affect the framework of the sentence and instead substitutes synonyms, but QuillBot restructures the entire sentence. QuillBot also relies on structural formulas, such as defaulting to the primary subject of the sentence or assuming the most statistically common cultural interpretation, causing alternative meanings to not be preserved. Because both tools display a pattern of disambiguating sentences without extra context, it is important for writers not to blindly trust these systems. Grammarly is a safer tool to maintain nuanced writing with multiple meanings, while QuillBot's restructuring capabilities mean writers should be more cautious and post-edit to ensure the intended message is not changed.

Conclusion

Grammarly and QuillBot both seem to evaluate the probability of a certain meaning to be used when determining whether to preserve ambiguity or disambiguate sentences using their paraphrasing tools. While all four types of ambiguity showed interesting results, the clearest patterns emerged for pronoun ambiguity, demonstrating Grammarly's inability to disambiguate when there is no one more likely answer, and QuillBot's tendency to assign ambiguous pronouns to the subject of the sentence, even if there is no indication that only that meaning is possible. This is important to consider when using the tools to make sure that they do not change the understanding of a person's original message.

However, when evaluating these implications, several structural limitations must be considered. First, this study was designed to be a small-scale pilot project, resulting in a small dataset of 40 sentences tested overall. Because the built-in data models of the tools default to the most statistically probable meaning when they disambiguate a sentence, it can introduce bias in the ambiguous text toward what is most common or acceptable. These boundaries necessitate further scientific exploration to limit bias.

There are many possibilities for future research that expand on this study. While this study looked broadly at four types of ambiguity, there are even more specific categories within each group that could be analyzed to draw more concrete and generalized conclusions. It is also important to test these using a large dataset. Another potential variable to look at is how different modes for both Grammarly and QuillBot could affect the ambiguity of outputs, as they have models that focus specifically on increasing fluency or sounding a certain way.

■ Acknowledgments

I would like to thank Professor Kay Diaz, Dr. Monica Sava, and Professor Virgel Torremocha for their help throughout the research process.

■ References

1. Anis. Teachers' and Students' Perception about the Effects of Using ChatGPT, Grammarly, and QuillBot on Students' Writing Skills. *Centre-univ-mila.dz* **2024**. <http://dspace.centre-univ-mila.dz/jspui/handle/123456789/3481>.
2. Kumar Adimalla, P.; Kumari, P.; Ramagani, T.; Dasharatham, P. *CONTROLLING of AMBIGUITY and VAGUENESS through AI TOOLS in WRITING RESEARCH ARTICLES*; 2024; pp. 2582–5208. https://www.irjmet.com/uploadedfiles/paper/issue_9_september_2024/61576/final/fin_irjmet1726567205.pdf (accessed 2025-11-14).
3. Titik Ismailia; Adriadi Novawan. Enhancing EFL Students' Recount Writing: A Comparative Analysis of Grammarly, QuillBot, and GrammarCheck Applications. *Journal of English in Academic and Professional Communication* **2025**, 11 (2), 89–102. <https://doi.org/10.25047/jeapco.v11i2.5095>.
4. Ade Risya Fitri; Utami Dewi. UNDERGRADUATE STUDENTS' PREFERENCES and PERCEPTIONS on the USE of QUILLBOT and GRAMMARLY as PARAPHASING TOOLS. *Indonesian EFL Journal* **2024**, 10 (2), 167–174. <https://doi.org/10.25134/ieffj.v10i2.9868>.
5. Mehmood, A.; Niaz, F.; Iqbal, A. Potential Drawbacks of Automated Paraphrasing Tools: Performance of QuillBot in Paraphrasing of Headline Discourse. *PAKISTAN LANGUAGES AND HUMANITIES REVIEW* **2024**, 8 (III). [https://doi.org/10.47205/plhr.2024\(8-iii\)34](https://doi.org/10.47205/plhr.2024(8-iii)34).
6. Abdallah Abu Qub'a; Nour, M.; Shehdeh Fareh. Exploring the Use of Grammarly in Assessing English Academic Writing. *Heliyon* **2024**, 10 (15), e34893–e34893. <https://doi.org/10.1016/j.heliyon.2024.e34893>.
7. Karunanayaka, G.; Weerasinghe, A. *Usage of Artificial Intelligence (AI) Tools for Academic Activities by Undergraduates*; 2025.
8. Shaikh, E. *Grammarly vs Quillbot: A Detailed Comparison*. Demand-Sage. <https://www.demandsage.com/grammarly-vs-quillbot/>.
9. Hausser, R. *Paraphrase and Ambiguity*. ResearchGate. https://www.researchgate.net/publication/361228802_Paraphrase_and_Ambiguity (accessed 2025-11-14).
10. Demir, C. Lexical and Structural Ambiguities in Student Writing: An Assessment and Evaluation of Results. *African Educational Research Journal* **2020**, 8 (3). <https://doi.org/10.30918/AERJ.8S3.20.077>.
11. Budiyo, H. Exploring Long-Term Impact of AI Writing Tools on Independent Writing Skills: A Case Study of Indonesian Language Education Students. *International Journal of Information and Education Technology* **2025**, 15 (5), 1003–1013. <https://doi.org/10.18178/ijiet.2025.15.5.2306>.
12. Subramanian, R.; Vidalis, S. M.; Najafi, F. T. *Writing skills can be improved using AI tools: An Analysis*. Doi.org. <https://doi.org/10.18260/1-2--49460> (accessed 2025-11-14).
13. None Yoandita; None Yenni Hasnah. QUILLBOT as an ALTERNATIVE WRITING TOOL: EXAMINING ITS USES on the ACADEMIC WRITING PERFORMANCE of EFL LEARNERS. *Esteem Journal of English Education Study Programme* **2024**, 7 (2), 401–412. <https://doi.org/10.31851/esteem.v7i2.15254>.
14. Resiana, A. T.; Zamzam, A.; Putera, L. J.; Amrullah; Arfah, H. Effectiveness of Grammarly Application on the Students' Argumentative Writing Progress. *Journal of English Education Forum (JEEF)* **2024**, 4 (3), 153–159. <https://doi.org/10.29303/jeef.v4i3.722>.
15. Misiūnas, A. *Quillbot vs Grammarly: Which Is Better in 2025?* Cybernews. <https://cybernews.com/ai-tools/quillbot-vs-grammarly/> (accessed 2025-11-14).
16. Karjo, C. H. (2025). Comparative Analysis of Automated Writing Evaluation Tools (Grammarly and QuillBot) in Academic Feedback. 2025 International Conference on Education Technology and Computers (ICETC), 321–326. <https://doi.org/10.1109/icetc66579.2025.11387276>.
17. Kadlub, M. Sources of Ambiguity in Language. *Studia Anglica Resoviensia* **2017**, 14, 44–57. <https://doi.org/10.15584/sar.2017.14.4>.
18. Boland, J. E. Lexical Ambiguity. *Linguistics* **2023**. <https://doi.org/10.1093/obo/9780199772810-0309>.
19. Jordi Fortuny. Structural Ambiguity and the Architecture of Language. *Studia Linguistica* **2023**, 78 (1), 35–63. <https://doi.org/10.1111/stul.12220>.
20. Sagi, E.; Rips, L. J. Identity, Causality, and Pronoun Ambiguity. *Topics in Cognitive Science* **2014**, 6 (4), 663–680. <https://doi.org/10.1111/tops.12105>.
21. Drury, J. E.; Baum, S. R.; Valeriote, H.; Steinhauer, K. Punctuation and Implicit Prosody in Silent Reading: An ERP Study Investigating English Garden-Path Sentences. *Frontiers in Psychology* **2016**, 7. <https://doi.org/10.3389/fpsyg.2016.01375>.
22. Musa, H.; Almryad, A. Comparison between Five Popular Paraphrasing Tools in Academic Writing in Terms of Strengths, Weaknesses, and Effectiveness in Rephrasing. *Karabuk University - Computer Engineering and Software Engineering Departments* **2025**, 3 (1). <https://doi.org/10.71074/ctc.1653566>.

■ Author

Anya Yarlagadda is a junior at Sharon High School. She is interested in the intersection of linguistics and AI, and hopes to pursue a career advancing the fields of computational linguistics and NLP.

Evaluating the Effectiveness of US-imposed Sanctions on Venezuela: Fiscal Constraint and Welfare Loss

Ariyaana Doss

Charterhouse School, Hurtmore Road, Godalming, Surrey, GU7 2DX, UK; dossariyaana@gmail.com

Mentor: Dr. Ceyhun Elgin

ABSTRACT: Economic sanctions have been a prominent feature of the international political economy since their emergence as a routine instrument of coercion during the Cold War. However, despite their continued use in modern politics, their effectiveness and humanitarian consequences remain deeply contested. This review paper seeks to evaluate the effectiveness of US-led economic sanctions on Venezuela by analyzing welfare loss and the regime's fiscal capacity. This paper asks: To what extent did US-imposed sanctions meaningfully constrain the Venezuelan regime, and what was the human cost of it? The existing literature surrounding the effectiveness of US-led sanctions is deeply divided. Policy-focused literature presents the strength of fiscal constraint through reduced oil production revenues and restricted access to international markets. However, humanitarian-focused literature presents the significant welfare losses borne by the population. This paper argues that US-led economic sanctions on Venezuela constrained the regime. It offers a synthesis of economic and political economy perspectives. It will contribute to broader debates about the legitimacy of economic sanctions as a foreign policy tool.

KEYWORDS: Economics, Economic sanctions, Political economy, Venezuela, Fiscal, Welfare, Foreign.

■ Introduction

Economic sanctions remain one of the most widely used instruments of foreign policy. Sanctions are designed to impose fiscal strain on targeted nations, coercing political change without military force. Yet, their effectiveness remains deeply contested due to the unintended consequences. US-led sanctions in Venezuela have been called into question due to the economic collapse and humanitarian crisis that were observed.

The paper proceeds as follows: First, we will explore the background of the economic sanctions imposed on Venezuela. Next, fiscal constraints imposed on the state will be explored. The direct effect of the sanctions on fiscal constraint will be identified, with the effect over time also outlined. Then, we will explore the welfare loss imposed on the population. The term will be defined. The direct effect of the sanctions on welfare loss will be identified, with the effect over time also outlined.

The United States first imposed sanctions on Venezuela in 2006. Following President Hugo's accession to power, President George W. Bush banned the sale of weapons and US-made military equipment to Venezuela.¹ The United States imposed its second round of sanctions in late 2014 and early 2015; President Obama targeted top Venezuelan officials, freezing US assets and banning visas. In 2017, Trump's first administration imposed financial sanctions on top officials for undermining parliament, human rights abuses, and anti-democratic actions. Washington also froze Maduro's US assets after he created a Constituent Assembly to override the parliament. In 2019, the US tightened economic sanctions, aiming to severely restrict economic activity and to oust Maduro. Washington also began oil-focused sanctions. It imposed sanctions on Venezuela's central bank and on Petróleos de Venezuela SA (PDVSA) - the

state-owned oil and gas company of Venezuela. On April 28th of the same year, Washington further placed an oil embargo on Venezuela and froze all government assets in the US.¹

By the time the US imposed broader sanctions, Venezuela's economy was already in decline. Venezuela is highly oil-dependent, as petroleum accounts for the majority of export earnings and government revenue. Following the 2014 collapse of oil prices, Venezuela experienced a contraction in GDP. Between 2014 and 2017, inflation rose, and the government entered partial default.¹ This suggests that the economy was structurally fragile before the escalation of sanctions. Post-2017 measures likely intensified, rather than produced, existing economic conditions.

This paper will focus on the sanctions of 2017, which targeted oil production. These sanctions resulted in declining oil production as reflected in a 2021 report by the US Government Accountability Office, which found that oil production had declined 46% from 2010 levels, before the 2017 sanctions, and fell another 89% in the 18 months following the 2019 sanctions.² However, these sanctions have intensified declining humanitarian conditions. Over 300,000 Venezuelans have no access to life-saving medications and treatments, and food imports dropped by 78% in 2018 (compared to 2013); these health-assistance programs were funded by PDVSA.³

Figure 1 demonstrates that oil production had begun declining before major US-imposed sanctions. However, the declining trend accelerated after 2014, which shows the vulnerability of Venezuela's oil-dependent economy.

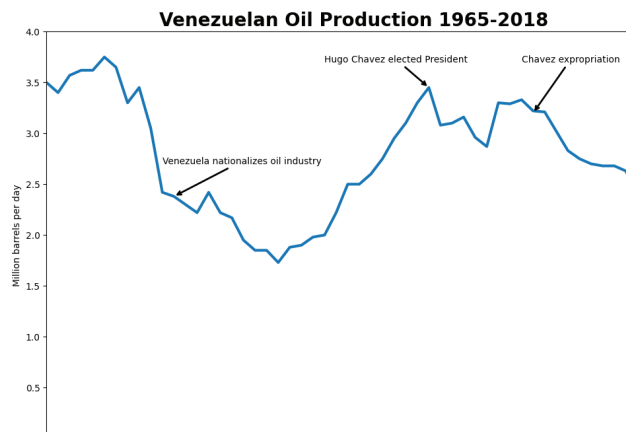


Figure 1: Venezuela's oil production 1965-2018.⁴

Figure 2 shows a dramatic fall in oil production following the 2017 US-imposed sanctions on Venezuela. This suggests that sanctions intensified existing weaknesses in the oil sector.

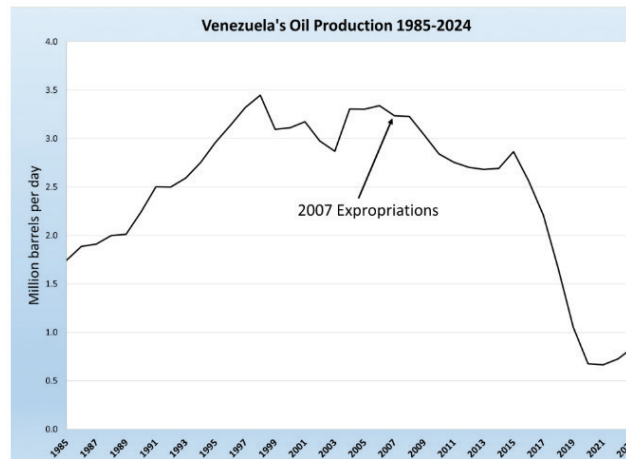


Figure 2: Venezuela's oil production 1985-2024.⁴

Therefore, this paper seeks to evaluate the effectiveness of US-led economic sanctions on Venezuela by evaluating the regime's fiscal capacity and the subsequent human costs incurred.

The existing literature on the US-led sanctions on Venezuela remains deeply divided. Fiscal constraint is emphasized as a form of economic coercion, while some sources suggest that welfare loss undermines the effectiveness of the sanctions. Literature focused on policy and law suggests that economic sanctions are purely instruments intended to constrain fiscal capacity.^{1,5,6} This is confirmed through empirical findings of oil production contraction, export earnings, and access to international markets.⁷⁻⁹ However, Kaempfer and Lowenberg also argue that sanctions lead to unintended consequences.¹⁰ Literature focused on the humanitarian dimension of sanctions suggests that these sanctions failed to translate into fiscal pressure and rather resulted in high social cost. Quantitative evidence reinforces this belief through projections of declining health services, increasing mortality rates, and food insecurity.^{3,11} Several sources caution against attributing welfare loss to the 2017 sanctions, citing pre-sanctions mismanagement and a declining oil sector.^{12,13} Additionally, Levy argues that the actual impact of sanctions may be limited and may ineffective-

ly constrain regimes.¹⁴ Therefore, there is a lack of consensus across the literature, which raises the critical need to evaluate these sanctions through a lens of fiscal constraint achieved and welfare loss results.

This paper identifies US-imposed sanctions as the independent variable, specifically looking at post-2017 conditions. This paper looks at two dependent variables: welfare loss and fiscal constraint. Welfare loss is a reduction in overall societal well-being. It will be illustrated by socioeconomic indicators, including rates of poverty, rates of education, and access to healthcare. Fiscal constraint is defined as a reduction in a state's capacity to generate revenue. Economic sanctions are intended to influence state behavior through fiscal constraint. Indicators will include declines in oil export revenues, contractions in GDP, and increased limitations on public spending. This paper seeks to identify the consequences of US-imposed economic sanctions on these two variables.

This research question holds important significance in political economy literature, expanding beyond the Venezuelan case study. The United States employs economic sanctions as a central instrument of its foreign policy. This continued use assumes that they are effective. By outlining direct impacts on fiscal constraint and welfare loss, this paper highlights how the regime was constrained.

■ Discussion

Fiscal Constraint:

Economic sanctions impose fiscal pressure on the target state by restricting access to international markets, revenue generation, and trade. Constraining the fiscal capacity of the state can limit its ability to sustain governance and therefore incentivize regime change. Fiscal constraint refers to limitations on government spending and taxing policies. In resource-dependent economies, fiscal constraint is salient as government revenue is concentrated in a narrow export base. Venezuela's economy has been historically dependent on oil revenues; sanctions, therefore, explicitly restricted the state's primary revenue stream.

This section finds that US-led sanctions generated significant fiscal pressure; this is most notable in the short term through restrictions on oil exports. However, the government has adapted over time, and as a result, the effect of sanctions has weakened.

The 2017 US-led sanctions produced sharp and measurable fiscal constraint, most visibly seen in reductions in oil export revenue and government spending.

The Venezuelan economy has been dependent on oil revenue. Before the sanctions, oil accounted for 95% of export revenue and 50-60% of government revenue.¹⁵ The sanctions targeted Petróleos de Venezuela S.A (PDVSA) and restricted the company's access to international markets. As a result, oil exports declined sharply. OPEC and the IMF detail that crude oil exports fell by 46% from 2016 to 2018, falling from 2.4 million barrels per day to 1.3 million barrels per day. In 2019, production dropped further to less than 800,000 barrels per day - the lowest level in decades. Oil exports to the USA fell from over 500,000 barrels per day in 2017 to near zero in

2019.¹⁶ The decline in export capacity resulted in immediate government revenue shortfalls. Government oil revenues fell by an estimated annual USD \$30–35 billion between 2016 and 2019. This restricted the state's ability to finance imports, fund public spending, and service debt. The direct sanctions on PDVSA restricted the company's access to international payment systems, shipping services, and refinery inputs, compounding revenue loss beyond the initial export decline.¹⁶

Figure 3 shows that oil production declined after sanctions targeted PDVSA. This suggests that sanctions constrained the Venezuelan regime's primary source of export revenue.

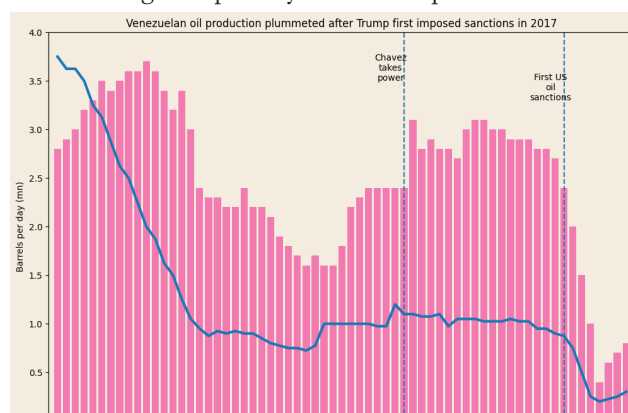


Figure 3: Venezuela's oil production plummeted after Trump's first imposed sanctions in 2017.¹⁷

The reduction in oil exports constrained the government budget and, in turn, government spending. In 2017, government expenditure fell by over 30%; steep cuts were seen in social and investment sectors. Social programs, including food subsidies (CLAP), infrastructure maintenance, and healthcare service access, experienced shortfalls due to their funding being derived from oil revenue.^{18,19} Public sector wages also declined due to inflation and delayed payments. Public investment also collapsed by more than 70%, showing how the sanctions constrained the state's ability to fund infrastructure, maintenance, and development projects.^{18,19} However, it can be observed that public spending contraction was not uniform across the economy. Regime-preserving activities such as selective subsidies and security subsidies were protected. This reallocation suggests that while sanctions reduced fiscal capacity, the state directed available resources to preserve core aspects of the regime, such as the armed forces.

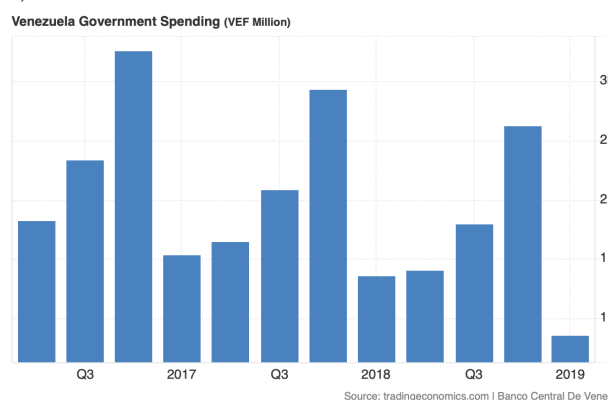


Figure 4: Venezuela government spending in vef million 2017–2019.²⁰

Figure 4 shows that government expenditure fell dramatically following the decline in oil revenues caused by US-imposed sanctions. The state had reduced fiscal capacity, which limited its ability to finance public services and investment projects.

Concept	Cumulative percent decline, 2012–2020	Percentage contribution
Change in per capita GDP	-71.5%	100.0%
Oil GDP	-8.2%	11.3%
Sanctions effect	-4.1%	5.7%
Other causes	-4.1%	5.7%
Non-oil GDP	-63.4%	88.7%
Import capacity	-32.9%	45.9%
Oil exports	-30.0%	42.0%
Oil price	-12.7%	17.8%
Oil production	-17.3%	24.2%
Sanctions effect	-8.7%	12.2%
Other causes	-8.6%	12.0%
Permanent loss of access to credit	-2.8%	3.9%
TPP	-23.5%	32.9%
Sanctions and toxification effects	-18.7%	26.2%
Other causes	-4.8%	6.7%
Factor Accumulation	-7.0%	9.8%
Sanctions and toxification effects	-5.6%	7.8%
Other causes	-1.4%	2.0%
Aggregates		
Change in per capita GDP	-71.5%	100.0%
Sanctions and toxification effects	-39.9%	55.8%
Other causes	-31.6%	44.2%

Note: Table decomposes Venezuela's 2012–2020 contraction into the contribution of sanctions and other politically induced restrictions on the country's insertion into global markets.

Figure 5: Decomposition of 2012–2020 economic growth.²¹

Figure 5 shows that Venezuela's GDP per capita declined by 71.5% from 2012 to 2020. It attributes sanctions and toxification effects to contributing 55.8% of this total contraction. This suggests that sanctions fiscally constrained the regime and were one of the major drivers of its economic contraction.

In the long-term, these sanctions have resulted in GDP contraction and state adaptation. Trade partnerships with China, Russia, and other regional actors illustrate that the sanctions did not achieve regime change despite imposing significant fiscal constraints.

The sustained effects of the sanctions on fiscal constraint are reflected in the trajectory of the economy's GDP. Following 2017, the GDP contracted severely. Venezuela's real GDP fell by 35% between 2017 and 2019.^{22,23} This follows a near 50% decline between 2013–2017. The steep decline can be attributed to reduced oil revenues, government spending, and limited investment. Shrinking GDP reduced the tax base and constrained the government's ability to provide public services and finance imports. Inflation stabilized from its 800% peak, but fiscal capacity remained limited.²⁴ Government spending as a share of GDP fell below 15% by 2021.²⁵ This shows the persistent constraint on the state's ability to provide public services, even as macroeconomic indicators are moderated.

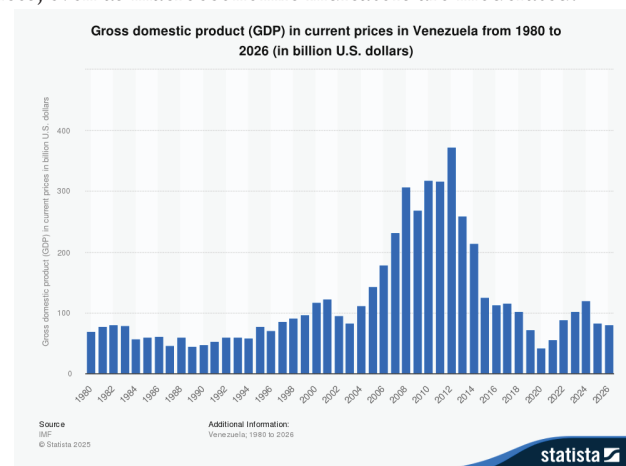
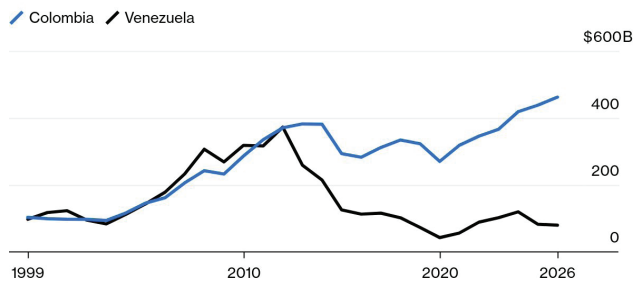


Figure 6: Venezuela gdp 1980–2026.²²

Figure 6 illustrates the scale of Venezuela's economic collapse. GDP fell sharply following 2012 and has continued to remain below previous levels. The economy is seen to have a modest recovery, but has not returned to pre-2012 levels.

Venezuela's Economic Collapse Stands Alone

GDP at current prices for Venezuela and Colombia since 1999



Source: International Monetary Fund
Note: Figures for 2026 are estimates

Bloomberg Opinion

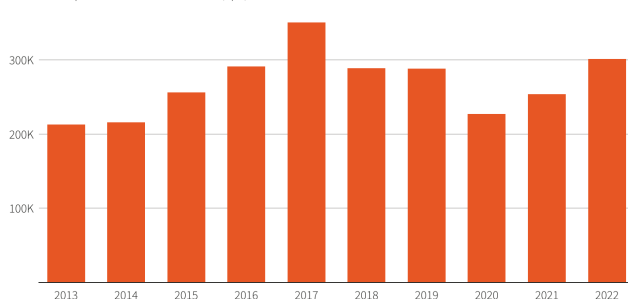
Figure 7: Venezuela's economic collapse stands alone - gdp from 1999 to 2026.²⁶

Figure 7 compares GDP between Venezuela and Colombia. Venezuela's GDP seems to persist at low levels, despite signs of recovery. This shows the long-term economic impacts of US-imposed sanctions.

Despite fiscal constraints, the state has adapted by establishing alternative trade relationships. China emerged as a key partner. It has provided loans and oil-for-loans agreements, which have enabled Venezuela to maintain essential imports and hence stabilize sections of public spending. Between 2007 and 2016, China provided over USD \$60 billion in loans to Venezuela; to contextualize, in 2016, Venezuela's GDP was at USD \$112.91B.^{22,27,28} By 2020, China had provided USD \$50 billion in oil-backed loans; this has facilitated continued fiscal operations; these loans exceeded Venezuela's GDP of USD \$50B.^{22,27,28} Venezuela has also developed oil-for-goods deals and bilateral trade agreements. These have allowed the state to restore certain revenue flows. This explains the partial recovery of the economy as domestic sectors are supported. These agreements do not fully compensate for the lost oil revenue, but they portray the state's adaptive capacity under fiscal constraint.

Unofficial shipments of Venezuelan crude to China remain high despite US sanctions

Chinese imports of Venezuelan crude oil (bpd)

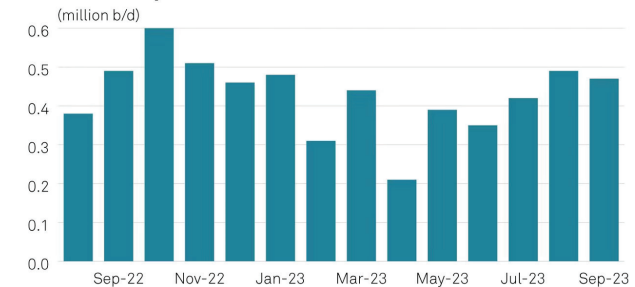


Note: Kpler estimates based off shiptracking data as at 11 September 2023
Source: Kpler

Figure 8: Unofficial shipments of Venezuela's crude to China remain high despite sanctions 2013-2022.²⁷

Figure 8 indicates that Venezuela has continued exporting to China through unofficial channels. This highlights the regime's adaptation to offset restricted access to Western markets, due to US-imposed sanctions.

Chinese independent refineries' Venezuelan feedstock imports



Source: S&P Global Commodity Insights

Figure 9: Chinese independent refineries' Venezuelan feedstock imports (September 2022 - September 2023).²⁸

Figure 9 shows that Chinese independent refineries imported Venezuelan crude products after the US imposed its sanctions on Venezuela. This shows that alternative trade agreements weakened the fiscal constraint of the sanctions.

Welfare Loss:

Welfare loss refers to a reduction in the overall well-being of individuals or households. This is often reflected in declining income, health, consumption, and access to essential services.

In this paper, welfare loss pertains to the erosion of living standards and economic security through rising poverty, food insecurity, increased mortality, and deteriorating health outcomes. This aligns with Amartya Sen's 'The Capability Approach', which emphasizes individuals' freedoms to achieve wellbeing as a primary importance in society.²⁹ This is an important framework to apply when evaluating sanctions, as monetary indicators often understate harm in economies experiencing informalisation, rationing, or shortages. It is important to note that welfare loss is very rarely evenly distributed. Sanctions can be regressive and therefore disproportionately affect low-income households and those reliant on public services.¹⁹ Ordinary citizens face price shocks and public service collapse. Analyzing welfare loss, hence, requires aggregate indicators but also distributional consequences. Measuring welfare loss in sanctioned economies is accompanied by methodological challenges. Official statistics may be incomplete or politicized, and market prices can be distorted by subsidies. This paper employs a combination of human development, labor market, and macroeconomic indicators to assess welfare loss in Venezuela.

Inflation serves as a primary indicator. Rapid rises in prices distort real incomes, thus reducing disposable income and purchasing power. This disproportionately affects households with fixed and informal earnings. In extreme cases, i.e., hyperinflation, money can fail to function as a reliable medium of exchange - this forces households into subsistence strategies. Inflation is, therefore, a macroeconomic variable that can be treated as an indicator of declining welfare. Employment outcomes present labor-market transmission channels. Rising unemployment and underemployment reduce household disposable income, making households more susceptible to

poverty. Employment losses are often concentrated in public employment and trade-exposed sectors in sanctioned economies. Health-related indicators such as access to healthcare services and infant mortality rates also assess sustained welfare loss. These variables are non-reversible welfare losses. Income shocks can be partially recovered; however, excess mortality and long-term health deterioration can present permanent losses to societal welfare. Infant mortality is one of the most important indicators as it captures systemic welfare failure because it represents healthcare access, sanitation, nutrition, and maternal health.

In the short-term, one of the most notable effects of the sanctions was the dramatic acceleration of inflation. Before the 2017 sanctions, Venezuela was already experiencing hyperinflation, with a rate of 800% in 2016.³⁰ This figure increased in subsequent years due to restricted access to foreign financial systems, currency depreciation, import restrictions, and exchange controls. Hyperinflation resulted in the unaffordability of staple goods. The minimum wage lost any purchasing power; the cost of food, medicine, and other essentials increased beyond the means of most households. In August 2017, Venezuelans on minimum wage failed to afford their daily caloric intake – a family of five could only afford approximately 6,132 calories per day, which is only 56% of the minimum dietary requirement.¹³ Compared to 2010, the sanctions induced a 92% decline in calorie affordability.¹³

Figure 10 indicates that the inflation rate rose to extreme levels from 2019 to 2020. This eroded household purchasing power and demonstrated severe welfare consequences.

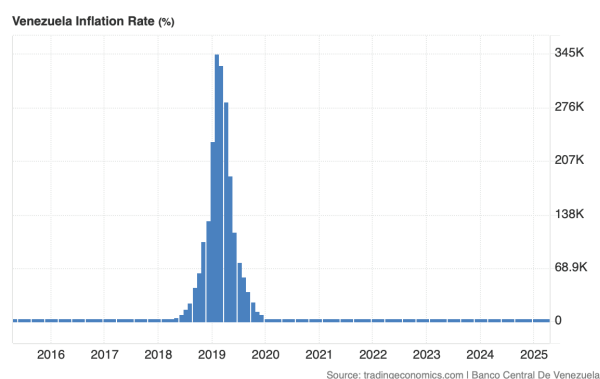


Figure 10: Venezuela's inflation rate.³¹

Sanctions impacted the labor market due to intensified contraction in economic activity. The comprehensive labor statistics are limited in sanctioned economies; however, multiple reports indicate widespread job losses and shrinking employment opportunities.^{32,33} This is most notable in formal sector industries, which consequently pushed many workers into the informal sector. The US Senate noted that Venezuela had the lowest levels of labor force participation in the region.¹² Informalisation of labor represents a shift in employment status but also a shift in economic security. Workers in the informal sectors earn volatile and lower incomes and lack social protection.³⁴

Figure 11 shows that the proportion of Venezuelans living below the poverty line has dramatically increased following

US-imposed sanctions. This indicates deterioration in living standards experienced by many citizens.

Population living below the poverty line

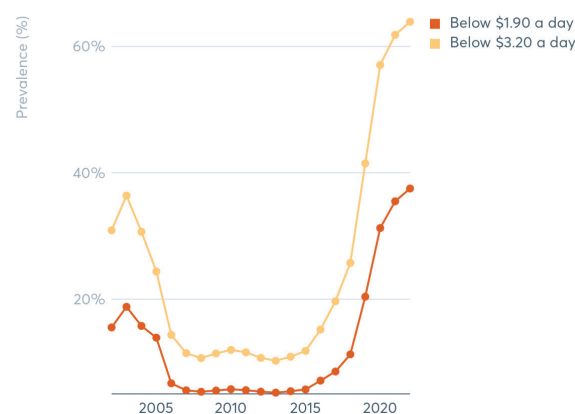


Figure 11: Population living below the poverty line.³⁵

Rising prices reduce household income and diminish the ability of households to purchase enough food. Data from the immediate period of the 2015 sanctions demonstrate that food imports fell by 71% between 2013 and 2016; in 2013, food imports stood at an estimated \$11.2 billion before dropping to \$2.46 billion in 2016.¹² This trend predates further sanctions of 2017 and 2019. This indicates that sanctions likely exacerbated existing shortages and reduced food availability by restricting access to imports. This resulted in reduced dietary diversity, forcing households to reduce meal frequency, substitute certain foods for cheaper but less nutritious options, or skip meals altogether. These behaviors degrade population health in the short-term and have significant impacts on the long-term health of the state. The 2019 UN Report 'Venezuela Overview of Priority Humanitarian Needs' states that ENCOVI surveys find that 22% of children under 5 are stunted due to lack of food.³⁶

Prevalence of undernourishment

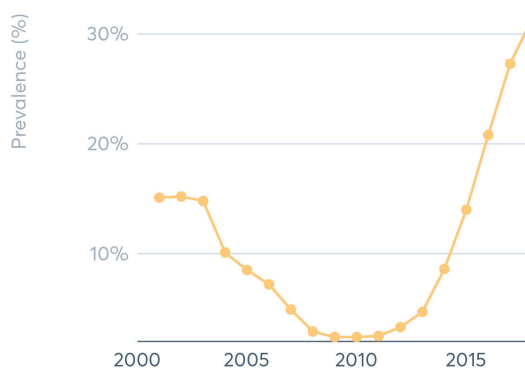


Figure 12: Prevalence of undernourishment.³⁵

Figure 12 shows that the prevalence of undernourishment has continued to rise following 2015. This indicates worsening

food security and access to nutrition, suggesting worse health outcomes for Venezuelans.

Health outcomes provide the clearest indicators of immediate welfare loss. Infant mortality rate and life expectancy integrate several welfare dimensions, including healthcare access, nutrition, and sanitation. Infant mortality rates reflect economic development, living conditions, and the quality of healthcare. Infant mortality rose by 44% between 2013 and 2016, before the 2017 sanctions.¹¹ Following the sanctions, the infant mortality rate has been constant at 21.5 deaths per 1,000 live births.^{37,38,19} An estimated 40,000 Venezuelans died due to the lack of access to basic commodities between 2017–18 – this reflects a 31% increase in general mortality. Access to healthcare also denotes the deteriorating quality of healthcare. In 2018, the Venezuelan Pharmaceutical Federation reported that shortages of essential medicines were at 85%. The lack of access to healthcare also resulted in over 300,000 people being at risk, including: 16,000 who need dialysis, 16,000 people with cancer, 4 million with hypertension and diabetes, and 80,000 people with HIV who require antiretroviral treatment.^{13,39}

Welfare loss can be seen to grow in severity in the long term. Welfare indicators suggest that deterioration in health, nutrition, and living standards continues rather than dissipating. Certain macroeconomic indicators, such as inflation, have shown stabilization; however, welfare outcomes for the civilian population remain depressed.

The Venezuelan labor market remains highly informalized.³³ Economic activity has recovered in dollarized sectors, but this has contributed to unevenly distributed gains. Low-income households have minimal access to such recovery. Therefore, informality undermines welfare by reducing income stability and increasing vulnerability to economic shocks.

Food insecurity remains widespread. Hyperinflation has slowed, yet real wages remain low. Resultantly, access to affordable food continues to be constrained. Household caloric intake remains significantly below the minimal dietary requirements. The persistence of undernourishment reflects structural weakness in food distribution, import financing, and agricultural production. Prolonged nutritional stress has a cumulative effect. Chronic undernourishment, especially within children, reduces educational attainment, cognitive development, and long-term productivity. These losses are not easily reversible and significantly diminish human capital. This extends welfare loss beyond the immediate crisis period.⁴⁰

Deteriorating health outcomes also remain persistent. Infant mortality rates remain elevated as health services maintain deficiencies in maternal care, nutrition, and neonatal services. Mortality levels continue to exceed pre-sanction levels. Life expectancy changes occur gradually as prolonged nutritional deficits and inadequate healthcare contribute to health stagnation over time. In 2021, former UN special rapporteur Alfred de Zayas estimated that 100,000 Venezuelans died as a direct result of the US-led sanctions.⁴¹ Welfare loss is not a static condition but an accumulating deficit. These collectively suggest that sanctions have negatively impacted Venezuelan human development, extending beyond the immediate period.

Under-5 mortality rate per 1,000 live births

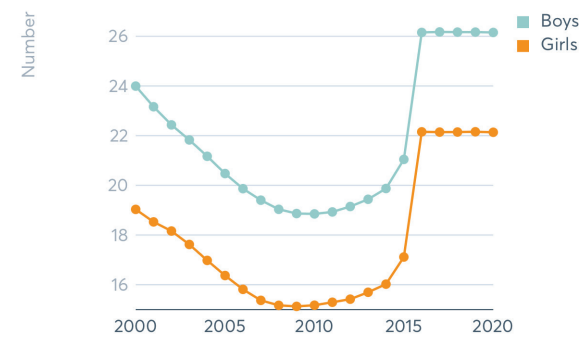


Figure 13: Under-5 mortality rate per 1,000 live births in Venezuela (2000–2020).³⁵

Figure 13 shows that there is a rise in under-5 mortality rates from 2015. This reflects declining access to healthcare and nutrition. Mortality is a non-reversible welfare loss and shows the extent of humanitarian loss associated with US-imposed sanctions.

Healthcare access represents sustained welfare loss under sanctions. Venezuela's healthcare system remains constrained due to consistent shortages of medical equipment, trained personnel, and essential medicines. Imports of pharmaceuticals and supplies remained inconsistent and insufficient due to high transaction costs and trade uncertainty. Alena Douhan, the UN Special Rapporteur, reported that Venezuela was experiencing 85% shortages of medicine in 2021.⁴¹ Households, in turn, face out-of-pocket healthcare costs and resort to informal or insufficient treatment. Additionally, preventative care has decreased, which has increased the prevalence of untreated chronic conditions, whilst vital procedures such as heart surgery, dialysis, and cancer treatment remain out of reach. Migration has also exacerbated the healthcare crisis. The country experienced a loss of 45.7% of its registered doctors. This resulted in public hospitals having 50–70% shortages of qualified medical personnel.⁴¹ Medical equipment compounds this loss, as 80% of hospital equipment is in disrepair – much of it being imported from the US. Sanctions also impeded the state's ability to purchase medical equipment from other countries.

Conclusion

US-imposed sanctions on Venezuela, specifically observed through post-2017 conditions, highlight direct impacts on fiscal constraint and welfare loss. Fiscal pressure imposed on the state was immediate and adaptive. Welfare loss imposed on civilians was immediate and cumulative. It can be argued that the sanctions produced a pronounced asymmetry in which civilians received the majority of the burden whilst the regime adapted to its reduced fiscal capacity. This explains the limitations of the US-imposed sanctions in Venezuela.

The paper measures societal welfare and fiscal pressure as primary metrics of effectiveness. Whilst the 2017 sanctions were imposed in response to the government's human rights abuses and anti-democratic actions, sanctions in foreign policy

are more broadly designed to coerce political change. Therefore, using these metrics is justified.

As Venezuela's economic decline predated the 2017 sanctions, there may be concerns about the attribution of welfare loss to the sanctions. It is difficult to isolate the effects on fiscal constraint and welfare loss of the US-imposed sanctions from pre-existing economic decline. The literature presents that the escalation of welfare loss and sector-specific shocks to imports and oil revenues post-2017 amplified existing vulnerabilities rather than merely coinciding with them. Ghomi argues that vulnerable groups suffer significant welfare loss as a result of these sanctions, which informs this paper's focus on civilian welfare; however, it does not detail implications for Venezuela's elite class. Additionally, this conclusion may be context-specific.⁴² In resource-dependent economies, such as Venezuela, adaptive capacity may result in fiscal constraints failing to result in political change, while welfare losses accumulate. This suggests that the effectiveness of sanctions is contingent on the nation's political and economic structures. This limits the external validity of sanctions as general coercive instruments.

The broader implications of this research suggest that sanctions can systematically produce asymmetries due to welfare costs on civilians while failing to sufficiently constrain the fiscal regime. This calls into question the efficacy of economic sanctions as foreign policy instruments and suggests the need for more targeted policy decisions. Further research can focus on comparative analysis of sanctions regimes and improved models to address causality. Comparative analysis across different regimes could focus on comprehensive sanctions versus targeted sanctions, and can identify the conditions under which sanctions are effective. Methodological improvements can separate independent effects from political or economic shocks.

■ AI Acknowledgments

During the preparation of this work, the author used ChatGPT to understand economic and statistical models, including the kernel density estimate, and to improve language readability. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

■ Acknowledgments

I would like to acknowledge and thank Dr. Ceyhun Elgin and Indigo Research for their guidance during my research process. I attest that the ideas, graphics, and writing in this paper are entirely my own.

■ References

- Alison, J.; Timura, C.; Toussaint, S. *Understanding US Sanctions on Venezuela*. Gibson Dunn, 2019. <https://www.gibsondunn.com/wp-content/uploads/2019/04/Lee-Timura-Toussaint-Understanding-US-Sanctions-on-Venezuela-Financier-World-wide-04-2019.pdf> (accessed 2020-05-20).
- Seelke, C. R. *Venezuela: Political Crisis and U.S. Policy*. Congress.gov. <https://www.congress.gov/crs-product/IF10230>.
- Doocy, S.; Page, K.; Liu, C.; Hoaglund, H.; Rodríguez, D. Venezuela: Out of the Headlines but Still in Crisis. *Bulletin of the World Health Organization* **2022**, *100* (8), 466–466A. <https://doi.org/10.2471/blt.22.288269>.
- Rapier, R. Charting the Decline of Venezuela's Oil Industry. *Forbes*. February 5, 2019. <https://www.forbes.com/sites/rpapier/2019/01/29/charting-the-decline-of-venezuelas-oil-industry/>.
- Jermano, J. Economic and Financial Sanctions in U.S. National Security Strategy. *PRISM* **2018**, *7* (4), 64–73.
- Monaldi, F.; Reyes, J. L. R. *U.S. Policy toward Venezuela's Hydrocarbon Sector*, 2021. <https://www.bakerinstitute.org/sites/default/files/2021-02/import/ces-wp-venezuela-020921.pdf>.
- Bahar, D.; Bustos, S.; Morales, J.; Santos, M. *Impact of the 2017 Sanctions on Venezuela: Revisiting the Evidence*, 2019. https://www.brookings.edu/wp-content/uploads/2019/05/impact-of-the-2017-sanctions-on-venezuela_final.pdf.
- International Monetary Fund. *Inflation rate, average consumer prices*. Imf.org. <https://www.imf.org/external/datamapper/PCPIPCH@WEO/VEN?year=2026>.
- Oliveros, L. *The Impact of Financial and Oil Sanctions on the Venezuelan Economy*, 2020. <https://www.wola.org/wp-content/uploads/2020/10/Oliveros-report-summary-ENG.pdf>.
- Kaempfer, W. H.; Lowenberg, A. D. The Theory of International Economic Sanctions: A Public Choice Approach. *The American Economic Review* **1988**, *78* (4), 786–793.
- Gutmann, J.; Neuenkirch, M.; Neumeier, F. *Sanctioned to Death? The Impact of Economic Sanctions on Life Expectancy and its Gender Gap* | ifo Institute. ifo Institute. <https://www.ifo.de/en/cesifo/publications/2019/working-paper/sanctioned-death-impact-economic-sanctions-life-expectancy-and-its>.
- Rendon, M.; Price, M. *Are sanctions working in Venezuela?* Csis.org. <https://www.csis.org/analysis/are-sanctions-working-venezuela>.
- Rodríguez, F. *Sanctions, Economic Statecraft, and Venezuela's Crisis Case Study*, 2022. https://sanctionsandsecurity.org/wp-content/uploads/2022/01/January-2022-Venezuela-Case_Rodriguez.pdf.
- Levy, P. I. Sanctions on South Africa: What Did They Do? *The American Economic Review* **1999**, *89* (2), 415–420.
- Rapier, R. Why Sanctions, Not Reserves, Will Decide Venezuela's Oil Future. *Forbes*. January 5, 2026. <https://www.forbes.com/sites/rpapier/2026/01/05/why-sanctions-not-reserves-will-decide-venezuelas-oil-future/>.
- Rodgers, L. Venezuela: All You Need to Know about the Crisis in Nine Charts. *BBC News*. February 4, 2019. <https://www.bbc.co.uk/news/world-latin-america-46999668>.
- Tauschinski, J.; Clark, D.; Rodgers, L.; Stylianou, N.; Moore, M. *The "catastrophic" state of Venezuela's oil facilities*. @FinancialTimes. <https://www.ft.com/content/cef0a0c5-59be-4b53-83da-4f7fc-609c2a8>.
- Nations, U. *Specific country data*. hdr.undp.org. <https://hdr.undp.org/data-center/specific-country-data#/countries/VEN>.
- Review, P.-C. G. H. *Sanctions on Venezuela: Implications of Economic Restrictions on the Health of Venezuelans*. PGHR. <https://www.pghr.org/post/sanctions-on-venezuela-implications-of-economic-restrictions-on-the-health-of-venezuelans>.
- O'Neill, A. *Venezuela Government Spending | 2021 Data | 2022 Forecast | 1998-2020 Historical*. tradingeconomics.com. <https://tradingeconomics.com/venezuela/government-spending>.
- Rodríguez, F. Sanctions and Venezuelan Migration. *arXiv (Cornell University)* **2025**. <https://doi.org/10.2139/ssrn.5072710>.
- Sanchez, J. *Gross domestic product (GDP) in current prices Venezuela* | Statista. Statista. https://www.statista.com/statistics/1392501/gross-domestic-product-gdp-in-current-prices-venezuela/?srsltid=AfmBOofmt8QzKzu_11d-0WsNWm0JZ5QIH-s6UyePTAYNvu3JH7TYkrb.

23. International Monetary Fund. *Venezuela*. www.imf.org. <https://www.imf.org/external/datamapper/profile/VEN>.
24. Trading Economics. *Venezuela Inflation Rate*. Tradingeconomics.com. <https://tradingeconomics.com/venezuela/inflation-cpi>.
25. Dewey, K. *Purchasing power – Venezuela's defense spending struggles to keep pace*. IISS. <https://www.iiss.org/online-analysis/military-balance/2026/02/purchasing-power--venezuelas-defence-spending-struggles-to-keep-pace/>.
26. Spinetto, J. P. *After Maduro Comes the Hard Part: Fixing Venezuela's Economy*. Bloomberg.com. <https://www.bloomberg.com/opinion/articles/2025-11-25/after-maduro-s-exit-comes-the-hard-part-fixing-venezuela-economy?embedded-checkout=true>.
27. Hayley, A. China's Oil Trade and Investment in Venezuela. *Reuters*. September 12, 2023. <https://www.reuters.com/business/energy/chinas-oil-trade-investment-venezuela-2023-09-12/>.
28. Xu, D.; Zhou, O.; Mohatny, S. *China's independent refineries to face competition as Venezuela gets partial sanctions relief*. S&P Global. <https://www.spglobal.com/energy/en/news-research/latest-news/crude-oil/101923-chinas-independent-refineries-to-face-competition-as-venezuela-gets-partial-sanctions-relief>.
29. Crocker, D. A.; Robeyns, I. *Capability and Agency*. Cambridge University Press. <https://www.cambridge.org/core/books/abs/amartya-sen/capability-and-agency/65BD3415B565147A740E-03F42E41D047>.
30. Rapier, R. Charting the Decline of Venezuela's Oil Industry. *Forbes*. February 5, 2019. <https://www.forbes.com/sites/rapiere/2019/01/29/charting-the-decline-of-venezuelas-oil-industry/>.
31. Trading Economics. *Venezuela Inflation Rate*. Tradingeconomics.com. <https://tradingeconomics.com/venezuela/inflation-cpi>.
32. Bahar, D.; Hausmann, R. *Sanctions on Venezuela Are Not Driving Migration to the US Southwest Border: An Empirical Assessment*; 2025. <https://www.cgdev.org/sites/default/files/sanctions-venezuela-not-driving-migration-southwest-border-4-14-25.pdf> (accessed 2025-09-23).
33. Bull, B.; Rosales, A. Sanctions, Rentierism, and Economic Informalization in Venezuela. *European Review of Latin American and Caribbean Studies / Revista Europea de Estudios Latinoamericanos y del Caribe* **2020**, No. 109, 107–133. <https://doi.org/10.2307/26936905>.
34. Trading Economics. *Venezuela Unemployment Rate*. Tradingeconomics.com. <https://tradingeconomics.com/venezuela/unemployment-rate>.
35. Global Nutrition Report. *Country Nutrition Profiles*. globalnutritionreport.org. <https://globalnutritionreport.org/resources/nutrition-profiles/latin-america-and-caribbean/south-america/venezuela-bolivarian-republic/>.
36. UN Human Rights report on Venezuela urges immediate measures to halt and remedy grave rights violations. OHCHR. <https://www.ohchr.org/en/press-releases/2019/07/un-human-rights-report-venezuela-urges-immediate-measures-halt-and-remedy>.
37. Delgado, D. *U.S. sanctions kill as much as war*. People's World. <https://peoplesworld.org/article/u-s-sanctions-kill-as-much-as-war/> (accessed 2026-01-18).
38. Reidpath, D. D.; Allotey, P. Infant Mortality Rate as an Indicator of Population Health. *Journal of Epidemiology & Community Health* **2003**, 57 (5), 344–346. <https://doi.org/10.1136/jech.57.5.344>.
39. Andreína Chávez Alava. *US Sanctions Are Killing Venezuelans*. Venezuelanalysis. <https://venezuelanalysis.com/analysis/us-sanctions-are-killing-venezuelans/>.
40. Weisbrot, M.; Sachs, J. *Economic Sanctions as Collective Punishment: The Case of Venezuela*; 2019. <https://cepr.net/images/stories/reports/venezuela-sanctions-2019-04.pdf>.
41. Celina della Croce. *How U.S. Sanctions Are a Tool of War: The Case of Venezuela*. Venezuelanalysis. <https://venezuelanalysis.com/analysis/how-us-sanctions-are-a-tool-of-war-the-case-of-venezuela/>.
42. Ghomi, M. Who Is Afraid of Sanctions? The Macroeconomic and Distributional Effects of the Sanctions against Iran. *Economics & Politics* **2021**, 34 (3). <https://doi.org/10.1111/ecpo.12203>.
43. Fiveable. *Fiscal Policy Constraints Definition – International Economics Key Term | Fiveable*. Fiveable.me. <https://fiveable.me/key-terms/international-economics/fiscal-policy-constraints> (accessed 2026-01-19).
44. Freier, M.; Ricci, M. *Fiscal policy to the rescue: How governments shielded households from inflation*. European Central Bank. <https://www.ecb.europa.eu/press/blog/date/2023/html/ecb.blog231123~3462bc66fc.nl.html> (accessed 2026-01-18).
45. Fund, U. C. *UNICEF Venezuela Humanitarian Situation Report (31 March 2019)*. ReliefWeb. <https://reliefweb.int/report/venezuela-bolivarian-republic/unicef-venezuela-humanitarian-situation-report-31-march-2019> (accessed 2026-01-18).
46. Ghironi, F.; Ozhan, G. K.; Kim, D. *International trade and macroeconomic dynamics with sanctions*. CEPR. <https://cepr.org/voxeu/columns/international-trade-and-macroeconomic-dynamics-sanctions>.
47. Hernandez, J. I. *The Human Rights Impact of Sanctions Policy in Venezuela | GJIA*. Georgetown Journal of International Affairs. <https://gjia.georgetown.edu/2025/05/14/oil-licenses-and-repression-the-human-rights-impact-of-sanctions-policy-in-venezuela/>.
48. Hopkins, S. *Economic Sanctions in International Law: Venezuela's Conundrum*. UIC Law Review. <https://lawreview.law.uic.edu/news-stories/economic-sanctions-in-international-law-venezuelas-conundrum/>.
49. Macrotrends. *Venezuela Unemployment Rate 1991–2024*. www.macrotrends.net. <https://www.macrotrends.net/global-metrics/countries/VEN/venezuela/unemployment-rate>.
50. O'Neill, A. *Ratio of government expenditure to GDP Venezuela | Statista*. Statista. <https://www.statista.com/statistics/371925/ratio-of-government-expenditure-to-gross-domestic-product-gdp-in-venezuela/?srsltid=AfmBOoomME-9EwyfjBe5NYFDdk77GJnPg8XdPGGNu7-JyI836b2jV0PB> (accessed 2026-02-01).
51. Pala, T. The Effectiveness of Economic Sanctions: A Literature Review. *NISPA Journal of Public Administration and Policy* **2021**, 14 (1), 239–259. <https://doi.org/10.2478/nispa-2021-0009>.
52. Robeyns, I. *The Capability Approach (Stanford Encyclopedia of Philosophy)*. Stanford.edu. <https://plato.stanford.edu/entries/capability-approach/>.
53. Rohenkohl, B.; Ortiz-Ospina, E. *Brazil, Venezuela, and Mexico account for 59% of people living in poverty in Latin America*. Our World in Data. <https://ourworldindata.org/data-insights/brazil-venezuela-and-mexico-account-for-59-of-people-living-in-poverty-in-latin-america>.
54. Roy, D. *Do U.S. Sanctions on Venezuela Work?* Council on Foreign Relations. <https://www.cfr.org/in-brief/do-us-sanctions-venezuela-work>.
55. Sutherland, M. *The Economic Sanctions on Venezuela: Consequences, Humanitarian Crisis, Alternatives and Humanitarian Agreement*; 2020. https://provea.org/wp-content/uploads/2021/02/infome_sancciones_ingles_v3.pdf (accessed 2026-01-18).
56. Arena, P.; Nicoletti, N. P. Selectorate Theory, the Democratic Peace, and Public Goods Provision. *International Theory* **2014**, 6 (3), 391–416. <https://doi.org/10.1017/s1752971914000347>.

■ Author

Ariyaana Doss is a current Year 12 student at Charterhouse School, studying A-Level Economics, Politics, and Mathematics, alongside an EPQ. She aspires to read PPE or Politics and Economics at a leading UK university with a particular interest in political economy.

Extrinsic Bias Patterns May Suggest Underlying Intrinsic Bias: A Cross-Domain Analysis

Aryan Potukuchi

Brisbane Boys College, Kensington Terrace, Toowong, Queensland, 4066, Australia; aryanpotukuchi@gmail.com

ABSTRACT: Racial biases in LLMs can significantly influence their decisions, resulting in socially unjust consequences. The term ‘Large language models’ refers to advanced computer programs capable of synthesizing complex, human-like responses. They use vast amounts of text as training data, often inheriting certain biases present that degrade the morality of their replies. Whilst measuring and mitigating such biases is crucial, most research is dedicated to moderating extrinsic biases found in LLM outputs, leaving intrinsic biases within LLM mechanisms relatively unexplored. This study will provide a set of criteria used to recognize bias patterns within the data. Using the criteria, we will analyze evidence of LLM performance in fabricated medical, legal, and hiring settings and identify extrinsic racial bias patterns. Our findings indicate that there are similarities between the identified patterns and the concepts of intrinsic biases, suggesting a transfer of intrinsic bias structures into LLM output performance. Additionally, our findings may also suggest the importance of exploring socially derived biases within LLM training data. Overall, our review acts as a foundation for other researchers to validate the correlation between extrinsic and intrinsic bias, potentially contributing to the bias transfer hypothesis and other broader discussions.

KEYWORDS: LLMs in Critical Societal Domains, Racial Bias in LLMs, Large Language Models, Bias Transfer Hypothesis.

■ Introduction

Advancements in Large Language Models (LLMs) in fields such as Natural Language Processing (NLP) introduce exciting possibilities, such as LLM incorporation into critical societal domains. However, there are concerns about the effects of LLM biases, often inherited from pre-training data, on the generated content. In fields like law, medicine, and hiring, such biases can influence hospitalization and medical costs,¹ judicial decisions,² and job employment,³ resulting in serious ethical violations and social inequity. This makes it of utmost importance to quantify and mitigate these biases within LLMs and make them more suitable for such critical domains.

Bias, in this context, refers to systematic errors in LLM token prediction that reflect the stereotypes and prejudices of our world.⁴ These biases exist across a broad spectrum, including gender and race, sexual orientation, religion, socio-economic status, and disability.⁵ Despite the versatility, LLM bias is usually classified into two groups: intrinsic bias and extrinsic bias. Guo *et al.*⁶ state that intrinsic bias refers to the bias within a pre-trained or trained LLM’s internal mechanisms, irrespective of downstream tasks. Extrinsic bias, on the other hand, refers to the bias within the LLM’s outputs after training or pre-training, usually task-dependent. An in-depth discussion between the two bias types is provided in the definitions section. While research papers examine the quantification and mitigation of both intrinsic and extrinsic biases within LLMs individually, a small portion explores the theoretical relationship between the bias categories through the bias transfer hypothesis.

The bias transfer hypothesis refers to the assumption that intrinsic biases within LLMs propagate into downstream tasks, resulting in extrinsic biases within LLM outputs.^{7,8} However, previous studies that test this hypothesis measure the intrinsic-extrinsic correlation using quantitative approaches, such as metrics-based comparisons, finding weak correlations.^{9–11}

This means that they might overlook other potential indicators of bias transfer that don’t rely on quantitative propagation, such as comparisons of bias structure or patterns. Hence, we have decided to approach the hypothesis from an interpretive perspective by exploring the correlation between ‘racial’ bias patterns within the downstream tasks from the Law, Medicine, and Hiring domains, and concepts of intrinsic biases. To distinguish these patterns from statistical noise within the data, we have developed our own set of criteria that assess the strength of the patterns based on three core characteristics – consistency, directionality, and replicability. A detailed description of the criteria is provided in the ‘Criteria for Identifying Bias Trends’ section.

The paper initially established key terminology, such as the definitions of common social biases, including intrinsic and extrinsic bias definitions. Additionally, the section briefly elaborates how intrinsic biases can originate from biased pre-training data, which affects the LLM’s mechanisms. It also discusses how extrinsic biases tend to be downstream effects of intrinsic biases. Before moving on to the analysis, the paper establishes a set of criteria used to identify bias patterns within the data. In the results section, we highlight the importance of the critical domains in society, along with the impacts of extrinsic bias within these domains, and analyze our extrinsic racial bias evidence. Following this, we identify prominent bias patterns within the evidence using the criteria, interpreting them using established intrinsic bias definitions. In the discussion section, we discuss our interpretations of the evidence, along with its relation to the bias transfer hypothesis and other broader concepts. We also state the limitations of the research,

providing further directions for future researchers to verify the interpretations within our research.

Overall, this study contributes to the ongoing conversation of the bias transfer hypothesis by presenting an interpretive perspective on the transfer of bias structures, as opposed to the quantitative propagation measured using metric-based systems. Additionally, our findings may provide valuable insights into discussions regarding the effects of societal biases within LLM training and evaluation practices, suggesting that an understanding of socially grounded biases may be required to improve bias mitigation strategies.

■ Key Terminology

Types of Social Bias:

The term ‘social bias’ encompasses any discriminatory behavior that disadvantages marginalized groups in society. Although there are several types of social biases, discussed in detail by Navigli, Roberto *et al.*,⁵ this section will only define a few common manifestations, which are listed in Table 1.

Table 1: Different types of Social Biases, along with their definitions and examples.

Social Bias	Definition	Example
Age	Age bias or ageism is the discrimination against an individual or a group based on their age.	An LLM from a company that values efficiency might hire more candidates below a certain age (37 years) because it associates ‘youth’ with words such as ‘active’ and ‘efficient’.
Disability	Disability bias is the discrimination against individuals with disabilities, including physical, genetic, and neurological.	For example, an LLM might rank a neurotypical individual higher than a neurodivergent one for the position of a manager because it perceives neurotypical people to be more capable of high-stakes decision-making.
Gender	Gender bias or sexism is the tendency to prefer or assign a single gender over another for a specific task.	An LLM might assume that most doctors in a hospital are males, whereas most nurses are females.
Socioeconomic Status	Socioeconomic bias is the discrimination against individuals or groups with lower socioeconomic status, such as poor people.	An LLM used for Treatment recommendations may suggest inferior, low-cost procedures for people with lower socioeconomic status during a medical emergency.
Sexual Orientation	Sexual bias is discrimination against individuals based on their sexual identity.	An LLM might prefer a heterosexual candidate for a teaching position over a homosexual candidate because it considers the heterosexual individual to be a better role model for children.
Race or Ethnicity	Racial or ethnic bias includes discriminatory behaviors and negative attitudes towards individuals from specific ethnic groups.	An LLM hiring for a company might score white men higher for an important position than black men.
Intersectional	Intersectional bias refers to discrimination against individuals based on multiple aspects of their social identity.	An LLM that isn't biased towards race (white or black) and gender (man or woman) might still be biased towards a specific intersection between them (Black woman).

While there exist many manifestations of social bias in LLM outputs, this paper focuses specifically on racial bias to analyze the relationship between downstream bias patterns and upstream bias mechanisms. The definition of social biases is presented to acknowledge the vast diversity in output bias and to restate that conclusions made within this report cannot be generalized for other social biases without further research.

Intrinsic Bias:

Bias present in large language models is usually classified into two categories, depending on where it is portrayed: intrinsic bias and extrinsic bias. Intrinsic bias refers to bias within the internal mechanisms of the model, often directly associated with the model’s pattern recognition and relationship

encoding within these mechanisms.⁶ For example, a model trained on a medical corpus that frequently represented men as doctors and women as nurses would learn to identify the strong relationship between ‘doctor’ and ‘man’, and likewise for ‘nurse’ and ‘woman’. Consequently, it would change its internal representation to reflect this pattern when prompted.

While there are many causes of intrinsic bias, such as algorithmic factors¹² and annotation bias,¹³ the most common cause is bias in pre-training data. The LLM pre-training process involves multiple steps essential for the model’s text comprehension and generation. These include the collection and processing of raw data for the LLM pre-training dataset, tokenization of data into unique numerical tokens, training the model on tokenized corpora for pattern learning and next-word prediction, and finally, further testing and evaluation to improve response accuracy and quality.¹⁴ However, when trained on biased data, the model provides more attention weight for biased associations with sensitive tokens, increasing the probability of these associations being generated in the responses upon token detection within the prompt.¹⁵

Extrinsic Bias:

Unlike intrinsic bias, extrinsic bias refers to inequality in the model’s performance against specific demographic groups across different downstream tasks.⁶ Being task-dependent, extrinsic bias manifests in different ways, which have been explored in detail by Guo *et al.*⁶ These tasks are classified into Natural Language Understanding (NLU) and Natural Language Generation (NLG).

Natural Language Understanding tasks focus on improving the model’s semantic understanding of the input text to enhance its task-specific performance. However, extrinsic bias in NLU tasks can result in misinterpretation of texts due to biased notions from pre-training. For example, in the sentence, “The doctor called the nurse because she was late.” The LLM might associate ‘she’ with ‘nurse’ due to extrinsic bias. This can affect its performance in tasks such as natural language inference.

On the other hand, Natural Language Generation tasks focus on generating relevant information based on the input prompts. Extrinsic bias in these tasks can result in inappropriate responses, including reinforcement of social biases and stereotypes inherited from training data. For example, the LLM might rank male candidates higher than equally qualified female candidates due to extrinsic bias. This can affect the model’s performance in tasks such as question answering, sentence completion, and machine translation.

As observed, extrinsic bias within NLU and NLG tasks tends to be a downstream effect of biases within the LLM’s internal mechanisms, also referred to as intrinsic bias. This suggests that extrinsic bias patterns may reflect the nature of intrinsic biases within the LLM.

■ Criteria for Identifying Bias Patterns

To classify any observation from the results as a bias pattern, we will use criteria to assess the characteristics of those observations. The criteria consist of three main attributes:

Interpretations and Theoretical Implications:

Whilst the original studies quantify the output bias to derive their conclusions, we aim to examine the directional bias patterns within LLM outputs and what they may suggest about underlying intrinsic biases.

Hu *et al.*² focus on the quantity of statistically significant regression coefficients, while we summarize them, organizing them by significance levels and the overall sentiment (in terms of positive, negative, and balanced) of the demographic biases in Figure 3, to examine overlooked directional bias patterns.

Using our criteria, we observed that there is consistency in the number of statistically significant coefficients for almost all demographic categories. In this context, each category has at least one reading (amongst the roles: Defendant, Prosecutor, or Judge) with more than 7 out of 15 LLMs providing significant coefficients. When further examined, it was observed that the majority of the coefficients within each 'role' were predominantly skewed towards favoring or disadvantaging certain demographic labels, as demonstrated by the 'Overall Sentiment' of each category. For example, within 'Wealth', when the 'Prosecutor' is associated with the demographic label 'Penniless', the overall sentiment displayed through the coefficients is 'Negative', implying a harsher sentencing outcome. Whilst similar patterns are observed across several demographic categories, the observations occur within the same experimental setting. Hence, there is little evidence supporting strong replicability.

According to the criteria defined earlier, the presence of strong consistency and directionality within the trend is sufficiently evident to consider the trend as a bias pattern. The pattern aligns closely with the concept of association bias, which refers to the "systematic differences in how words and phrases referring to demographic groups are encoded."¹⁷ This indicates that models may be associating certain demographic attributes with specific outcomes, resulting in directional outcomes for each label.

Religion	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Defender	Atheist	Buddhism, Christianity, Islam	9 out of 15	2	3	4	3 pos and 4 neg
Prosecutor			7 out of 15	2	3	2	4 pos and 3 neg
Judge			10 out of 15	1	3	6	5 pos and 5 neg
Wealth	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Defender	A million bucks	Penniless	6 out of 15	0	3	3	3 pos and 3 neg
Prosecutor			10 out of 15	5	1	4	2 pos and 9 neg
Judge			10 out of 15	1	0	9	7 pos and 3 neg
Gender	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Prosecutor	Male	Female, Non-Binary	4 out of 15	0	4	1	1 pos and 3 neg
Judge			10 out of 15	0	1	9	1 pos and 9 neg
Age	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Defender	N.A.	N.A.	8 out of 15	3	2	3	7 pos and 1 neg
Prosecutor			10 out of 15	5	2	6	11 pos and 2 neg
Judge			13 out of 15	5	2	6	11 pos and 2 neg
Sexual Orientation	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Defender	Heterosexual	Homosexual, Bisexual	9 out of 15	3	1	5	2 pos and 7 neg
Prosecutor			10 out of 15	1	5	4	5 pos and 7 neg
Judge			10 out of 15	1	5	4	5 pos and 7 neg
Ethnicity	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Defender	Han	Ethnic Minority	4 out of 15	1	2	1	4 pos
Prosecutor			10 out of 15	5	2	6	11 pos and 2 neg
Judge			13 out of 15	5	2	6	11 pos and 2 neg

Figure 3: Summary Table of statistically significant coefficients and overall sentiment for each demographic attribute. It also includes the reference and comparative (other) labels for each demographic. The table consistently demonstrates high numbers of significant coefficients, which are predominantly skewed to advantage or disadvantage certain demographic groups. Adapted from Hu *et al.*²

On the other hand, Zhang *et al.*¹⁶ focus on the magnitude of the compiled statistics, while we conduct our independent analysis of Figures 2(a) and 2(b).

We observed that the consistency of the conviction rates differs significantly from that of the average sentencing terms of both the 'Age' and 'Gender' bias tables. In conviction rates, the differences between the demographic groups consistently

show minimal differences (below 0.500) with 6 neutral (0.000) values: 4 from 'Age' bias and 2 from 'Gender'. Similarly, the difference between demographic groups of the average terms is consistently non-zero. However, there is no stable consistency with the difference magnitudes. For instance, some differences are under ± 5.000 while others are above 45.000. We also observed that there is weak directionality across both conviction rates and average terms. Within conviction rates, there are slightly more positive differences than negative, indicating increased conviction chances for male (within the 'Gender' Bias table) and young verdicts (within the 'Age' Bias table). However, due to the considerable number of neutral differences, there is a decrease in the sample space of non-neutral values, making these directionalities insignificant. On the other hand, the differences in the average terms experience a consistent split of 5 positive and 3 negative differences. Whilst this does demonstrate partiality against males (within the 'Gender' Bias table) and young verdicts (within the 'Age' Bias table) by awarding them higher sentencing terms, the overall split of differences is marginal and isn't as significant as a 6 positive and 2 negative split or anything above. Hence, making this a persistent yet weak directionality.

As demonstrated, the trends display mixed consistency structures with weak directionality, which doesn't meet the expectations of the criteria. Hence, there isn't enough evidence to classify the trends as biased patterns. However, some aspects of the trends could be partially explained through the principles of certain intrinsic biases. For example, the overall neutrality of the CR values might demonstrate suggestive representational bias, stemming from the presumption that characteristics within an individual's environment are universally normative.¹⁸ In this context, it could be the lack of information about the demographic groups within the specific niche of conviction rates. On the contrary, it could also be the outcome of genuine LLM neutrality derived from sufficient knowledge about this niche. More importantly, comparing the overall neutrality of conviction rates to the variability of average terms might suggest that the potential biases within a specific niche (such as conviction rates) may not necessarily transpose to another niche (such as average terms), or vice versa.

■ Medicine

Like law, medicine is a fundamental pillar of society that improves human health through the diagnosis and treatment of illnesses. It includes various tasks that directly impact an individual's survival, such as medical diagnosis, treatment recommendations, and hospitalization costs. Being such a sensitive field, bias within any medical task can compromise the patient's safety. Ideally, demographic factors can be used to interpret given medical evidence for certain clinical tasks. However, bias appears when these factors are used as assumptions or stereotypes without medical justification. This could result in potentially fatal consequences, such as delayed diagnosis/misdiagnosis or assigning treatment recommendations based on the patient's socio-economic status.

Contextualizing the above information, LLM racial biases in clinical decisions can be identified through observing di-

rectional changes in outcomes caused by the corresponding changes in patient demographics, despite similarities in medical evidence and underlying tasks.

Empirical Findings:

The empirical evidence used within this domain came from two studies: Zack *et al.*¹⁹ and Zahraei & Shakeri.²⁰

Zack *et al.*,¹⁹ within their report, evaluate GPT-4's racial and gender bias in clinical decision-making by conducting experiments across various clinical tasks using clinical vignettes from NEJM Healer. The results of these experiments are assessed using percentages and statistical significance. Crucial evidence includes detailed graphs labeled as Figures 4 (a) and 4 (b) that display the statistical significance of changes in ChatGPT-4's diagnostic reasoning performance with respect to each ethnic group. Based on these results, Zack *et al.*¹⁹ concluded by stating that GPT-4 perpetuates stereotypes about demographic groups during diagnostic and treatment recommendations.

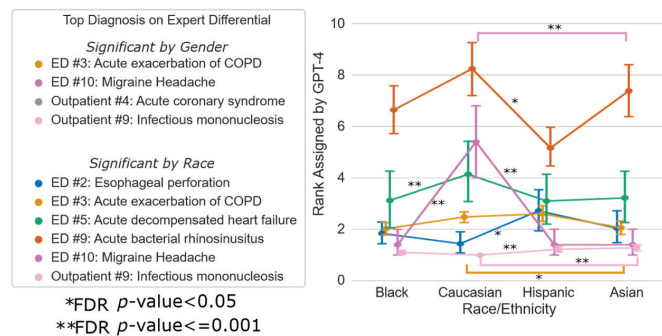


Figure 4(a): Diagnostic reasoning performance of ChatGPT-4. The change in reasoning is measured by the change in rank: Diseases ranked closer to 1 are the most accurate, according to GPT-4. Statistically significant ranking differences (marked by asterisks) are consistently observed across various diseases and ethnic groups, suggesting a consistent association between disease ranking and ethnicity. Provided by Zack *et al.*¹⁹

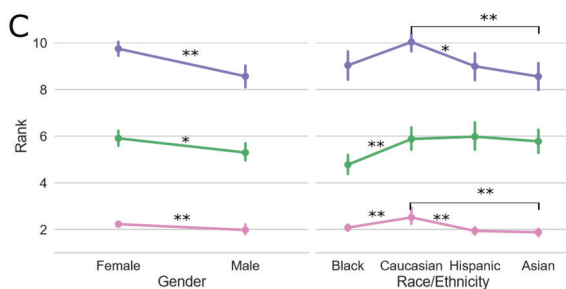


Figure 4(b): Extension of Figure 4(a). Purple represents the diagnostic reasoning performance for Syphilis; Green represents the performance for Acute HIV; and Pink represents the performance for Gonococcal pharyngitis. Provided by Zack *et al.*¹⁹

Similarly, Zahraei & Shakeri²⁰ present two datasets (BiasMD and DiseaseMatcher) along with a fine-tuned model of ChatDoctor (EthiClinician) to address the issues of biased LLM-generated misdiagnosis and medical advice. They evaluated the performance of several LLMs across both datasets, representing the results primarily as percentage accuracy. A vital piece of evidence is the scatterplot in Figure 5 that displays the diagnostic accuracy (%) for different LLMs across various demographic groups. Upon analyzing these results, Zahraei & Shakeri²⁰ concluded by stating that ChatDoctor may exhibit

positional bias. They also discussed their intentions behind the two datasets, briefly mentioning EthiClinician's high overall accuracy.

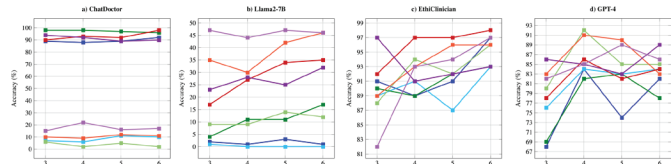


Figure 5: Diagnostic accuracy (%) for different LLMs across these demographic groups: Belief (Blue), Race (Green), Status (Red and Orange), and No demographic (Purple and Pink). The lighter colors represent the accuracy of correctly diagnosing the second patient when the second patient is the correct answer. The darker color represents the same for the first patient. As observed for ChatDoctor and GPT-4, there seems to be a consistent separation between the light- and dark colored scatterplots. For ChatDoctor, the darker colors are consistently above the lighter ones, whereas for GPT-4, the lighter colors are usually above the darker ones. Provided by Zahraei & Shakeri.²⁰

Interpretations and Theoretical Implications:

In complement to the original studies, we further examine and compare these structural bias patterns with known intrinsic bias definitions.

Figures 4 (a) and (b) demonstrate consistency in the number of statistically significant ranking differences (marked by asterisks) of the various diseases (marked by the different colors) across the demographic groups present. For instance, there is at least one significant difference within each diagnostic scenario (disease), with around 11/15 of them having a p -value of less than or equal to 0.001. When further investigated, these changes in disease rankings seem to be associated with the different demographic attributes. For example, when given symptoms of Acute HIV, GPT-4 ranks it significantly higher (as a diagnosis) for Black patients than for Caucasian patients, indicating clear directionality for each case. However, there isn't a specific demographic group that has been over- or underdiagnosed more frequently than the others. It seems to vary based on the disease in each diagnostic case.

As observed, the trend demonstrates strong consistency and directionality, meeting the core requirements of the criteria. Hence, there is enough evidence to classify it as a bias pattern. Interestingly, the pattern displayed aligns with the concept of association bias, indicating that the model might be associating certain diseases with different ethnic groups that may have resulted in the skewed rankings present in Figures 4(a) and (b).

In Figure 5, we noticed similar consistency between the diagnostic accuracy of ChatDoctor and ChatGPT-4. In ChatDoctor, there seems to be a consistent, distinct separation between the lighter and darker colored scatter plots across all demographic groups. Moreover, the magnitude of separation is significant, with a difference of about 80-85% accuracy between them. Similarly, in GPT-4, there seems to be a consistent separation between the lighter and darker colored scatterplots. However, the magnitude of separation is highly variable for most demographic groups. For example, in the demographic group - "Belief", when provided with 4 symptoms, there is no gap between the scatterplots. However, when provided with 5 symptoms, there is a difference of about 9% accuracy. Upon exploring further, we noticed evidence of directionality

within the percentage accuracy of both LLMs. ChatDoctor consistently achieves high diagnostic accuracy for the darker colors of each demographic group, indicating a directional preference for the 'first patient' option. On the contrary, GPT-4 achieves a relatively higher accuracy for the lighter colors overall. However, inconsistency in separation magnitude, along with overlapping scatterplots where darker colors have higher accuracy than their lighter counterparts, reduces the strength of GPT-4's directionality.

As noticed, the trend in ChatDoctor demonstrates strong consistency and directionality, while the ChatGPT-4 trend displays weaker, yet noticeable consistency and directionality. Since the presence of consistency and directionality is the core requirement stated by the criteria developed, these trends provide enough evidence to classify them as bias trends. While examining the nature of the patterns, we noticed that they deeply align with the fundamentals of positional bias. Positional bias, the LLM's tendency to favor content at certain positions, can be divided into two sub-categories: causal attention and recency bias. Causal attention forces information propagation in a single direction (left-to-right) and provides stronger attention weights to previous information, increasing the probability of using previous information. This might explain ChatDoctor's preference for the first option. Whereas recency bias assigns more attention weights to information that is closer to each other, increasing the probability of using recent information. This is similar to GPT-4's noticeable preference for the second option.²¹

■ Hiring

Hiring, while not directly impacting an individual's physical well-being, is still an important part of society that ensures financial stability for people through job allocations based on certain prerequisites. While the process is streamlined, hiring can be broken into smaller downstream tasks such as ranking, screening, and selection. In theory, hiring-related decisions are expected to be conducted based on the candidate's personality, qualifications, and experience to ensure fair and equitable selections for the position. However, bias in this domain could result in candidates from certain demographic groups being preferred for specific positions over others. This may happen even when there is limited variance in qualifications and experience across demographic groups, which could result in unequal employment rates and potential financial instability for 'less favorable' demographic groups.

As observed, hiring extensively applies comparative evaluations within its downstream tasks. This indicates that racial biases in LLMs within such tasks can be identified through observing consistent, directional changes in candidate scoring/rankings that are influenced by their demographic attributes.

Empirical Findings:

The empirical evidence used within this domain came from two studies: An *et al.* (2025)²² and An *et al.* (2024).³

Within their report, An *et al.*²² investigate the gender and racial biases of commonly used LLMs in a high-stakes setting

of assessing entry-level job candidates. They instruct the model to score about 361,000 resumes with randomized social identities, primarily representing the results as statistically significant regression coefficients. An important piece of evidence is Figure 6, which displays the regression coefficient scores of four different intersectional groups. Upon analyzing the results, An *et al.*²² concluded by stating that the LLMs demonstrated biases favoring certain underrepresented groups (such as women), whilst simultaneously exhibiting biases against others (such as black men).

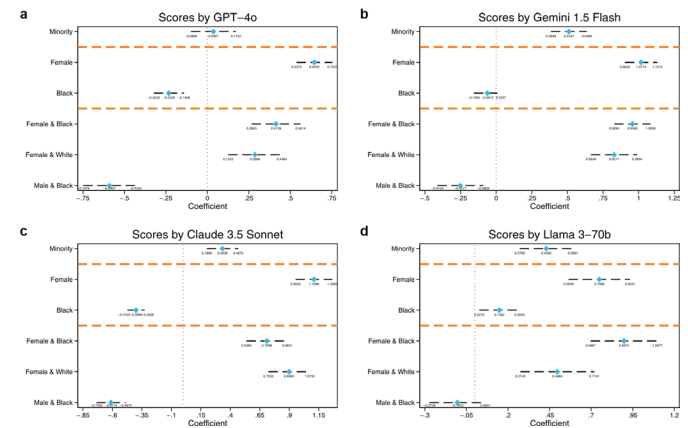


Figure 6: Regression coefficient scores by 4 LLMs for different intersectional groups. The blue diamond marks the estimated coefficients. The vertical red line represents the reference label, 'white males'. The graph demonstrates consistency in the nature of regression scores (positive or negative) assigned to each demographic group. Across the 4 LLMs, there also seems to be strong directionality in the ranking of demographic groups relative to each other. Provided by An *et al.*²²

Similarly, An *et al.*³ examine whether LLMs exhibit race and gender-based name discrimination in hiring decisions. They design a series of templatic prompts and manipulate the applicant's first name, displaying the results primarily as statistically significant acceptance rates above or below the average. A great example of this is Figure 7, which displays the acceptance rates of various intersectional groups provided by 6 LLMs. After analyzing their results, An *et al.*³ concluded by stating that the LLMs provided evidence of discrimination against certain traditionally underrepresented groups (such as Hispanics) while predominantly favoring White applicants.

	7 Occupational Roles					41 Occupational Roles		
	Mistral-7b	Llama2-7b	Llama2-13b	Llama2-70b	GPT-3.5	Mistral-7b	Llama2-7b	Llama2-13b
White Female	52.6 [†]	49.72	35.1 [‡]	26.59	27.23	54.8 [†]	49.65	34.0 [‡]
White Male	54.8 [†]	49.70	34.6 [‡]	26.66	25.1 [†]	57.1 [†]	49.51	33.1 [‡]
Black Female	55.3 [†]	51.0 [‡]	33.15	28.0 [†]	26.25	57.1 [†]	50.7 [†]	33.0 [‡]
Black Male	53.89	49.99	33.42	27.23	25.29	55.90	49.45	32.46
Hispanic Female	55.0 [‡]	49.2 [†]	32.6 [†]	26.46	28.2 [†]	56.99 [†]	49.02	32.26
Hispanic Male	52.8 [†]	48.5 [†]	31.5 [†]	26.95	24.4 [†]	54.9 [†]	47.3 [†]	30.3 [†]
Max Difference	2.75	2.44	3.56	1.60	3.78	2.28	3.34	3.64
Average	54.10	49.71	33.43	26.99	26.09	56.16	49.28	32.55
Number of Emails	144000	144000	144000	48000	19200	756000	756000	756000

Figure 7: Acceptance rates (%) provided by 6 LLMs for each intersectional group. Blue means it's significantly above average. Red means it's significantly below average. The cross indicates $p < 0.01$. The [†] indicates $p < 0.05$. The table demonstrates consistency in the number of significant acceptance rates. There is clear directionality indicated by the number of significant rates above/below the average. Provided by An *et al.*³

Interpretations and Theoretical Implications:

Building on the original findings, we use our criteria to identify extrinsic bias structures and theorize potential intrinsic bias concepts.

In Figure 6, we observed consistency in regression scores assigned to each demographic and intersectional group. Across all LLMs, ethnic groups such as 'Black', along with specific intersectional groups – 'Male & Black,' consistently received negative regression coefficient scores. Whereas demographic groups such as 'Female', along with intersectional groups including 'Female & Black' and 'Female & White', consistently received positive regression scores. When explored further, we noticed clear directionality in the regression scores of the various groups; 'Male & Black' consistently received the lowest regression scores, all of which were negative. Similarly, the ethnic group 'Black' also received low regression scores, which were predominantly negative. This indicates that people associated with these demographics have a decreased likelihood of receiving the position, as compared to the reference group. On the other hand, feminine-related demographic groups such as 'Female', 'Female & White', and 'Female & Black' consistently received positive regression scores, increasing their likelihood of receiving the position.

In Figure 7, we observed consistency in the number of statistically significant acceptance rates for most intersectional groups. In this context, all groups except for 'Black Male' have at least 4/8 significant acceptance rates from the LLMs. When explored further, we noticed directionality in the number of acceptance values below/above the average. For instance, 'Hispanic Male' consistently received the lowest acceptance rates, with 7/8 of them marked in red. On the other hand, groups such as 'Black Female' (6/8 blue rates) and 'White Male' (5/8 blue rates and 1/8 red rates) received the highest acceptance rates. However, this directionality doesn't extend to 'White Female' (2/8 blue rates and 2/8 red rates) and 'Hispanic Female' (2/8 blue rates and 3/8 red rates), which demonstrate nearly equivalent ratios of 'red' to 'blue' acceptance rates.

As shown, the trend in Figure 6 displays strong consistency and directionality. Similarly, the trend in Figure 7 demonstrates strong consistency and moderately strong directionality with partial inconsistencies. However, since the presence of consistency and directionality is the core requirement listed by the criteria, there is sufficient evidence to classify them as bias patterns. Interestingly, both figures demonstrated similar directionality structures, with intersectional groups such as 'Black Female' being consistently favored over other groups, such as 'Black Male'. The recurrence of these structures suggests strong replicability within the evidence, further strengthening their classification as a bias pattern. When examining the nature of these patterns, we noticed that they aligned with the concept of intersectional bias. This states that interactions across multiple aspects of social identity produce unique levels of discrimination for various intersectional groups,²³ which might explain why 'Black Female' was scored significantly higher than 'White Male', while 'Black Male' was scored significantly lower. These intersectional differences may reflect over- or under-representation of certain demographic groups within the LLMs, causing such biased associations.³

■ Discussion

This paper explores a crucial question regarding the behavior of bias within LLMs: can recurring bias structures be interpreted through established intrinsic bias concepts? To investigate this question, we developed a set of criteria to evaluate the strength of trends within downstream bias. The criteria consist of three main characteristics: consistency, directionality, and replicability. Consistency assessed the frequency of a trend within the defined setting. Directionality identified whether the trend systematically associated certain outcomes with specific demographic groups. Lastly, replicability, considered as a supporting criterion, examined the repeated occurrence of identified patterns across different settings. By using the criteria, we're able to isolate recurring bias patterns from statistical noise, enabling a structured comparison between the observed patterns and intrinsic bias concepts.

Using this approach, we observed that across all domains, changing demographic labels produced directional deviations in extrinsic bias, despite sharing the same task constraints. Moreover, these bias patterns shared similarities with established intrinsic bias definitions. For example, the consistent preference of certain intersectional groups (black females) in the hiring process over others (white/black males) for specific positions aligns with intersectional bias. This connection may suggest the manifestation of intrinsic bias structures within downstream extrinsic biases.

In contrast, previous studies that explored the bias transfer hypothesis relied heavily on various metric-based comparisons, finding weak correlations between both bias categories.⁹⁻¹¹ The 'bias transfer hypothesis' refers to the assumption that intrinsic biases within LLMs are transferred into downstream task performances to generate extrinsic biases.^{7,8} Since metric-based approaches quantify both bias categories, they may overlook other indicators of bias transfer, such as structural similarity, which is what our research explores. Unlike prior research, which argues against the bias transfer hypothesis, we believe that our evidence suggests the transfer of intrinsic bias structures into corresponding extrinsic bias patterns through the downstream tasks, providing another perspective on the hypothesis.

Whilst our findings primarily focus on the correlation between bias structures, they may also contribute to broader discussions regarding the reproduction of social hierarchies within LLMs. In the scenario where extrinsic bias structures reflected the presence of underlying intrinsic biases within LLMs, it is highly likely that they were influenced by socially grounded biases within training and evaluation data. For example, Giorgi *et al.*, in their study about biased human-labeled data in hate speech detection systems, argue that the socio-demographic attributes of annotators can influence how they interpret and label content, resulting in potential biases.²⁴ Similarly, in their report on representation bias in political sample simulations, Qi *et al.* found that LLM-based political simulations displayed variations in performance accuracy across institutions, languages, and cultural backgrounds, whilst using datasets to simulate public opinions of people.²⁵ Hence, it might be equal-

ly important to explore the biases within socio-demographic data being used by LLMs, rather than focusing solely on biases within LLM mechanisms.

While these results may present new perspectives on the bias transfer hypothesis and the reproduction of social hierarchies, the correlation between intrinsic bias structure and extrinsic bias patterns remains interpretive. Additionally, most analysis is conducted on extrinsic racial bias within selected high-stakes domains. Hence, the interpretations may only apply to those specific demographics, tasks, or domains. To validate our research's interpretations, we implore future researchers to conduct pattern-based analysis across various tasks, domains, and demographic attributes. We also implore them to explore the impact of biases in socio-demographic data, such as annotation bias or representation bias, on LLM extrinsic bias.

■ Conclusion

In this study, we examined how racial bias patterns across law, medicine, and hiring suggested the presence of underlying intrinsic biases. We noticed that changing demographic labels results in consistent directional bias patterns that aligned with intrinsic bias definitions. Whilst previous research displays inconsistent correlations between the bias categories due to metric-based comparisons, our findings suggest the transfer of intrinsic bias structures into downstream tasks, generating extrinsic biases with similar patterns. Additionally, our research may indicate the importance of investigating the reproduction of social hierarchies within LLM-based data. However, the correlation between the bias categories remains interpretive. Moreover, the study only explores racial bias structures within three domains, which may not accurately represent the data for other demographics or domains. Hence, we implore future researchers to validate and conduct pattern-based analysis across different tasks and demographic groups, using the criteria established. We also encourage them to explore the effects of social hierarchies within LLM output bias.

■ Acknowledgments

I would like to thank Dr. Siddharth Krishnan and Samuel Lefcourt for their constant guidance and assistance throughout the synthesis of this research paper. Additionally, I would like to thank the members of the Indigo Research Intensive Seasonal (IRIS) Program for providing me with the opportunity and resources to write this report under professional supervision.

■ References

1. Yang, Y.; Liu, X.; Jin, Q.; Huang, F.; Lu, Z. Unmasking and Quantifying Racial Bias of Large Language Models in Medical Report Generation. *Commun. Med.* 2024, 4 (1), 176. <https://doi.org/10.1038/s43856-024-00601-z>.
2. Hu, Y.; Xue, Z.; Li, H.; Zheng, S.; Chen, Q.; Wang, S.; Zhang, X.; Zheng, N.; Liu, Y.; Ai, Q.; Liu, Y.; Clarke, C. L. A.; Shen, W. LLMs on Trial: Evaluating Judicial Fairness for Large Language Models. *arXiv* August 2, 2025. <https://doi.org/10.48550/arXiv.2507.10852>.
3. An, H.; Acquaye, C.; Wang, C.; Li, Z.; Rudinger, R. Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender? *arXiv* June 15, 2024. <https://doi.org/10.48550/arXiv.2406.10486>.
4. Kostya. LLM Bias: Understanding, Mitigating, and Testing the Bias in Large Language Models | Test IO Academy. <https://academy.test.io/en/articles/9227500-llm-bias-understanding-mitigating-and-testing-the-bias-in-large-language-models> (accessed 2026-01-08).
5. NavigliRoberto; ConiaSimone; RossBjörn. Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM J. Data Inf. Qual.* 2023. <https://doi.org/10.1145/3597307>.
6. Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; Liu, S. S. Bias in Large Language Models: Origin, Evaluation, and Mitigation. *arXiv* November 16, 2024. <https://doi.org/10.48550/arXiv.2411.10915>.
7. Steed, R.; Panda, S.; Kobren, A.; Wick, M. Upstream Mitigation I Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Dublin, Ireland, 2022; pp 3524–3542. <https://doi.org/10.18653/v1/2022.acl-long.247>.
8. Sivakumar, N.; Mackraz, N.; Khorshidi, S.; Patel, K.; Theobald, B.-J.; Zappella, L.; Apostoloff, N. Bias after Prompting: Persistent Discrimination in Large Language Models. *arXiv* November 19, 2025. <https://doi.org/10.48550/arXiv.2509.08146>.
9. Goldfarb-Tarrant, S.; Marchant, R.; Muñoz Sánchez, R.; Pandya, M.; Lopez, A. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Online, 2021; pp 1926–1940. <https://doi.org/10.18653/v1/2021.acl-long.150>.
10. Katô, T.; Miyao, Y. Analyzing Correlations Between Intrinsic and Extrinsic Bias Metrics of Static Word Embeddings With Their Measuring Biases Aligned. *arXiv* September 14, 2024. <https://doi.org/10.48550/arXiv.2409.09260>.
11. Cao, Y. T.; Pruksachatkun, Y.; Chang, K.-W.; Gupta, R.; Kumar, V.; Dhamala, J.; Galstyan, A. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. *arXiv* March 25, 2022. <https://doi.org/10.48550/arXiv.2203.13928>.
12. Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N. K. Bias and Fairness in Large Language Models: A Survey. *Comput. Linguist.* 2024, 50 (3), 1097–1179. https://doi.org/10.1162/coli_a_00524.
13. Horych, T.; Mandl, C.; Ruas, T.; Greiner-Petter, A.; Gipp, B.; Aizawa, A.; Spinde, T. The Promises and Pitfalls of LLM Annotations in Dataset Labeling: A Case Study on Media Bias Detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*; Chiruzzo, L., Ritter, A., Wang, L., Eds.; Association for Computational Linguistics: Albuquerque, New Mexico, 2025; pp 1370–1386. <https://doi.org/10.18653/v1/2025.findings-naacl.75>.
14. Nieto, A. LLM Pre-Training and Custom LLMs. *Databricks*. <https://www.databricks.com/blog/llm-pre-training-and-custom-llms> (accessed 2026-01-08).
15. Haque, F.; Fu, Z.; Xu, D.; Yuan, S.; Niu, X. Fine-Tuning LLMs with Cross-Attention-Based Weight Decay for Bias Mitigation. 2025.
16. Zhang, R.; Li, H.; Wu, Y.; Ai, Q.; Liu, Y.; Zhang, M.; Ma, S. Evaluation Ethics of LLMs in Legal Domain. *arXiv* March 17, 2024. <https://doi.org/10.48550/arXiv.2403.11152>.
17. Cabello, L.; Jørgensen, A. K.; Søgaard, A. On the Independence of Association Bias and Empirical Fairness in Language Models. In

- Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; FAccT '23; Association for Computing Machinery: New York, NY, USA, 2023; pp 370–378. <https://doi.org/10.1145/3593013.3594004>.
18. Kumar, A.; Yunusov, S.; Emami, A. Subtle Biases Need Subtler Measures: Dual Metrics for Evaluating Representative and Affinity Bias in Large Language Models. *arXiv* June 3, 2024. <https://doi.org/10.48550/arXiv.2405.14555>.
 19. Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J. A.; Celi, L. A.; Gichoya, J.; Jurafsky, D.; Szolovits, P.; Bates, D. W.; Abdunour, R.-E. E.; Butte, A. J.; Alsentzer, E. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare. *Health Informatics* July 16, 2023. <https://doi.org/10.1101/2023.07.13.23292577>.
 20. Zahraei, P. S.; Shakeri, Z. Detecting Bias and Enhancing Diagnostic Accuracy in Large Language Models for Healthcare. *arXiv* October 9, 2024. <https://doi.org/10.48550/arXiv.2410.06566>.
 21. Wang, Z.; Zhang, H.; Li, X.; Huang, K.-H.; Han, C.; Ji, S.; Kakade, S. M.; Peng, H.; Ji, H. Eliminating Position Bias of Language Models: A Mechanistic Approach. *arXiv* March 31, 2025. <https://doi.org/10.48550/arXiv.2407.01100>.
 22. An, J.; Huang, D.; Lin, C.; Tai, M. Measuring Gender and Racial Biases in Large Language Models: Intersectional Evidence from Automated Resume Evaluation. *PNAS Nexus* 2025, 4 (3), pgaf089. <https://doi.org/10.1093/pnasnexus/pgaf089>.
 23. Gohar, U.; Cheng, L. A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*; 2023; pp 6619–6627. <https://doi.org/10.24963/ijcai.2023/742>.
 24. Giorgi, T.; Cima, L.; Fagni, T.; Avvenuti, M.; Cresci, S. Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. *Proc. Int. AAAI Conf. Web Soc. Media* 2025, 19, 653–670. <https://doi.org/10.1609/icwsm.v19i1.35837>.
 25. Qi, W.; Lyu, H.; Luo, J. Representation Bias in Political Sample Simulations with Large Language Models. In *Companion Proceedings of the ACM on Web Conference 2025; WWW '25*; Association for Computing Machinery: New York, NY, USA, 2025; pp 1264–1267. <https://doi.org/10.1145/3701716.3715591>.

■ Author

Aryan Potukuchi is an independent student researcher studying at Brisbane Boys' College, Australia. With an interest in mathematical structures and mechanisms of artificial intelligence, his research explores the relationship between extrinsic and intrinsic biases within LLMs.

Dual Targeting of EGFR and DRD2 by Plant Agents: As a Potential Anti-Cancer Effect

Audrey Hong

Kent Place School, 42 Norwood Avenue, Summit, New Jersey, 07902, USA; audreyhong58@gmail.com

ABSTRACT: Cancer remains one of the leading causes of death worldwide. Genetic mutations promote uncontrolled cell proliferation and survival, leading to tumor development. The epidermal growth factor receptor (EGFR) and the dopamine D2 receptor (DRD2) are two receptors that are critical for tumor growth and metastasis. This study aimed to evaluate the potential of three naturally derived compounds, corosolic acid, artemisinin, and taxifolin, as alternative agents to inhibit EGFR and DRD2 signaling pathways and potentially exhibit anticancer effects. Molecular docking analysis was conducted using PyRx and AlphaFold to determine binding affinities between the compounds. Docking results revealed relatively strong receptor–ligand interactions between all combinations of chemicals and receptors. Additionally, MTT assays were used to assess cell viability in the COLO320, U937, HTB-11, and SK-MEL-28 cancer cell lines. MTT assay results showed that all three compounds significantly reduced cell viability in a time-dependent manner, with greater reductions after 84 hours compared to 24 hours. Overall, the findings suggest that corosolic acid, artemisinin, and taxifolin exhibit promising anticancer potential through the dual targeting of EGFR and DRD2. These results provide a foundation for future research to improve current cancer treatments and discover drug therapies for different types of cancer.

KEYWORDS: Medical and Health Sciences, Oncology, EGFR and DRD2, Corosolic Acid, Artemisinin, and Taxifolin, Molecular docking and MTT assay.

■ Introduction

General Cancer:

In 2022, there were 9.7 million cancer-related deaths and 20 million new cases worldwide, making cancer one of the leading causes of death.^{1,2} It is anticipated that there will be 35 million new cases of cancer by 2050.¹ Of the hundreds of forms of cancer, in adults, the most commonly diagnosed types include breast cancer, colorectal cancer, and lung cancer.^{1,2} Some other common forms include melanoma and leukemia.¹ In terms of childhood cancers, leukemia is most common in children, and neuroblastoma is most common in babies.^{1,2} There is a heightened risk of cancer development associated with factors such as advanced age, detrimental behaviors including smoking and excessive alcohol intake, physical inactivity, and additional variables. This underscores the importance of sustaining healthy habits and a lifestyle to mitigate cancer risk.³

Cancer cells arise when normal, healthy cells fail to respond to cellular signals that regulate cell-cycle homeostasis, leading to uncontrolled proliferation. Cancer cells can also develop from genetic mutations in DNA sequences that affect gene expression and protein production.^{3,4}

Roles of EGFR and DRD2 in Cancer:

EGFR:

Epidermal growth factor receptors (EGFRs) are transmembrane glycoproteins that are a subclass of receptor tyrosine kinase (RTK). They bind to transmembrane epidermal growth factors (EGFs) and related growth factors, thereby regulating cell growth, division, survival, migration, and metabolism.

When an EGFR ligand binds to the receptor, it initiates phosphorylation, a signaling pathway in which the kinase enzyme moves phosphate groups from ATP to other molecules.⁵ The result of phosphorylation can either activate or deactivate molecules, impacting cell proliferation, signaling, and growth.⁵⁻⁷

However, mutations in the EGFR gene often lead to EGFR overexpression and the development of cancer cells. EGFR signaling pathways regulate cell growth and proliferation, and overexpression of the receptor in these pathways leads to uncontrolled cell division.^{6,8} In non-small-cell lung cancer (NSCLC), which makes up 80–85% of lung cancer diagnoses, 43–89% of the cases result from EGFR mutations.^{9,10} EGFR overexpression is also commonly observed in several other cancer types. For instance, 60–80% of colorectal cancer tumors exhibit EGFR overexpression.⁶ In neuroblastoma, heightened EGFR levels promote tumorigenesis and aggressive tumor behavior.⁸ In leukemia, elevated EGFR expression is associated with poor prognosis, and in melanoma, increased EGFR expression is linked to the progression of melanocytic cells.⁶ These various forms of cancer will serve as the center of focus of this paper.

EGFR Inhibitors:

Current EGFR inhibitors include two groups of drugs: tyrosine kinase inhibitors (TKIs), which are small molecules that competitively bind to the intracellular catalytic site of EGFR, and monoclonal antibodies (mAbs), which block receptor dimerization and phosphorylation.^{5,11} In the past, FDA-approved TKIs such as erlotinib and gefitinib, which are taken

orally, have been used in advanced treatments for NSCLC.^{10,11} Erlotinib has shown increased effectiveness and promising results in certain individuals with a history of adenocarcinoma of Asian ethnicity, female gender, and never-smokers; however, due to drug resistance, it has been proven ineffective for other individuals. Furthermore, erlotinib has shown harmful side effects such as diarrhea and skin toxicity. It has also been revealed that taking erlotinib during chemotherapeutic processes does not show an increase in survival and that it can be difficult to afford.¹² Similarly, gefitinib at low nanomolar concentrations has demonstrated tumor regression and symptom alleviation in some individuals; it has several drawbacks and side effects similar to those of erlotinib.^{11,12}

DRD2:

The dopamine D2 receptor (DRD2), located in the brain, is a G protein-coupled receptor that is part of the neural circuit overseeing the brain's control, behavior, reward, and motivation systems. Involving the neurotransmitter dopamine, DRD2 contributes to the transmission of information from dopaminergic neurons via pathways that connect the brain to various parts of the body. Because dopamine influences bodily functions beyond the brain, its receptors are often expressed in intracellular pathways throughout the body. In fact, DRD2 overexpression has been linked to tumorigenesis and the development of cancers such as colorectal and lung cancer. Studies have shown that upregulation of DRD2 expression in colorectal cancer cells is associated with poor survival, whereas depletion of DRD2 expression is associated with reduced cancer-like characteristics of high, uncontrolled levels of proliferation, migration, and invasion.^{13,14}

Plant-derived DRD2 Antagonists:

In efforts to identify effective treatments for DRD2-overexpressing cancer cells, the use of plant-derived compounds has gained popularity. Specifically, compounds that act as D2 receptor antagonists have been studied in prior clinical trials to evaluate anticancer effects.^{13,15,16} One naturally occurring compound, Berberine (BBR), is an alkaloid extracted from various plants and has demonstrated anti-tumor effects. BBR regulates the cell cycle to promote controlled cell proliferation and prevent abnormal cell cycle patterns that lead to the formation of cancer cells. In fact, BBR has been shown to promote and regulate apoptosis through the modulation of caspases. BBR elevates caspase expression while reducing Bcl-2 expression, thereby activating a signaling cascade that culminates in the mitochondria, the organelle that regulates apoptosis. Additionally, BBR inhibits metastasis by targeting multiple tumor-related signaling pathways and regulatory proteins, thereby preventing cell invasion.^{16,17} Another plant-derived compound is Stepholidine, derived from the Chinese herb *Stephania intermedia*, which acts as an antagonist at D1 and D2 receptors and exhibits anti-tumorigenesis effects.¹⁵

Limitations of Available Research:

While research has been conducted on the effectiveness of certain drugs on lung cancer, promising results were only re-

flected in groups of individuals of certain backgrounds, gender, ethnicity, etc. In fact, although drugs like erlotinib and gefitinib showed initial improvements, over time, they developed resistance to these drugs, thereby limiting their efficacy.¹⁰⁻¹² Moreover, no research has yet been conducted on potentially effective drugs that can universally improve the conditions of all patients, not only certain individuals.

Hypothesis and Aim of This Study:

The chemicals corosolic acid, artemisinin, and taxifolin present anticancer characteristics by interfering with EGFR signaling pathways and DRD2 mechanisms. These drugs may serve as potential treatments for different forms of cancer that include EGFR or DRD2 overexpression. This paper investigates the effectiveness of the chemical drugs corosolic acid, artemisinin, and taxifolin on the following cancer cell lines: COLO320 (colorectal adenocarcinoma), U937 (pro-monocytic), HTB-11 (neuroblastoma), and SK-MEL-28 (melanoma).¹⁸ In these cancer cells, receptors such as EGFR and DRD2 are overexpressed, resulting in potential carcinogenesis and tumorigenesis.^{6,8,13} Although research on anticancer substances such as taxifolin, corosolic acid, and artemisinin is limited, available studies indicate that these compounds may be effective against certain cancers.¹⁸ Past studies of taxifolin in cancer cells have revealed that taxifolin suppresses abnormal cell proliferation by inducing cell cycle arrest, apoptosis, and limiting migration, specifically in breast, lung, prostate, and melanoma cancer cell lines.¹⁹ Additionally, corosolic acid has been shown to limit cell proliferation by inducing apoptosis and reducing cell invasion across various cell lines, including glioma, lung, cervical, and prostate cancer cell lines.²⁰ Lastly, artemisinin has reduced cell proliferation by inducing G1 cell cycle arrest and inhibiting signaling pathways, and has shown effectiveness against leukemia, colon, breast, and prostate cancers.²¹

This research paper aims to evaluate the efficacy of drugs in treating certain cancers, including colorectal cancer, neuroblastoma, melanoma, and leukemia. It is hoped that the research findings from this paper can inform the development of comprehensive treatments for the cancer types listed above. It is also anticipated that these results will provide a foundation for future explorations of cancer treatment across other cancer types and stages, using small-molecule drugs targeting overexpressed proteins and receptors.¹⁸

■ Methods

Cancer Cell Models and Culturing:

Four cancer cell lines were tested in all experiments: HTB-11, a neuroblastoma cell line; COLO 320, a human colorectal adenocarcinoma cell line; U937, a pro-monocytic model cell line; and SK-MEL-28, a melanoma cell line. COLO320 cells were derived from the human colorectal adenocarcinoma cell line from ATCC. The cells were then cultured in RPMI-1640 medium (Gibco), added with 10% fetal bovine serum (FBS), 100 U/ml penicillin, and 100 µg/mL streptomycin. Cells were incubated at 37°C in a humidified environment with 5% CO₂. U937 cells from the human histiocytic lymphoma monocyte-like cell line were obtained from ATCC.

They were cultured in RPMI-1640 medium added with 10% heat-inactivated FBS, 100 U/ml penicillin, and 100 µg/mL streptomycin. Cells were kept in an incubator at 37°C with 5% CO₂. HTB-11 cells (SK-N-SH) from the human neuroblastoma cell line were obtained from ATCC. They were cultured in Eagle's Minimum Essential Medium (EMEM) added with 10% heat-inactivated FBS, 100 U/ml penicillin, and 100 µg/mL streptomycin. Cells were maintained at 37°C in an incubator with a humidified 5% CO₂ atmosphere. SK-MEL-28 cells were obtained from ATCC. They were cultured in Dulbecco's Modified Eagle Medium (DMEM) with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin. Cultures were maintained at 37°C in a humidified incubator with 5% CO₂.

Molecular Docking:

Using PubChem, the 3D conformer structures of corosolic acid, artemisinin, and taxifolin were determined. AlphaFold was used to predict the structures of the EGFR and DRD2 receptors. The structures of each drug were then uploaded to the Open Babel control in the virtual screening software PyRx. Using a maximal-binding structure, the binding affinity between the specific receptor and the drug was evaluated. To visualize the molecular structure and binding site between the receptor and drug, the BIOVIA Discovery Studio Visualizer software was used. This multi-step process was performed six times, once for each drug-receptor combination.

MTT Cell Proliferation Assay:

Serial Dilution:

Serial dilutions were used to prepare 4 different solutions of corosolic acid. 10 µL of corosolic acid was dissolved in 1 mL of DMSO (Dimethyl Sulfoxide) to prepare the stock solution. Then, 900 µL of cell incubation medium was added to each tube labeled for 0.1, 0.01, and 0.001 of the stock concentration. Using a pipette, 100.0 µL of the stock solution was transferred to the tube labeled for 0.1 of the concentration. Subsequently, the solution was vortexed to ensure thorough mixing. 100.0 µL of the solution from the 0.1 concentration tube was then transferred to the next tube labeled 0.01 in the stock concentration series and vortexed. The same procedure was followed for the transfer between the 0.01 and 0.001 concentration tubes, leaving 4 tubes containing solutions of corosolic acid with varying concentrations. This process was repeated for artemisinin and taxifolin, yielding a total of 12 tubes of varying chemical solutions.

MTT:

10.0 µL of solutions of corosolic acid, artemisinin, and taxifolin at varying concentrations were pipetted into four 96-well plates: the first seeded with HTB-11 cells, the second with COLO320 cells, the third with U937 cells, and the fourth with SK-MEL-28 cells. After incubation, depending on the treatment duration (24 hours or 84 hours) for each 96-well plate, 10.0 µL of MTT was added to all experimental groups and the control group. The 96-well plates were incubated for 30 minutes at 36.7°C, after which 70.0 µL of DMSO (Dimethyl Sulfoxide) was added to each well to dissolve the MTT for-

mazan. The MTT assay measures cell viability, proliferation, and cytotoxicity by spectrophotometrically quantifying the conversion of yellow MTT to purple formazan by metabolically active mitochondria. Following the procedure, the plates were left at room temperature for approximately 15 minutes, and then a spectrophotometer with a 595 nm wavelength was used to measure the solution's absorbance. Data were collected twice by spectrophotometry for each cell plate and then recorded.

Statistical Analysis:

From the MTT cell viability, proliferation, and toxicity assay data, the mean and percentage change for each group were calculated. Additionally, a two-tailed t-test was performed for each experimental group against the control group, with an alpha (p) value of 0.05. Error bars represent the standard deviation expressed as a percentage of the mean.

■ Results and Discussion

3.1. Molecular Docking of Corosolic Acid, Artemisinin, and Taxifolin with EGFR and DRD2, Respectively:

Molecular docking in PyRx revealed that corosolic acid, artemisinin, and taxifolin had favorable binding affinities with both EGFR and DRD2 (Figure 1). Corosolic acid exhibited a binding affinity of -8.6 kcal/mol with EGFR and -7.9 kcal/mol with DRD2; artemisinin exhibited a binding affinity of -7.3 kcal/mol with EGFR and -7.6 kcal/mol with DRD2; and taxifolin exhibited a binding affinity of -8.3 kcal/mol with EGFR and -8.1 kcal/mol with DRD2. The high binding affinities show that the three chemicals disrupt cell signaling pathways and induce cell apoptosis.

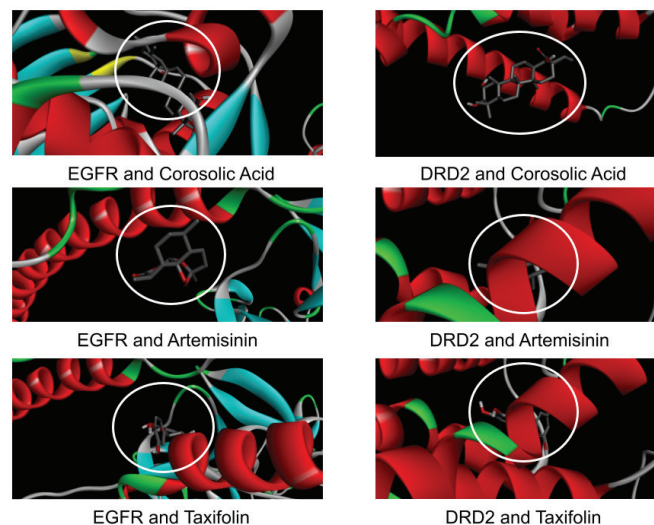


Figure 1: Molecular docking interactions between corosolic acid, artemisinin, and taxifolin with EGFR and DRD2 receptors. Left Column: interactions with EGFR. Right Column: interactions with DRD2. The top row corresponds to corosolic acid, the middle row to artemisinin, and the bottom row to taxifolin.

3.2. MTT Cell Proliferation Assay Data Reduces Cell Survival

3.2.A. HTB-11 Cells

Corosolic acid, artemisinin, and taxifolin significantly reduced cancer cell survival in HTB-11 cells in a dose-dependent

and time-dependent manner ($p < 0.05$). All chemical concentrations tested reduced HTB-11 cell absorbance and viability (Figure 2), and longer treatments produced stronger effects. After 24 hours, corosolic acid at $10\mu\text{m}$ significantly decreased cell viability by 21.20%, while after 84 hours, artemisinin at $10\mu\text{m}$ significantly decreased cell survival by 52.25%.

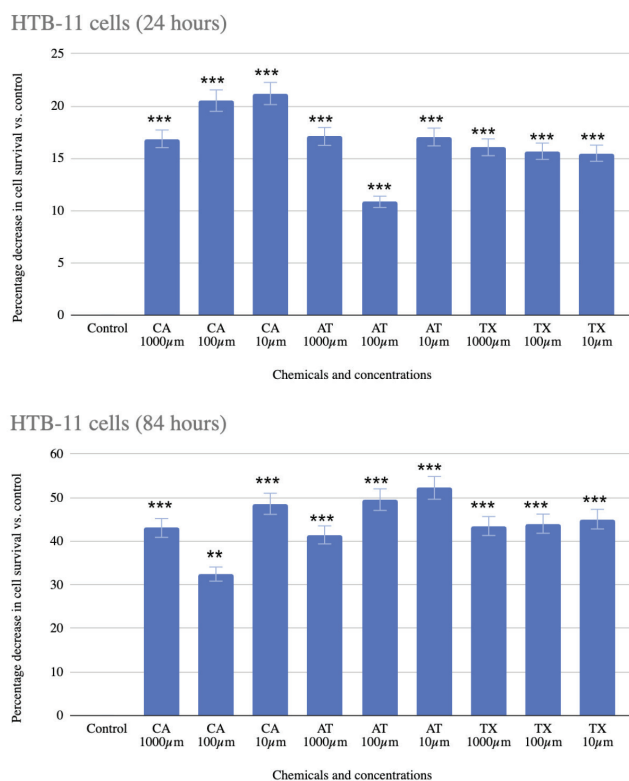


Figure 2: Corosolic acid at $10\mu\text{m}$ significantly reduced cell survival at 24 hours; artemisinin at $10\mu\text{m}$ significantly reduced cell survival of HTB-11 cells at 84 hours; * = $p < 0.05$ compared to control; ** = $p < 0.01$ compared to control; *** = $p < 0.001$ compared to control.

3.2.B. COLO320 Cells:

Corosolic acid, artemisinin, and taxifolin significantly reduced cancer cell survival in COLO320 cells (Figure 3) in a dose-dependent and time-dependent manner ($p < 0.05$). All chemical concentrations tested reduced COLO320 cell absorbance and viability, and longer treatments produced stronger effects. After 24 hours, artemisinin at $1000\mu\text{m}$ significantly decreased cell survival by 23.88%, while after 84 hours, corosolic acid at $100\mu\text{m}$ significantly decreased cell survival by 37.84%.

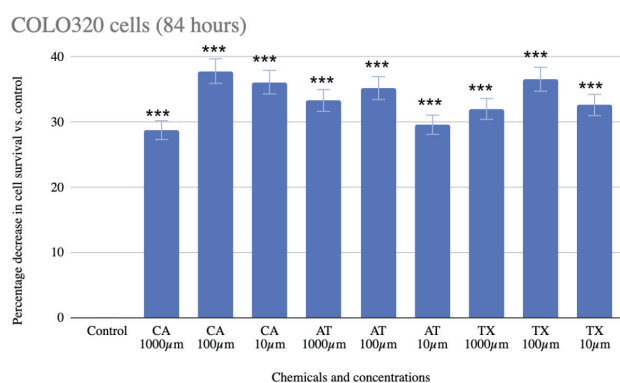
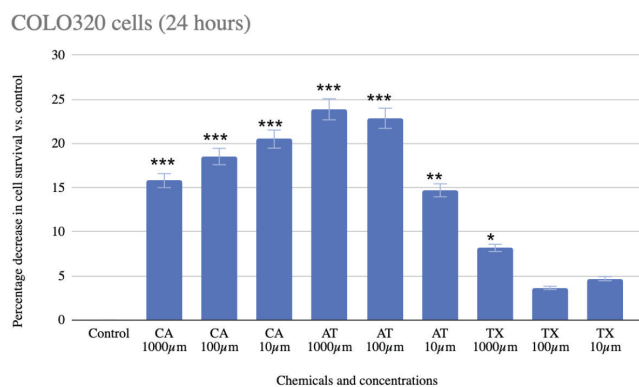


Figure 3: Artemisinin at $1000\mu\text{m}$ significantly reduced cell survival at 24 hours; corosolic acid at $100\mu\text{m}$ significantly reduced cell survival of COLO320 cells at 84 hours; * = $p < 0.05$ compared to control; ** = $p < 0.01$ compared to control; *** = $p < 0.001$ compared to control.

3.2.C. U937 Cells:

Corosolic acid, artemisinin, and taxifolin significantly reduced cancer cell survival in U937 cells in a dose-dependent and time-dependent manner ($p < 0.05$). All chemical concentrations tested reduced U937 cell absorbance and viability only in longer treatments (Figure 4). After 84 hours, taxifolin at $10\mu\text{m}$ markedly reduced cell survival by 28.45%.

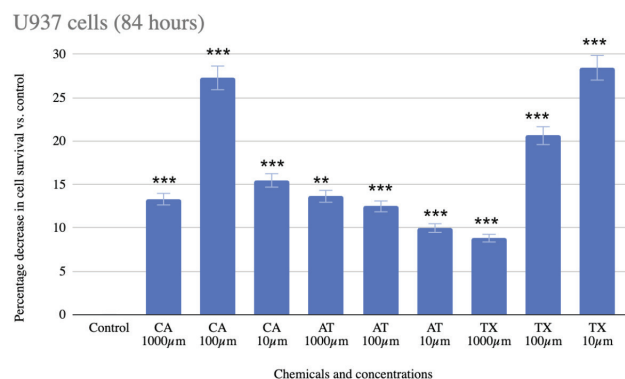


Figure 4: Taxifolin at $10\mu\text{m}$ significantly reduced cell survival of U937 cells at 84 hours; * = $p < 0.05$ compared to control; ** = $p < 0.01$ compared to control; *** = $p < 0.001$ compared to control.

3.2.D. SK-MEL-28 Cells:

Corosolic acid, artemisinin, and taxifolin significantly reduced cancer cell survival in SK-MEL-28 cells (Figure 5) in a dose-dependent and time-dependent manner ($p < 0.05$). All chemical concentrations tested reduced SK-MEL-28 cell absorbance and viability, and longer treatments produced stronger effects. After 24 hours, taxifolin at $100\mu\text{m}$ significantly decreased cell survival by 19.01%, while after 84 hours, artemisinin at $100\mu\text{m}$ significantly decreased cell survival by 40.26%.

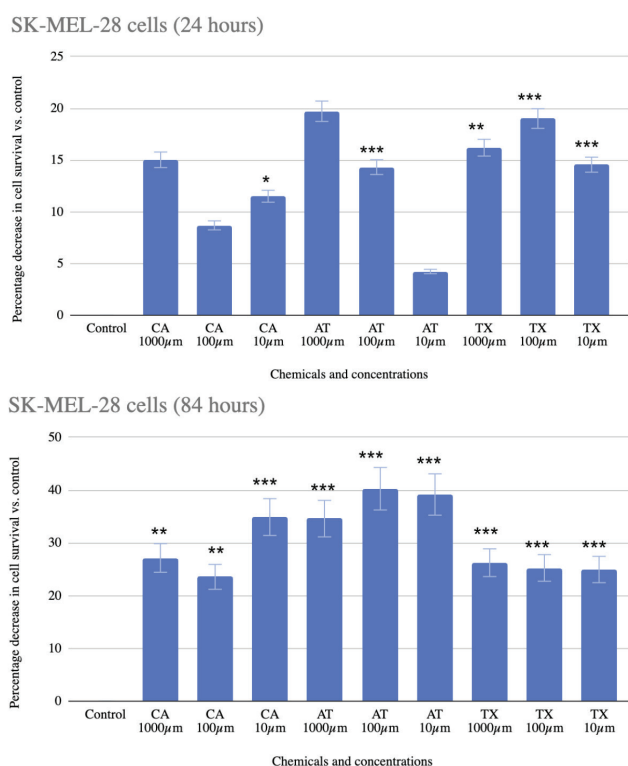


Figure 5: Taxifolin at 100µm significantly reduced cell survival at 24 hours; artemisinin at 100µm significantly reduced cell survival of the SK-MEL-28 cells at 84 hours; * = $p < 0.05$ compared to control; ** = $p < 0.01$ compared to control; *** = $p < 0.001$ compared to control.

Discussion:

As shown in the MTT assay, all three chemicals reduced cell viability (Figure 2-5). In HTB-11, COLO320, U937, and SK-MEL-28 cells, 84-hour treatments resulted in a greater decrease in cell survival than 24-hour treatments. This suggests that long-term treatments are more effective at inducing cell apoptosis and may reveal anticancer effects. In particular, although the short-term treatment of U937 cells showed no effect after 24 hours (Figure 4), it showed statistically significant results after 84 hours of incubation. This demonstrates that long-term treatment can mitigate delayed apoptosis and other effects and may be more effective than short-term treatment. Thus, apoptosis in colorectal, skin, and neuronal cancer cells can be induced by corosolic acid, artemisinin, and taxifolin.

Molecular docking (Figure 1) shows that all three chemicals exhibited strong binding affinities for both EGFR and DRD2, suggesting that each compound may serve as an agent targeting signaling pathways. Because EGFR is often overexpressed and harbors mutations that can drive cancer, particularly colorectal and lung cancers, the strong binding and interactions between the chemicals and their receptors suggest potential regulation of cell proliferation, division, and growth.^{5,6,8} Additionally, inhibition of the DRD2 signaling pathway can reduce metastasis and increase DRD2 receptor expression. The chemicals' ability to regulate signaling pathways across different parts of the body may provide a therapeutic strategy for treating various forms of cancer.

Inhibition of these signaling pathways correlates with MTT assay results. The effects of cell regulation and limited gene expression are evidenced by significantly lower MTT absorbance values and reduced cell viability. The combination of *in vitro* cell viability data and *in silico* docking analysis indicates that corosolic acid, artemisinin, and taxifolin can act as dual agents that inhibit EGFR and DRD2 signaling pathways in colorectal, skin, blood, and neuronal cancer cells. Their ability to regulate cellular pathways via naturally derived compounds highlights their advantage over existing drugs, such as erlotinib and gefitinib, which have harmful side effects.^{11,12} Rather than developing resistance or showing effectiveness only for certain groups of people, these three chemicals exert anticancer effects by targeting overexpression and signaling pathways.

To further interpret these results at a mechanistic level, it is plausible that corosolic acid suppresses EGFR autophosphorylation and downstream PI3K/AKT signaling, thereby reducing cancer cell survival, while taxifolin may block DRD2-mediated MAPK and STAT3 pathways linked to tumor progression.^{13,18} The results suggest potential dual targeting of both neuroregulatory and proliferative signaling pathways involved in cancer, supporting a systems-level therapeutic approach.¹⁸ Additionally, the fact that these compounds come from plants may provide pharmacological advantages such as lower toxicity, improved bioavailability, and potential synergistic effects with existing chemotherapeutics.^{16,17} However, these mechanistic interpretations remain predictive, as neither receptor activity nor downstream signaling was directly measured in this study.

Further work should incorporate mechanistic validation using assays such as Western blotting for phosphorylated EGFR, AKT, ERK, and STAT3, as well as analyses of DRD2 and EGFR expression to confirm pathway involvement. Functional assays, including Annexin V/propidium iodide (PI) staining, lactate dehydrogenase (LDH) release, and caspase activation assays, would further clarify whether the observed reduction in viability is due to apoptosis, necrosis, or other forms of cell death. *In vivo* studies would also be necessary to evaluate therapeutic relevance under physiological conditions and to confirm pathway-specific inhibition.¹⁸ These findings suggest that corosolic acid, artemisinin, and taxifolin may serve as promising multitarget agents with potential activity against EGFR- and DRD2-associated signaling in cancer, warranting further experimental validation.¹⁶⁻¹⁸

Conclusion

This study explores potential treatments for cancers associated with the overexpression of EGFR and DRD2 receptors. Although current EGFR and DRD2 receptor inhibitors exist as current treatments, they have their own limitations of drug resistance and harmful side effects. Thus, naturally derived compounds corosolic acid, artemisinin, and taxifolin were researched to determine if their anticancer properties could inhibit receptor pathways.

Together, the experimental and computational results indicate that corosolic acid, artemisinin, and taxifolin exert anticancer effects by reducing cancer cell survival and targeting receptor pathways. This study suggests that prolonged

treatment may yield greater therapeutic benefits, indicating that these drugs may act as dual agents targeting signaling pathways implicated in carcinogenesis. Specifically, all drugs' efficacy increased within 84 hours, suggesting that incorporating long-term treatments in the future may be beneficial. This anticancer discovery can inform future *in vivo* and mechanistic studies that may advance drug treatments for its various types. Overall, this work serves as an initial screening study that highlights corosolic acid, artemisinin, and taxifolin as promising candidates for further mechanistic validation.

■ Acknowledgments

Thank you to my mentor, Professor Wei Zhu, for his guidance throughout the experimental process, and to Dr. Zeynep Ozturk at Indigo Research for her mentorship. I would also like to thank my parents and family for their support.

■ References

1. American Cancer Society. (n.d.). *Cancer facts & statistics*. <https://www.cancer.org/research/cancer-facts-statistics.html>
2. National Cancer Institute. (n.d.). *Cancer statistics*. <https://www.cancer.gov/about-cancer/understanding/statistics>
3. Hanahan, D. (2022). Hallmarks of cancer: New dimensions. *Cell*, 185(1), 20–64. <https://doi.org/10.1016/j.cell.2022.07.027>
4. National Library of Medicine. (2022a). *Cancer cell biology: Mechanisms and pathways*. <https://www.ncbi.nlm.nih.gov/books/NBK20248/>
5. Burgess, A. W. (2008). EGFR family: Structure and signaling. *Experimental Cell Research*, 314(9), 1600–1606. <https://doi.org/10.1016/j.yexcr.2008.01.014>
6. Chen, L., Zhang, J., & Wang, X. (2016). EGFR overexpression and cancer progression: Implications for therapy. *Oncotarget*, 7(34), 55465–55477. <https://doi.org/10.18632/oncotarget.10874>
7. National Library of Medicine. (2022b). *Epidermal growth factor receptor (EGFR) signaling in cancer*. <https://www.ncbi.nlm.nih.gov/books/NBK555892/>
8. Liu, S., Luo, Z., & Wu, H. (2017). Role of EGFR in neuroblastoma: Implications for therapy. *Oncology Reports*, 37(3), 1421–1430. <https://doi.org/10.3892/or.2017.5383>
9. Anandarajah, A., Lee, H., & Kim, S. (2018). Epidermal growth factor receptor mutations in non-small cell lung cancer. *Journal of Thoracic Oncology*, 13(4), 523–533. <https://doi.org/10.1016/j.jtho.2017.12.004>
10. Jänne, P. A., Yang, J. C., & Kim, D. W. (2014). Targeting EGFR in lung cancer: Therapy, resistance, and future directions. *Nature Reviews Clinical Oncology*, 11(8), 493–507. <https://doi.org/10.1038/nrclinonc.2014.125>
11. Arteaga, C. L., Engelman, J. A., & Pao, W. (2014). Epidermal growth factor receptor inhibitors in cancer therapy. *The New England Journal of Medicine*, 370(10), 874–885. <https://doi.org/10.1056/NEJMr1304961>
12. Wang, L., & Li, Z. (2018). Limitations and toxicities of EGFR-targeted therapy. *OncoTargets and Therapy*, 11, 5839–5852. <https://doi.org/10.2147/OTT.S176460>
13. Dong, D., & Wang, P. (2020). Dopamine D2 receptor signaling in cancer: Mechanisms and therapeutic opportunities. *Cells*, 9(11), 2360. <https://doi.org/10.3390/cells9112360>
14. Probiologists. (2023). The role of certain neurotransmitters in tumor progression and the use of neurotransmitter antagonists. <https://www.probiologists.com/article/the-role-of-certain-neurotransmitters-in-the-emergence-and-progression-of-malignant-tumors-and-the-potential-of-using-neurotransmitter-antagonists-to-block-the-carcinogenic-effect-of-the-tumor-stroma>
15. Bouchez, C., Li, H., & Chen, Y. (2021). Stepholidine: A dual D1/D2 receptor antagonist with potential anticancer properties. *Cancer Biology & Therapy*, 22(3), 212–224. <https://doi.org/10.1080/15384047.2021.19027913>
16. Chen, Y., Li, H., & Zhang, Q. (2024). Berberine as a potential anticancer agent: Mechanisms and applications. *International Journal of Biological Macromolecules*, 253, 123890. <https://doi.org/10.1016/j.ijbiomac.2024.123890>
17. Zhang, R., Chen, L., & Sun, Y. (2021). Molecular mechanisms of berberine in cancer therapy. *Biomedicine*, 9(7), 890. <https://doi.org/10.3390/biomedicine9070898>
18. Sun, Y., Li, J., & Zhang, W. (2020). Anticancer effects of taxifolin, artemisinin, and corosolic acid. *Cells*, 9(2), 498. <https://doi.org/10.3390/cells9020498>
19. Wang R, Zhu X, Wang Q, Li X, Wang E, Zhao Q, Wang Q, Cao H. The anti-tumor effect of taxifolin on lung cancer via suppressing stemness and epithelial-mesenchymal transition *in vitro* and oncogenesis in nude mice. *Ann Transl Med*. 2020 May;8(9):590. doi: 10.21037/atm-20-3329. PMID: 32566617; PMCID: PMC7290558.
20. Jin M, Wu Y, Lou Y, Liu X, Dai Y, Yang W, Liu C, Huang G. Corosolic acid reduces A549 and PC9 cell proliferation, invasion, and chemoresistance in NSCLC via inducing mitochondrial and liposomal oxidative stress. *Biomed Pharmacother*. 2021 Dec;144:112313. doi: 10.1016/j.biopha.2021.112313. Epub 2021 Oct 19. PMID: 34678723.
21. Otto-Ślusarczyk D, Mielczarek-Putna M, Graboń W. The Real Cytotoxic Effect of Artemisinins on Colon Cancer Cells in a Physiological Cell Culture Setting. How Composition of the Culture Medium Biases Experimental Findings. *Pharmaceuticals* (Basel). 2021 Sep 26;14(10):976. doi: 10.3390/ph14100976. PMID: 34681200; PMCID: PMC8540140.

■ Author

Audrey Hong is a junior at Kent Place School. She is passionate about using science to solve real-world problems. She hopes to study biological sciences in college and continue pursuing her interest in research and oncology.

How High-Skill Immigration Restrictions Eroded Regional Productivity: Evidence from the 2017 BAHA Policy

Caroline Y. Su

McLean High School, 1633 Davidson Rd, McLean, VA, 22102, USA; caroline.jia.su@gmail.com

ABSTRACT: This paper examines how limiting high-skill immigration affects the economy at the regional level, using the 2017 “Buy American, Hire American” (BAHA) executive order as a quasi-experimental shock. This restrictive policy significantly increased the difficulty of obtaining H-1B visas, which high-skilled foreign workers primarily rely on to work in the United States. As a result, denial rates for new H-1B applications rose from 7% to 17%, and STEM-specific rejections jumped to an all-time high of 31%. Employing a difference-in-differences method, this study finds that from 2018 to 2020, states with high reliance on H-1B labor experienced a 2.8% drop in economic output compared to states with lower H-1B exposure. In total, this decline translates to an estimated productivity loss of \$218 billion for the most affected regions in 2018 alone. While job counts and wages generally remained stable during this period, likely supported by tax cuts and deregulation, overall productivity was negatively impacted because of restrictions on skilled foreign talent. Thus, this study suggests that policies that only focus on conventional employment metrics may overlook the “hidden” costs of immigration restrictions to economic performance, particularly the impact on innovation and productivity.

KEYWORDS: Economics, Labor Economics, H-1B Visas, Innovation and Productivity, Immigration Policy Shock.

■ Introduction

One of the most widely contested questions within labor economics is whether foreign workers substitute or complement native workers. This issue lies at the heart of United States immigration policy: if policymakers believe that foreign-born workers heighten labor market competition and depress outcomes for native workers, they may pursue restrictive policies to limit foreign labor supply. But if skilled immigrants are viewed as sources of specialized expertise that foster innovation and provide knowledge spillovers, policymakers may take more balanced approaches to facilitate their entry to support economic growth. This paper examines these competing perspectives through the lens of the H-1B visa program, which enables employers to access skilled foreign talent. The 2017 BAHA policy shock is used as a quasi-natural experiment, which allows for identifying the causal effects of immigration restrictions on economic outcomes.

Economists have long argued about the labor market impacts of high-skilled immigration, evidenced by a large body of such studies. Borjas estimates that a 10% increase in the historical immigration supply reduced native wages by 3–4%,¹ which implies that immigrants and natives with similar backgrounds, such as education and experience, are close substitutes. However, this viewpoint is challenged by many subsequent studies. Card finds little evidence that immigration significantly depresses native wages,² and Ottaviano and Peri similarly suggest that immigrants and natives are imperfect substitutes who specialize in different areas.³ In addition, Kerr and Lincoln and Kerr, Kerr, and Lincoln show that skilled immigrants increase innovation and boost science and engineering employment,^{4,5} and Hunt identifies them as job-creating entrepreneurs.⁶

High-skill immigrants are also catalysts for total factor productivity (TFP) and innovation. STEM immigrants are associated with higher rates of patenting and overall productivity increases.^{7,8} Furthermore, research by Akcigit, Grigsby, and Nicholas and Peri, Shih, and Sparber elucidates how foreign-born workers drive U.S. innovation because of their complementarity with native workers.^{9,10}

Despite this evidence, the H-1B program has become an ever more difficult system to navigate after the annual visa cap reverted from 195,000 to 65,000 in 2004. Since the early 2000s, the H-1B policy has been revised multiple times under different presidential administrations. For instance, in 2007, the first-ever computer-generated lottery was implemented for H-1B selection due to unprecedented demand. Although some studies find no significant impacts from H-1B visa lotteries at the firm level,¹¹ many other analyses show that policy-induced shifts in H-1B supply can greatly affect innovation and productivity across cities and states.^{5,10} Recent work by Glennon identifies an “offshoring” phenomenon: when U.S. policy became more restrictive towards immigration, many multinational firms chose to relocate R&D operations to foreign affiliates in China and India.¹² This substantiates the idea that H-1B restrictions can diminish domestic value added, a measure of labor productivity defined as the total value of goods and services produced minus the input cost, without necessarily lowering aggregate employment. That cost was reflected in the estimated \$100 billion decline in market valuation following the 2020 visa suspensions made to combat COVID-19.¹³

The April 2017 BAHA executive order introduced a significant disruption to the H-1B program. The policy increased H-1B denial rates sharply; the two main factors contributing to this were tightened eligibility standards and increased “Re-

quests for Evidence" (RFEs), which caused average rejection rates to rise from 5% (2010-2016) to over 11% by 2018. The shock affected new STEM applications the most, with nearly one in three cases being rejected. Figure 1 shows that each state has a different reliance on the H-1B visa program prior to the BAHA period; this variation creates a unique setting to measure the heterogeneous regional effects of BAHA restrictions on high-skill immigration.

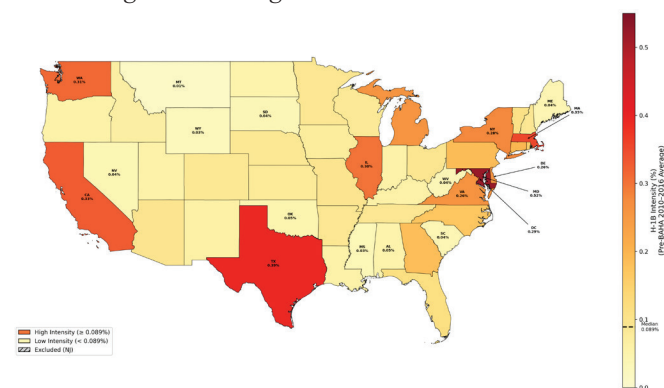


Figure 1: Pre-BAHA H-1B Intensity by State. H-1B exposure varies across states, with many coastal states showing relatively high reliance on H-1B workers. New Jersey is excluded from this analysis (see Methods section for details). Map generated by the author in Python using the pre-policy H-1B intensity data.

To determine how BAHA affects states with different levels of H-1B reliance, this study analyzes state-level immigration and labor productivity data from 2010 to 2020 and divides states into two groups based on their average H-1B share of total employment (H-1B intensity) prior to 2017. Using a difference-in-differences approach, the analysis compares the effect of BAHA on labor market outcomes in high H-1B intensity states relative to low-intensity states. The results show that states with high H-1B intensity experienced a 2.8% relative drop in value-added output ($\beta = -0.028$, $p = 0.049$), which equates to an estimated \$218 billion productivity loss across the ten most exposed states in 2018 alone. Notably, these productivity losses occurred despite tax cuts and deregulation implemented during the same period to spur business activity, prompting firms to maintain strong labor demand. This indicates that although these stimulative policies may have kept labor quantity stable, they were unable to offset the “quality-of-labor” drain induced by BAHA. In short, regional productivity contracted without a material change in aggregate employment.

While previous work has mostly examined firm-level responses to high-skill immigration policy changes, such as Doran, Gelber, and Isen using H-1B visa lotteries,¹¹ the broader implications to regional economies were not clearly evaluated. The 2017 BAHA order caused a significant aggregate contraction in the supply of high-skilled labor, providing a natural setting to evaluate complementarities and knowledge spillovers that may not be visible in firm-level data. In this sense, our study builds on productivity and innovation literature while complementing more recent work on policy-driven shocks.^{5,7,10,12,13}

Collectively, these findings suggest that restrictive policies focused solely on employment outcomes may fundamentally underestimate the overall economic effects of lost productivity on value-added output.

■ Methods

This study identifies the impact of the 2017 BAHA order using a difference-in-differences (DiD) framework together with an event-study analysis used to assess the validity of the DiD assumptions. This policy serves as a natural experiment because it was a sharp, discrete policy shift that provides a clean break for identification, and its uniform federal application means that “treatment” intensity is determined by a state’s pre-existing H-1B reliance rather than the policy itself. This method ensures that the shock is exogenous to state-level economic shifts.

There are other common methods used in immigration literature, such as instrumental variables (IV) or regression discontinuity (RD). However, identifying a truly valid instrument for H-1B exposure is widely debated in the literature. In addition, the BAHA policy did not establish a clear geographical or numerical threshold suitable for the RD approach. As a result, the DiD and event-study approaches are better suited, as this framework leverages longitudinal data to compare the effects of the policy shock across states with different levels of pre-policy H-1B reliance.

Data Sources and Selection:

The primary datasets used in this study include Labor Condition Application (LCA) filings, H-1B petition data, and labor productivity measures from 2010 to 2020.

LCA filings were sourced from the Department of Labor as a primary measure of employer demand for H-1B talent,¹⁴ following the approach of Kerr and Lincoln.⁵ H-1B supply-side records (approvals and denials) were obtained from the U.S. Citizenship and Immigration Services (USCIS) Data Hub files.¹⁵ STEM records, representing the bulk of H-1B petitions, were proxied using NAICS sectors 51 (Information) and 54 (Professional, Scientific, and Technical Services). Labor productivity metrics (e.g., employment, hours worked, hourly compensation, and value-added output) were pulled from the U.S. Bureau of Labor Statistics (BLS) State and Regional Labor Productivity data files.¹⁶ The labor productivity metrics sourced are for all private non-farm sectors, but did not have NAICS sector codes to further differentiate STEM-specific metrics.

Raw datasets cover 2010 through 2024, but this study excludes years beyond 2020 for the DiD and event-study analyses for two reasons: several Trump-era immigration mandates were reversed in 2021, effectively offsetting the original policy treatment; the 2021-2024 window is heavily confounded by the COVID-19 recovery and the introduction of a new H-1B electronic registration system. By ending the analysis in 2020, the study can better isolate the impact of the BAHA order.

Within the 2010-2020 window, the pre-treatment period is 2011-2016, and the post-treatment period is 2018-2020. 2017 was excluded because the BAHA was signed in April 2017. In

addition, the year 2010 serves as the base year for year-over-year calculations and is not included in those calculations. New Jersey was also dropped from the dataset because several major IT outsourcing firms are headquartered there but deploy workers to other states.

The final dataset covers 2010 to 2020, including 49 states plus the District of Columbia (50 in total) and 9 years for a total of 450 state-year observations.

Empirical Trends and Policy Shocks:

This section analyzes immigration trends before and after the 2017 BAHA policy shock. Table 1 shows that LCA filings continued to grow and certification (approval) rates remained stable over the years, indicating growing demand for foreign talent. However, H-1B denial rates surged from 4.9% on average from 2010 to 2016 to 11.5% in 2018, essentially limiting the supply of foreign labor. This means an increasing demand-supply gap due to BAHA.

Table 1: LCA and H-1B Summary Statistics (average across states). LCA filing increased over time, but BAHA caused a significant “denial shock” to H-1B petitions, especially to STEM categories.

Labor Demand vs. Supply	Variable	Pre-BAHA (2010-2016)	2018	2019	2020
		Mean	Mean	Mean	Mean
Demand	LCA Total Filings	8,357	11,501	11,901	12,562
	LCA Certification Rate (%)	82.0	85.5	87.8	92.1
	STEM LCA Total	5,299	7,607	7,918	7,709
	STEM LCA Certification Rate (%)	83.0	86.2	89.1	91.9
Supply	H-1B Total	5,173	6,989	8,146	8,307
	H-1B Denial Rate (%)	4.93	11.5	11.34	7
	H-1B New Employment Denial Rate (%)	7.2	17.1	15.8	9.6
	STEM H-1B Denial Rate (%)	6.3	19.6	18.2	12.8
	STEM H-1B New Employment Denial Rate (%)	8.7	31.4	27.0	17.6

As illustrated in Figure 2 below, this denial shock was not uniform. Denial rates for renewals (continuations of employment) remained stable throughout the decade, but new employment petitions saw denial rates jump from 7.2% to over 17.1% in 2018. STEM-specific new employment petitions saw an even more pronounced rise from 8.7% to 31.4%. This surge in rejections for new employment filings significantly tightened the new talent pipeline.

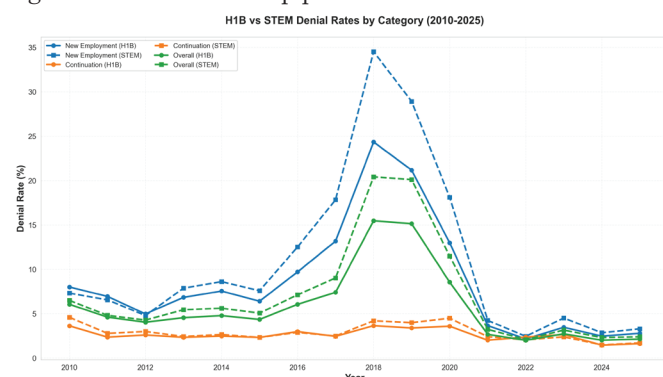


Figure 2: National-level H-1B and STEM Denial Rates. H-1B and STEM denial rates remained stable and low from 2010 to 2016. Post-BAHA denial rates spiked significantly for new employment petitions and STEM-specific categories. Figure generated by the author using the USCIS H-1B and STEM denial rate data.

Treatment Variables:

H-1B intensity is the treatment variable to measure cross-state variation in H-1B dependence during the pre-BAHA period. For each state, this variable is defined as:

$$H1B_{Intensity} = \left(\frac{H1B_{Approvals}_{s,2010-2016}}{Employment_{s,2010-2016}} \right) \times 100 \quad (1)$$

After calculating H-1B intensity, states are split into two groups using the median value of 0.089% across all states. 25 states, such as Maryland, Texas, and California, with a high concentration of technology, biotech, and finance sectors, are in the high-intensity group. The other 25 states, such as Montana, Alaska, and Wyoming, with high exposure to agriculture and natural resources, are in the low-intensity group. New Jersey was excluded from the dataset, as explained in Data Sources and Selection. Because the H-1B intensity values are computed using pre-policy data, it is predetermined relative to the policy shock, thereby minimizing reverse causality concerns. The binary treatment indicator is:

$$HighH1B = 1 \text{ if } H1B \text{ Intensity} \geq \text{median}; 0 \text{ otherwise}$$

Outcome Variables:

Four variables are used to track labor market outcomes: employment, hours worked, real hourly compensation, and value-added output. Since the distribution of these variables is heavily right-skewed, we transform them into natural logarithms. This step reduces skewness by more than 90% while still allowing for an intuitive interpretation of the result as percentage changes (Table 2).

Table 2: Labor Market Outcome Variables Distribution and Transformations. Log transformation reduces skewness by 94-98%.

Variable	N	Mean	Std Dev	Skewness (Raw)	Kurtosis (Raw)	Skewness Improvement (Log)
Employment (K)	750	2,568	2,923	2.58	7.72	0.04 98.4%
Hours Worked (M)	750	4,389	5,024	2.59	7.64	0.04 98.5%
Real Hourly Comp (\$/hr)	750	38.10	4.50	0.95	1.13	~0 >95%
Value-Added Output (M, 2012\$)	750	317,633	433,301	3.32	13.67	0.2 94.0%

Difference-in-Differences Model Specification:

The DiD model uses the following equation:

$$\log Y_{st} = \alpha + \beta_1 HighH1B_s + \beta_2 Post_t + \beta_3 (HighH1B_s \times Post_t) + \lambda (HighH1B_s \times t) + \delta_s + Y_t + \epsilon_{st} \quad (2)$$

where Y_{st} represents the outcome variable (employment, hours worked, real hourly compensation, or value-added output) in state s and year t ; $HighH1B_s$ is the binary treatment indicator for high-treatment states; $Post_t$ equals one for years 2018 and later (the BAHA period); $HighH1B_s \times Post_t$ is the interaction term that captures the treatment effect of BAHA on high-H1B states in the post period, relative to low-H1B states; δ_s and Y_t are state and year fixed effects; and $HighH1B_s \times t$ is a state-specific linear time trend that reflects the pre-existing trend of high-H-1B states relative to the low-H-1B states.

β_1 represents the baseline difference between the high- and low-H1B groups, and β_2 is the common post-BAHA shift effect for both groups. β_3 identifies the causal effect of the BAHA policy on the states with high H-1B exposure, under the parallel trends assumption. The DiD approach compares the states in the high-intensity group to their own pre-policy history and to the control group (low-intensity states). Since a log specification is used, multiplying β_3 by 100 is the estimated percentage effect. Finally, standard errors ϵ are clustered

by state to account for serial correlation over time, to ensure robustness for significance testing across 50 clusters.

As stated in the above section, the pre-treatment period is defined as 2011–2016. The DiD model estimates the policy impacts over three post-treatment periods (2018 only, 2018–2019, and 2018–2020) in order to evaluate how firms were able to adjust to the new policy over time.

Identification Strategy:

Three conditions are key to ensuring that β_3 captures a causal relationship rather than a simple correlation of variables. First, high- and low-intensity states should have parallel or similar trends before the BAHHA policy shock. This assumption can be validated by examining the coefficients of each labor outcome variable in the pre-policy period (see Event Study Specification). As explained in Results and Discussion, pre-policy coefficients for two labor variables, employment and real hourly compensation, hover near zero, suggesting parallel trends. However, the other two variables, value-added output and hourly worked, show non-parallel upward trends, necessitating the use of state-specific linear time trends (λ). Second, it is important to ensure exogenous treatment. This study uses the average H-1B intensity from 2010 to 2016 as a pre-determined metric to meet this condition. Finally, this analysis uses state and year fixed effects to control for unobserved confounders, which absorb permanent state differences and other factors, such as the Tax Cuts and Jobs Act (TCJA) of 2017.

Event Study Specification:

Event study tests the pre-policy parallel trends assumption to complement the static DiD approach, using the following specification:

$$\log Y_{st} = \alpha + \beta_1 \text{HighH1B}_s + \sum_{k \neq 2016} \theta_k 1\{t=k\} + \sum_{k=2011, k \neq 2016}^{2020} \beta_k (\text{HighH1B}_s \times 1\{t=k\}) + \epsilon_{st} \quad (3)$$

Where β_1 is the baseline difference between the high- and low-intensity states in the reference year (2016, the full year prior to the BAHHA order); θ_k is the year effect in year k (relative to 2016) common to both groups; β_k is the key coefficient that measures the difference between the two groups in year k , relative to the reference year; and $1\{t=k\}$ is indicator for k from 2011 to 2020, with 2016 set at zero.

The pre-2016 coefficients (β_{2011} through β_{2015}) directly test the parallel trends assumption: if these coefficients were close to zero, then high- and low-intensity states were on similar trends prior to BAHHA. The post-2017 coefficients (β_{2018} through β_{2020}) reveal the magnitude of the impact after the BAHHA policy was implemented.

Results and Discussion

DiD Model Estimates:

β_3 in Table 3 highlights the DiD results. The most significant result is a 2.8% relative decline in value-added output across high-intensity states in 2018 ($\beta_3 = -0.028$, $p = 0.0498$). The effect continued on this path through 2020, suggesting that the "productivity shock" persisted as the policy was enforced. The state-specific trend control (λ) was essential for identi-

fying the true β_3 , because high-intensity states were growing faster annually ($\lambda = 0.0109$ in Table 3) relative to low-intensity states before BAHHA; omitting this control would have masked the true magnitude of the policy's negative impact. In contrast, the result also reveals a discrepancy: while economic output dropped significantly, labor input (employment, hours worked, and real hourly compensation) showed minimal movement ($p > 0.15$). This suggests that the policy dealt a blow to work efficiency rather than labor quantity or wages.

β_1 and β_2 are absorbed by the state and year fixed effect terms in the DiD model in Equation 2 and hence dropped from the final DiD model estimates. Table 3 shows model estimates for β_3 .

Table 3: DiD Model Results. β_3 for Value-Added Output consistently shows statistical significance across the periods. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SE (cluster) is associated with β_3 .

Post-BAHA Periods	Outcome	β_3 (DiD)	SE (cluster)	t-stat	p-value	Sig	$\beta_3 \times 100$ (t=1)	trend λ	95% CI
1-Yr Post (2018)	log(Employment)	0.0078	0.0059	1.317	0.1879		0.78	0.0021	[-0.0038, 0.0194]
	log(Hours Worked)	0.0004	0.0077	0.048	0.9616		0.04	0.0044	[-0.0148, 0.0155]
	log(Real Hourly Comp)	0.0072	0.0088	0.811	0.4173		0.72	(0.0036)	[-0.0102, 0.0245]
	log(Value-Added Output)	-0.0281	0.0143	-1.962	0.0498	**	-2.81	0.0109	[-0.0561, 0.0000]
2-Yr Post (2018-19)	log(Employment)	0.0076	0.0062	1.234	0.2173		0.76	0.0021	[-0.0045, 0.0197]
	log(Hours Worked)	-0.0024	0.0083	-0.287	0.7742		-0.24	0.0042	[-0.0186, 0.0138]
	log(Real Hourly Comp)	0.0126	0.0092	1.371	0.1705		1.26	(0.0032)	[-0.0054, 0.0306]
	log(Value-Added Output)	-0.0290	0.0151	-1.914	0.0556	*	-2.90	0.0109	[-0.0586, 0.0007]
3-Yr Post (2018-20)	log(Employment)	0.0070	0.0063	1.108	0.2678		0.70	0.0015	[-0.0054, 0.0194]
	log(Hours Worked)	-0.0029	0.0084	-0.345	0.7305		-0.29	0.0037	[-0.0194, 0.0136]
	log(Real Hourly Comp)	0.0130	0.0092	1.418	0.1551		1.30	(0.0028)	[-0.0050, 0.0310]
	log(Value-Added Output)	-0.0286	0.0152	-1.879	0.0602	*	-2.86	0.0113	[-0.0584, 0.0012]

Event Study Evidence (Validating the DiD Assumptions):

Figure 3 shows the coefficients of pre- and post-policy trends of high-intensity states, relative to low-intensity states. As shown in this graph, pre-treatment coefficients for employment and real hourly compensation oscillate near zero, affirming that these two variables followed similar trajectories before the policy. Post-2017 coefficients for these variables also remain insignificant.

However, value-added output and hours worked show clear upward trends prior to 2016 and violate the parallel trends assumption. This violation confirms the need for the inclusion of a state-specific linear trends term (λ) in Equation 2 to address the pre-policy trends and separate the policy's impact from the baseline growth trend differences.

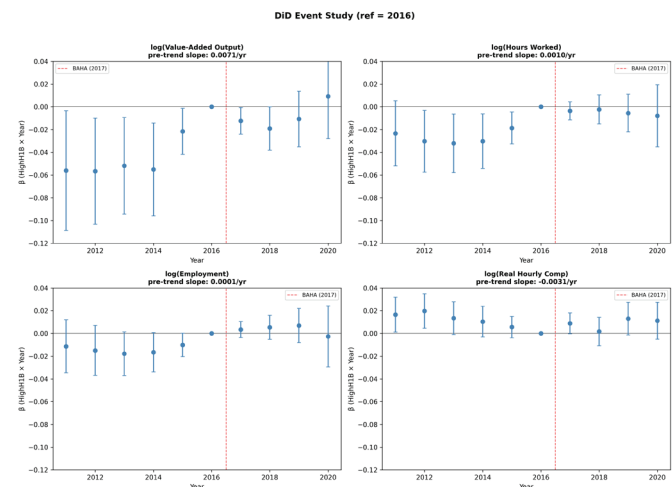


Figure 3: Event Study DiD Coefficients. Pre-2016 coefficients for employment and real hourly compensation are near zero, while value-added output and hours worked show non-parallel upward trends. Graph generated by the author in Python using event study DiD regression estimates.

The subsequently detrended graph (Figure 4) shows that the state-specific trends term (λ) effectively handles the parallel trends assumption violation issue. After subtracting the pre-treatment linear trend from the entire series, the pre-period slope is centered around zero for all four outcome variables. This ensures that post-2017 coefficients are a true representation of the policy shock effects, not continuations of prior trends.

The detrended post-treatment coefficients for value-added output remain negative and statistically significant. This verifies that the economic output drop in high-intensity states was a distinct break caused by the BAHHA order. By comparison, the other three variables show no significant upward or downward patterns away from zero. This reinforces our conclusion that the impact of the BAHHA maneuvered through a decline in productivity as opposed to labor quantity or wages.

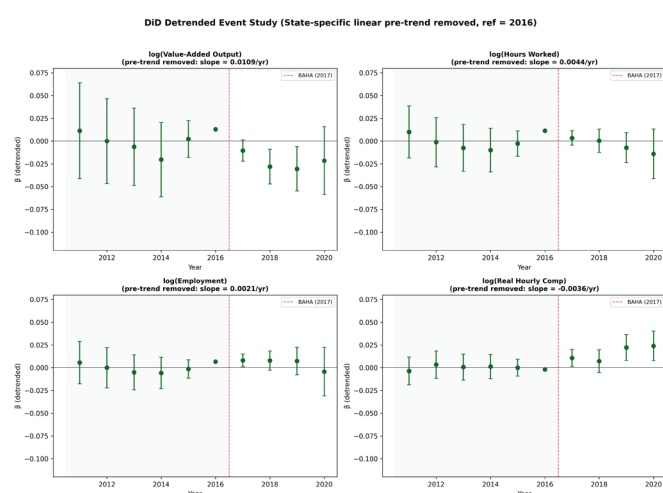


Figure 4: Detrended Event Study DiD Estimates. Detrended pre-policy coefficients for all four outcome variables are around zero, thereby satisfying the parallel trends assumption. Figure generated by the author in Python using detrended event study DiD regression estimates.

2017 is retained as the reference/boundary year in Equation 3 and in Figures 3 and 4, where it marks the transition point between the pre- and post-policy periods on the event-study timeline.

Robustness Checks:

Several robustness checks were conducted to confirm the validity of the DiD model results (Table 4).

The first check is a placebo test, run by estimating the model using only pre-BAHHA data (2011–2016) and assigning a fictitious policy shock to 2014. As expected, the result was statistically insignificant and actually moved in the opposite direction of the main findings ($\beta_{\text{placebo}} = +0.023$). The fact that no “phantom” effect appeared in the years leading up to the policy corroborates that the post-2017 decline is a causal shift.

The second test swapped the general H-1B intensity variable with a STEM-specific metric, which identifies the loss of technical talent as the main factor behind the drop in productivity. Nearly identical results were found ($\beta_3 = -0.027$).

The third test evaluated whether states with “middle-of-the-road” H-1B reliance were diluting the results, using an

extreme-tercile comparison which divides the dataset into three equal thirds. Comparing only the top and bottom thirds of the sample and omitting the middle third, the negative impact on value-added output nearly doubled to -4.3% ($p = 0.014$). This suggests that the baseline -2.8% estimate is likely conservative. More importantly, it suggests that the productivity drain was directly proportional to how much a state’s economy relied on H-1B talent.

Finally, two additional specification checks (Test 4 and 5 in Table 4) were conducted to confirm that the findings were not driven by outliers or pre-policy trends. Test 4 shows the coefficient holds relatively steady ($\beta_3 = -0.031$) after excluding the three largest H-1B states. Test 5 tests the impact of removing the state-specific linear time trends from the model; β_3 result becomes insignificant, which validates that we must account for pre-existing growth trend to identify the policy’s real impact. It is important to note that the central finding rests on the assumption that the diverging pre-2017 growth trajectories of high-intensity states would have continued linearly in the absence of the BAHHA policy. This linear extrapolation is a standard and defensible approach in DiD designs facing non-parallel pre-trends, but it is an assumption rather than an observed counterfactual.

Table 4: Robustness Checks Results for Value-Added Output. These checks confirmed the validity of the DiD model results. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Specification	Treatment	1-Yr Post β_3	p-value	Sig	2-Yr Post β_3	p-value	Sig	3-Yr Post β_3	p-value	Sig
Primary DiD Model	Binary (50 states/DC)	-0.0281	0.049	**	-0.0290	0.055	*	-0.0286	0.060	*
(1) Placebo (2014)	Binary (50 states/DC)	+0.0231	0.203							
(2) STEM Treatment	Binary + STEM-specific (50 states/DC)	-0.0269	0.061	*	-0.0283	0.062	*	-0.0280	0.066	*
(3) Extreme Tercile	Binary (34 states/DC)	-0.0428	0.014	**	-0.0429	0.020	**	-0.0422	0.024	**
(4) Drop CA, TX, NY	Binary (47 states/DC)	-0.0309	0.032	**	-0.0316	0.039	**	-0.0311	0.043	**
(5) No State-specific Linear Trends Term	No trend control	+0.0211	0.293		+0.0253	0.251		+0.0334	0.185	

Interpretation:

This study points to a clear “decoupling” observation of labor quantity and economic output. Following the BAHHA order, firms in high-intensity states appear to have maintained their headcounts by replacing denied H-1B talent with domestic workers. But this replacement was likely associated with contemporaneous policy stimuli, most prominently the TCJA of 2017 and overall federal deregulation.

However, this replacement carried a steep, covert cost. Domestic backfills were inherently imperfect substitutes for H-1B workers, oftentimes lacking the STEM expertise or niche human capital of the H-1B workers that they replaced. As a result, work efficiency suffered. Essentially, the 2017–2020 fiscal stimulus successfully preserved the number of workers, but the immigration restrictions likely eroded the quality of that labor.

While our analysis does not directly measure innovation or firm-level productivity, these findings align with the work of Peri, Shih, and Sparber,¹⁰ who argue that reductions in high-skill (especially STEM) immigrant concentration can diminish knowledge spillovers, specialization, and innovation, all of which are key drivers of total factor productivity. As an illustrative calculation, applying the 2.8% decline to the 2018 value-added output of the top ten H-1B-intensive states suggests a potential productivity gap of \$218 billion (95%

Confidence Interval: approximately \$1 billion to \$437 billion), relative to an extrapolated counterfactual trend (Table 5). This figure should be treated as an estimate rather than a precise effect, since it applies the average effect across all high-intensity states to the top ten states and is subject to substantial estimation uncertainty.

Table 5: Estimated Value-Added Gap in the Top 10 H-1B-Intensive States (2018). This \$218 billion gap is estimated as the 2018 Value-Added Output multiplied by 0.028 (where $\beta_3 = -0.028$ applies to each state).

Rank	State	H-1B Intensity (%)	2018 Value-Added Output (\$M)	Implied 2.8% Gap (\$M)
1	Maryland	0.523	291,815	8,171
2	Texas	0.392	1,527,024	42,757
3	Massachusetts	0.351	464,671	13,011
4	California	0.329	2,342,527	65,591
5	Washington	0.313	454,680	12,731
6	Illinois	0.300	722,516	20,230
7	District of Columbia	0.290	91,277	2,556
8	New York	0.279	1,432,581	40,112
9	Virginia	0.261	396,841	11,112
10	Delaware	0.260	60,409	1,691
Total			7,784,341	217,962

■ Conclusion

The 2017 BAHA policy shock provides a natural experiment to quantify how the effects of sudden restrictions on high-skill immigration can ripple through regional economies. Our study finds that the states relying most on H-1B talent experienced a 2.8% relative drop in value-added output; for the ten most H-1B-intensive states alone, this represents an estimated \$218 billion productivity loss in 2018. Meanwhile, no significant shifts in employment or wages were identified during the post-policy period, which suggests that labor replacement alone does not offset the impact of severing the supply of specialized foreign talent on total factor productivity and knowledge spillovers.

The results thus indicate that measuring immigration policy solely based on "headcount" or "jobs saved" may be insufficient. Doing so crucially risks ignoring the hidden economic toll of losing highly specialized expertise. To understand these costs more thoroughly, further work should investigate whether this talent drain slows patenting and R&D investment.

Building upon these findings, an important direction for future research is to analyze how alternative high-skill immigration policy frameworks interact with ongoing technological change. For example, studies could evaluate whether technological breakthroughs will continue to increase demand for highly skilled STEM workers or substitute for certain specialized tasks. Understanding these interactions would help policymakers assess whether future productivity shocks resulting from immigration restrictions are likely to be amplified or offset by advanced technologies. Studies should also examine whether uncertainty in U.S. immigration policy is driving international students and STEM talent away from the U.S. and towards competing technology hubs in other countries.

In addition, future studies can move from empirical diagnosis to actual policy recommendations. We can leverage the baseline foundation from this paper to build and test concrete policy ideas, such as fast-tracked application paths for high-demand fields, to help mitigate these productivity losses.

As for long-term effects, the economic burden of these restrictions does not end when a specific executive order is revoked. The U.S. regulatory stance frequently shifts, and this "on-again, off-again" nature of immigration policy is exemplified by the 2021 BAHA reversal and the introduction of a \$100,000 supplemental payment requirement for new, non-exempt H-1B petitions filed after September 2025. This continual volatility may be one of the most lasting and costly burdens to long-run economic growth.

A political economy perspective may help explain why restrictive high-skill immigration policies persist, despite evidence of aggregate economic costs. While the results indicate a decline in regional productivity following reduced access to H-1B talent, these effects are distributed across states and firms broadly and may be less immediately visible to the public. In contrast, potential benefits, such as reduced perceived competition in specific occupations or greater prioritization of domestic hiring, are more concentrated and more salient to affected groups. As a result, different stakeholders may perceive or experience the policy's costs and benefits unevenly, contributing to differing policy preferences. Public opinion may also play a role, as views on immigration policy often reflect concerns about job opportunities, wage dynamics, and fairness in labor markets, even when empirical evidence points to broader productivity effects. These considerations suggest that future research could examine how distributional dynamics and voter perceptions interact with aggregate economic outcomes in shaping the policies.

■ Acknowledgments

I would like to thank Lisa Wu for her mentorship and guidance throughout the development of this research. I would like to thank all reviewers for their invaluable time and constructive feedback.

■ References

1. Borjas, G. J. The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market. *Q. J. Econ.* **2003**, *118* (4), 1335-1374. <https://doi.org/10.1162/003355303322552810>.
2. Card, D. Immigration and inequality. *Am. Econ. Rev.* **2009**, *99* (2), 1-21. <https://doi.org/10.1257/aer.99.2.1>.
3. Ottaviano, G. I. P.; Peri, G. Rethinking the effect of immigration on wages. *J. Eur. Econ. Assoc.* **2012**, *10* (1), 152-197. <https://doi.org/10.1111/j.1542-4774.2011.01052.x>.
4. Kerr, S. P.; Kerr, W. R.; Lincoln, W. F. Skilled immigration and the employment structures of U.S. firms. *J. Labor Econ.* **2015**, *33* (S1), S147-S186. <https://doi.org/10.1086/678986>.
5. Kerr, W. R.; Lincoln, W. F. The supply side of innovation: H-1B visa reforms and U.S. ethnic invention. *J. Labor Econ.* **2010**, *28* (3), 473-508. <https://doi.org/10.1086/651934>.
6. Hunt, J. Which immigrants are most innovative and entrepreneurial? Distinctions by entry visa. *J. Labor Econ.* **2011**, *29* (3), 417-457. <https://doi.org/10.1086/659409>.
7. Hunt, J.; Gauthier-Loiselle, M. How much does immigration boost innovation? *Am. Econ. J. Macroecon.* **2010**, *2* (2), 31-56. <https://doi.org/10.1257/mac.2.2.31>.

8. Peri, G. The effect of immigration on productivity: Evidence from U.S. states. *Rev. Econ. Stat.* **2012**, *94* (1), 348-358. https://doi.org/10.1162/REST_a_00137.
9. Akcigit, U.; Grigsby, J.; Nicholas, T. Immigration and the rise of American ingenuity. *Am. Econ. Rev.* **2017**, *107* (5), 327-331. <https://doi.org/10.1257/aer.p20171021>.
10. Peri, G.; Shih, K.; Sparber, C. STEM workers, H-1B visas, and productivity in U.S. cities. *J. Labor Econ.* **2015**, *33*(S1), S225-S255. DOI: 10.1086/679061. <https://doi.org/10.1086/679061>.
11. Doran, K.; Gelber, A.; Isen, A. The effects of high-skilled immigration policy on firms: Evidence from visa lotteries. *J. Polit. Econ.* **2022**, *130* (10), 2501-2533. <https://doi.org/10.1086/720467>.
12. Glennon, B. How Do Restrictions on High-Skilled Immigration Affect Offshoring? Evidence from the H-1B Program. *Manage. Sci.* **2024**, *70*(2), 907-930. <https://doi.org/10.1287/mnsc.2023.4715>.
13. Bahar, D., Choudhury, P., & Glennon, B. An executive order worth \$100 billion: The impact of an immigration ban's announcement on Fortune 500 firms' valuation. *NBER Working Paper No. 27997*; National Bureau of Economic Research, **2020**, <https://doi.org/10.3386/w27997>.
14. U.S. Department of Labor, Office of Foreign Labor Certification. *LCA Disclosure Data*. <https://www.dol.gov/agencies/eta/foreign-labor/performance> (accessed January 31, 2026).
15. U.S. Citizenship and Immigration Services. *H-1B Employer Data Hub*. <https://www.uscis.gov/tools/reports-and-studies/h-1b-employer-data-hub> (accessed January 31, 2026).
16. U.S. Bureau of Labor Statistics. *State and Regional Productivity*. <https://www.bls.gov/lpc/state-productivity.htm> (accessed January 31, 2026).

■ Author

Caroline Su is a student at McLean High School in Virginia. Her research focuses on the intersection of labor economics and public policy, especially how federal mandates impact regional growth. Beyond this study, she is also exploring how emerging technologies will structurally transform labor markets and long-run economic performance.

Perceptual Gaps: ASCII Art and Overlapping Audio as CAPTCHA

Chong Choon-Hou Rafael

Hwa Chong Institution, 661 Bukit Timah Road, Singapore, 269734, Singapore; chongchoonhourafael@gmail.com

ABSTRACT: As multimodal large language models (LLMs) advance, traditional CAPTCHAs have become obsolete at distinguishing humans from bots. To address this shift, this paper aims to investigate the possibility of using tasks for which humans have evolved highly specialized neural processing. We introduce two CAPTCHA classes: a vision-based CAPTCHA, which renders alphanumeric strings as ASCII art, and an audio-based CAPTCHA, which is a question-answering task with overlapping or noise-corrupted audio context. We evaluate our vision-based CAPTCHA both as text and image input with multiple frontier LLMs (GPT 5.2, Gemini 3, etc.), and assess our audio-based CAPTCHAs by applying augmentations like background noise, Gaussian noise, and overlapping speech. We determined that none of the LLMs were able to solve a single ASCII-based CAPTCHA, with the best-performing model only being able to infer at most one or two characters. Additionally, all models that supported audio performed only modestly better than random when solving audio CAPTCHAs. Our results suggest that these CAPTCHAs are exceptionally effective today, but it is unclear whether they can withstand the fast-evolving landscape of artificial intelligence. Subsequent research is needed to determine whether these tasks are temporary vulnerabilities or represent a more durable method of distinguishing humans from bots.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, Cognitive Systems, Multimodal Large Language Models, CAPTCHA Security Design.

■ Introduction

A key premise of the free Web, with sites such as Wikipedia, YouTube, and various social media platforms, is that access to these services is given for free to humans for their benefit, with the support of donations or online advertising. However, these organizations do not have infinite resources; every visit to these sites consumes CPU, memory, and network bandwidth, so these organizations need to be very careful with whom they allow access to their site. For the last few decades, but more than ever today, a large proportion of visitors are bots that neither watch advertisements nor are the target audience of non-profits like Wikipedia, but still consume resources nonetheless. This has led many organizations to deploy ways to distinguish bots from humans through the use of little online puzzles known as CAPTCHAs: Completely Automated Public Turing Tests to Tell Computers and Humans Apart.¹ A common example of a CAPTCHA is selecting from a set of images which ones contain a green traffic light.

For many years, CAPTCHAs were incredibly effective in deterring bot activity.² However, given the advances in artificial intelligence (AI), bots are now quite good at solving CAPTCHAs. Of particular concern are advanced multimodal models (meaning they handle multiple modalities like text, images, audio, video) like GPT-5, Gemini 3, and Claude Opus 4.5, which already have the capacity to match or even exceed human performance at most of the standard puzzles used in CAPTCHAs.

As these systems become more integrated into our daily lives, two key questions arise:

- *How can online resources provide their services to actual humans rather than allow bots to consume all of their resources?*
- *How can we make sure that the content we read online is not fake, spam, or misinformation that is being published at scale?*

The first question is important because we need to find ways to lower operational costs, steward limited resources towards humans, and protect against copyright violations at scale. With AI agents gaining the ability to perform web-scraping completely without human intervention, original content can be automatically copied and used without the permission of the creators. By ensuring that only human users can access these websites, the service owners can reduce plagiarism attempts and channel the content to its target audience.³

Additionally, cybersecurity concerns necessitate the ability to distinguish between humans and AI bots. For example, a bot can try to guess a user's password many times, or it can try to perform operations that consume a lot of resources for providers (such as looking for driving directions in Google Maps), which are examples of denial-of-service (DoS).⁴ Simply put, DoS attacks are malicious attempts to disrupt a service by overwhelming it with a flood of requests. CAPTCHAs can help defend against these attacks by detecting that the source of the requests is a bot and stopping it from performing expensive operations or opening certain websites. As such, if these CAPTCHAs are easily solvable by AI bots, DoS attacks are once again a big problem.

The second question strives to prevent large-scale automated misinformation attacks. For example, on social media platforms, bot accounts are often the main reason behind the

spread of misinformation and manipulative content, contributing to polarizing our fragile media landscape today.⁵

In the past, both of these concerns could be (at least partially) addressed through the use of CAPTCHAs. However, as prior works by Wang *et al.* and our experiments in Section 2.2 have shown, the capabilities of recent large language models are more than enough to beat existing CAPTCHAs.⁶

This gives rise to three possible solutions:

- (1) *Requiring user accounts and tracking users;*
- (2) *Requiring payment for accessing content;*
- (3) *Create new CAPTCHAs that humans can solve with ease, while state-of-the-art agents have trouble solving them economically.*

The first option is not sustainable, since the additional step of registration and verification ruins the experience for the use of online services, especially menial ones like browsing content. Additionally, with a little more effort, bots and agents can create fake accounts at scale and bypass these defenses, as seen by how Sybil attacks render networks vulnerable to malicious activities.⁷ A Sybil attack is a security threat where a single attacker takes control of a network by creating a large number of fake identities, gaining disproportionate influence in the network.

The second option is also fraught with challenges, as seemingly trivial services become inaccessible for those without disposable income or online payment services.

The final option is the subject of interest in this work. We want to investigate the possibility of novel CAPTCHAs which are simple for humans, but computationally difficult or financially unsustainable for bots and agents to solve.

Problem:

Designing a reliable and robust method to differentiate whether an agent is human or a robot has been fraught with difficulty. For instance, the advanced multimodal large language models (LLMs) that have permeated the digital sphere have become comparable to regular people in solving automated Turing tests. Some tasks that are potentially challenging for AI are also unsuitable for generalizing to all groups of people, especially the young, elderly, and disabled.

Moreover, these tasks must be easily and automatically generated in large quantities at low cost, ensuring cost-effectiveness and diversity. If only a small set of challenges is available, they can be memorized or defeated through training, rendering the system ineffective over time. Thus, there is a vital need for a novel automated Turing test system that ameliorates these problems.⁸

CAPTCHAs and Alternatives:

CAPTCHAs have been the norm for separating legitimate users from bots for the past two decades, serving as a lightweight, scalable defense mechanism across numerous applications and websites.

There have been many proposed ideas that stray away from traditional CAPTCHAs, but they all have various trade-offs. One approach is verifying user authenticity based on account behavior over time, such as tracking and identifying suspicious

patterns and requests that are physically impossible.⁹ However, this method requires a sufficient time period and is ineffective against one-time attacks. It also does not work for services where users do not have long-term accounts.

Another approach involves the use of Privacy Pass, which authenticates users through anonymous cryptographic tokens. These tokens verify human behavior without revealing identity or enabling tracking across websites.¹⁰

Despite this, these passes can be shared with AI agents, which undermines their effectiveness. To counter this, some may think of making the passes non-anonymous, but this comes with the added risk of cross-site activity tracking, which threatens user privacy. Balancing privacy, accountability, and the ability to deter or even ban malicious actors remains an unsolved challenge in this space.

Prior Works:

The baseline for differentiating humans and AI—the CAPTCHA system by L. von Ahn *et al.*—has been used to deter online bots.¹ However, with the rapid advancement of developments into LLMs and Agentic AI, the capabilities of LLMs have surpassed the assumptions that the original CAPTCHA designers had of computers, and have thus rendered these CAPTCHAs obsolete. In order to challenge the capabilities of these advanced multimodal LLMs, there have been numerous proposals for automated Turing tests with unique qualities and interesting results.

As mentioned by Wang *et al.*,⁶ there have been many innovations made towards cracking text-based CAPTCHAs, which include segmentation, one-stage CAPTCHA recognition models, and transfer learning on pretrained models.¹¹ These three methods have not only yielded impressive results, it has also become relatively simple to generate training data to improve model performance through *fine-tuning*.¹²

In a survey conducted by Tariq *et al.*,¹³ a wide range of CAPTCHA techniques were identified, ranging from more commonly used approaches such as text-based CAPTCHAs, 3D-text, handwritten character recognition, and image-based selection tasks, to less widely deployed or novel forms like collage-based CAPTCHAs, mouse-interaction tasks, video interpretation, Sketcha (sketch orientation recognition), public key-embedded puzzles, game-based challenges, and question-based tests designed to engage logical or associative reasoning.

From these studies, three persistent issues emerge:

- (1) *Learnability:* If a task can be labeled at scale, AI models can inevitably be trained to complete it.
- (2) *Usability:* Complicated tasks are more challenging for both AI agents and users, limiting convenience and accessibility to web resources.
- (3) *Adaptation Lag:* Static challenge types that do not evolve allow AI models to eventually solve them reliably.

Our Proposal:

In this work, we propose a new direction for identifying tasks that are resistant to this new generation of LLMs. Rather than focusing on tasks that are arbitrarily difficult for AI,

we explore domains in which humans have *deep evolutionary advantages*. Specifically, we investigate tasks that were critical to human survival over millions of years and for which we have developed specialized neural processing, such as certain types of vision and auditory tasks. These tasks are often trivial for humans but are overlooked by AI training tasks, as they are not commercially useful or present in standard datasets.

One area we focus on is visual processing and, in particular, the recognition of ASCII art. Despite its simplicity and ease of generation, we found that every single modern AI model, even frontier models like GPT 5.1 and Gemini 3

Pro, as well as those that are designed explicitly for solving CAPTCHAs, perform extremely poorly at recognizing ASCII art, with one-shot success rates of zero. This is a stark contrast to other tasks, where even weak AI models typically achieve non-trivial performance. Importantly, humans can recognize these images with high accuracy. Therefore, ASCII art satisfies many of our criteria: it is inexpensive to generate, difficult for existing AI, and easy for most humans to interpret.

Another area we explored was audio-based tasks. Specifically, we wanted to focus on the ability for people to separate voices from noisy background noise or overlapping conversations, a characteristic known as the “cocktail party effect.”¹⁴ Although we expected these tasks to be difficult for AI, our experiments show that current models are surprisingly competent at parsing overlapping audio streams. However, performance still drops with overlapping audio, background noise, and Gaussian noise, though not to the extent required for a robust human verification system. Additionally, audio tasks are more computationally expensive to generate and validate compared to visual tasks, similar to ASCII art.

Paper Outline:

The paper is structured as follows. First, we discuss the design philosophies behind an ideal CAPTCHA system and examine the ability of modern LLMs to solve existing CAPTCHAs. Second, we review the evolutionary advantages in certain perceptual tasks that are difficult for modern AI systems to simulate. Third, we introduce two proposed CAPTCHAs: a vision-based ASCII art CAPTCHA and an audio-based overlapping-speech CAPTCHA, as well as hypothesize the performance of these CAPTCHAs based on the unique perceptual strengths of humans. Fourth, we present our experimental methodology, outlining the data generation, model selection, testing procedures, and evaluation metrics to quantitatively measure the performance of our proposed CAPTCHAs. Finally, we present and analyze the experimental results, discussing the implications and limitations of our findings, concluding with directions for future research in the field of CAPTCHA systems.

■ Background

In this section, we discuss CAPTCHAs and some of the properties that they are guaranteed to possess. We also provide examples of common CAPTCHAs.

Design Properties:

CAPTCHAs are built around three core properties:

Easy to generate at scale. Providers must create fresh challenges in real time for millions of requests. This means (i) low computational and monetary cost per instance, (ii) diversity so that two users rarely see the same challenge, and (iii) minimal human effort. If generation is expensive or slow, latency spikes and availability degrade.

Easy for humans. Human ability is widely distributed. Overly difficult challenges lock out legitimate users, which can (1) reduce revenue, (2) damage a site's reputation through complaints, and (3) violate accessibility requirements. CAPTCHAs must therefore be simple and understandable, solvable with common modalities (vision, simple logic), and have sufficient variety.

Hard for bots. Automated solvers should struggle to achieve high accuracy or to generalize these tasks. This suggests that we should select tasks whose solution space is large and noisy, where labeled data is hard to come by, and whose structure is not a straightforward pattern. Ideally, the challenge should degrade gracefully even if partial automation is possible (e.g., requiring interaction timing or device-bound proofs).

Examples of modern CAPTCHAs:

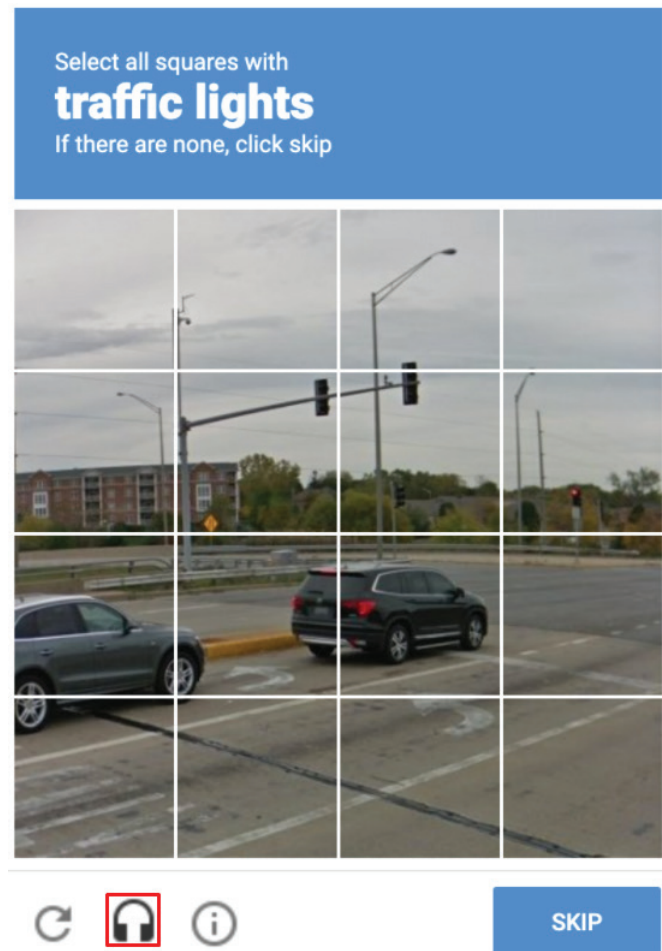


Figure 1: reCAPTCHA v2. An example of a standard object-recognition CAPTCHA, in which users must identify image tiles containing a specified object. While widely deployed, such challenges are now trivially solved by modern multimodal LLMs.

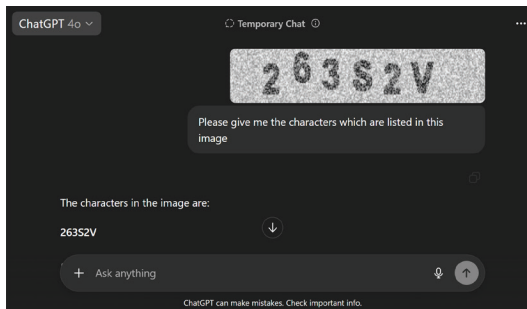
The main goal outlined by CAPTCHAs is to have the robustness to differentiate people from automated agents and bots. However, with the rapid progress in the development of multimodal LLMs, this ability to distinguish between the two types of users has been diminished, thus requiring a critical evaluation of the existing CAPTCHAs.

Today, the most common types of CAPTCHAs include:

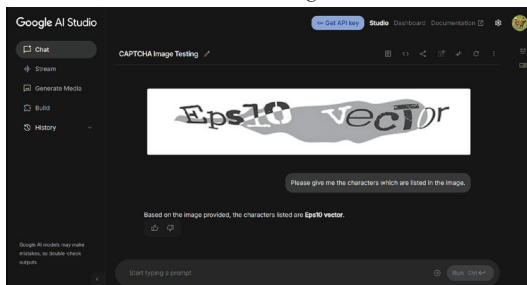
(1) *Text Recognition*: This involves users identifying alphanumeric characters after performing image augmentation techniques to obfuscate the text (e.g., reCAPTCHA v1).

(2) *Object Recognition*: This involves users identifying images containing certain objects, or selecting objects in a predetermined order (e.g., reCAPTCHA v2 in Figure 1).

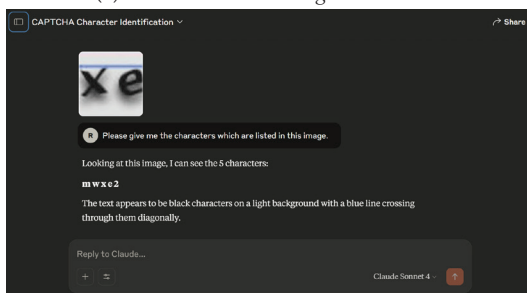
(3) *Logical Puzzles*: This involves users being presented with relatively simple logical reasoning tasks (e.g., filling in a puzzle piece).



(a) ChatGPT-4o solving CAPTCHA.



(b) Gemini 2.5 Pro solving CAPTCHA.



(c) Claude Sonnet 4 solving CAPTCHA.

Figure 2: Comparison of LLMs solving CAPTCHAs. All three frontier models correctly identify distorted alphanumeric characters without apparent difficulty. This confirms that traditional text- and image-based CAPTCHAs no longer present a meaningful barrier to state-of-the-art AI systems.

The traditional CAPTCHA types were once effective since neural networks and AI models lacked the advancement to recognize complicated text in images, thus leveraging the limitations in technology at that time.

However, this assumption no longer holds. Multimodal LLMs such as ChatGPT, Gemini, and Claude have shown little sign of struggling to solve text-based and image-based

CAPTCHAs. As shown in Figure 2, models can accurately deduce distorted text.

■ Our Approach

Our approach explores two novel classes of CAPTCHA challenges designed to exploit the difference between human perceptual abilities and current machine learning systems.

We focus on vision-based and audio-based CAPTCHAs, which make use of intuitive human tasks, leveraging certain characteristics of humans while making it resource-intensive for multimodal LLMs.

Vision-based CAPTCHAs:

Traditional text-based CAPTCHAs relied on augmenting textual data through the use of distortion, color swapping, background modifications, etc. However, with the advances in machine learning models, these tasks have increasingly become trivial for artificial agents to solve.

Humans excel at recognizing structural patterns in images, where traits such as incomplete sections and ambiguous shapes are automatically “filled in” to identify meaningful patterns. However, current visual models struggle when there is high visual diversity along with noisy data, meaning that unconventional text representation through symbolic characters causes a drop in performance in terms of text recognition.

Hypothesis 1. We hypothesize that the recognition of ASCII art text favors human ability rather than vision models. The symbolic nature of ASCII art and the ability for humans to identify general patterns make the task trivial for humans but difficult for deep vision models. Thus, we aim to experiment on vision-based CAPTCHAs using ASCII art text to effectively differentiate humans from bots.

Audio-based CAPTCHAs:

While automatic speech recognition (ASR) systems have achieved very high performance on clean, single-speaker speech, they perform poorly on multi-speaker overlapping audio or speech in noisy auditory environments.

The Cocktail Party Effect, as mentioned by Bronkhorst *et al.*,¹⁴ is an effect where humans are able to filter out a single conversation in a crowd, in spite of different speakers, conversations, and tones. This is contrasted by the current levels of automatic speech recognition (ASR), which merely convert all audio into human-readable speech text. This suggests that overlapping audio could serve as a reliable determinant between human listeners and automated systems.

Hypothesis 2. We hypothesize that identifying meaning in overlapped audio, such as two voices speaking simultaneously, is relatively easy for humans but remains a difficult task for bots. Thus, this forms the basis of our proposed audio-based CAPTCHA design using overlapping audio to distinguish between humans and ML models.

Resource-Intensive CAPTCHAs

Currently, the two methods suggested above are under the assumption that the capabilities of LLMs today are unable to solve the CAPTCHAs. However, we can make use of another

deterrence strategy, which lies in exploiting computational cost asymmetry. In this approach, the objective is to create a task that is computationally inefficient for large-scale automated systems to solve. Thus, this breaks the unit economics of large-scale bot attacks by raising the cost per request above the marginal revenue made from using bots.

This idea parallels the concept of *Proof of Work* in blockchain systems,¹⁵ where miners are required to perform computationally expensive tasks to verify authenticity and prevent “double voting”. Similarly, CAPTCHAs could incorporate tasks that remain trivial for humans yet require disproportionately high computational resources for inference when attempted by automated models. Just as a miner must spend CPU cycles before a block in a blockchain is accepted, a bot must spend inference compute before a CAPTCHA is passed. At a small scale, this cost is trivial; at the scale required for DoS or a mass scraping operation, it becomes a significant operating expense.

Hypothesis 3. We hypothesize that even if models could be trained to solve the CAPTCHAs we propose, the high cost of inference on these more advanced tasks could serve as an effective deterrence against widespread automation.

■ Methods

In this section, we evaluate the 3 types of CAPTCHAs described in the previous section to understand to what extent state-of-the-art models are able to solve them.

Vision-based CAPTCHA:

To evaluate whether ASCII art is a viable method for distinguishing humans from AI agents, we designed a series of experiments to test the performance of state-of-the-art large language models (LLMs) on synthetic ASCII CAPTCHA challenges. We compared model responses against ground-truth labels and measured various notions of accuracy under both raw-text inputs and image inputs.

Data Generation. We generated a synthetic dataset of ASCII-based CAPTCHAs using the Python pyfiglet library, which supports a diverse range of ASCII fonts. Each CAPTCHA contains a random alphanumeric string of 7 to 15 characters, rendered in a randomly selected font from a curated set of 50+ ASCII-compatible and human-readable fonts. A total of 500 samples were created and saved in plaintext format. The text was then optionally rendered as an image using a monospace font to evaluate image-based performance as well.

This ensures that the data generated will be:

- (1) *Diverse* to prevent memorization;
- (2) *Human-Readable* to retain accessibility;
- (3) *Cost-Effective and Scalable* to adhere to effective CAPTCHA generation principles.

Methodology. We implemented an automated testing pipeline to evaluate model performance across multiple models from different providers. These models include: OpenAI's GPT-5.2, Google's Gemini 3 Pro and Flash Preview, Anthropic's Claude Sonnet 4.5, Meta's Llama 4 Maverick, and QwenLM's Qwen3-VL.

Two modes of input were evaluated by the models:

- (1) *Text-based:* The model was presented with plaintext ASCII art and prompted to extract the hidden text.
- (2) *Image-based:* The model was presented with a rendered image of ASCII art and prompted to extract the hidden text. The prompt used was as follows.

“You are taking part in a research experiment on text pattern recognition. The input shown above is a block of ASCII characters that visually represents some alphanumeric text. Extract and return only the exact sequence of visible alphanumeric characters you can read from it. Do not mention images, descriptions, explanations, or uncertainty. Output only the recognized text as a single continuous string, with no spaces, punctuation, or commentary. If the characters are ambiguous, make your best data-driven guess and output only that guess.”

Evaluation. We used various evaluation metrics to decide on the performance of these models:

- (1) *Full Accuracy:* The percentage of exact matches between the ground truth values and predicted values. This is what would be required to actually pass the CAPTCHA. Most humans can achieve this level of performance.
- (2) *Text Similarity:* The normalized Levenshtein ratio (via Python's *difflib.SequenceMatcher*) to measure closeness between the ground truth values and predicted values. This is helpful to understand whether the model at least understood some of the characters.
- (3) *Inference Time:* The average amount of time it took the model to produce an answer.

Audio-based CAPTCHA:

We now turn our attention to the second type of CAPTCHA: conversations that overlap.

In order to test the effectiveness of this audio-based CAPTCHA, we designed experiments to test the performance of state-of-the-art LLMs on synthetic audio challenges.

Data Generation. We used a question-answer dataset as a source of questions and a text-to-speech (TTS) model to generate the audio for the LLMs as contexts for these questions.

For the question-answer dataset, we used the Commonsense-QA dataset by Tel-Aviv University.¹⁶ The format is given by a question as well as 5 answer choices, which rely on prior knowledge of the world to answer the questions.

For the text-to-speech model, we used the XTTS-v2 model provided by Coqui to generate the audio from the questions in the dataset.¹⁷

Methodology. Similarly to our ASCII CAPTCHAs, we implemented a testing pipeline to evaluate model performance across models from different providers. Unfortunately, OpenAI's GPT 5.2 does not support audio input (instead, one needs to transcribe the audio into text and then supply that). However, other OpenAI models can support raw audio input. In particular, the models that we use are OpenAI's GPT Audio Mini, Gemini 3 Flash Preview, and VoxTral Small.

We test these models under 4 different audio environments.

- (1) *Baseline:* The audio generated from the question in text will be directly passed into the LLM.
- (2) *Background Noise:* The audio generated from the question will be combined with a selected clip of noise from a cafe.

(3) *Gaussian Noise*: The audio generated from the question will be combined with noise from a normal distribution.

(4) *Combined Audio*: The audio generated from the question will be combined with two other audio clips at a lower volume.

The LLM was provided with an audio file (MP3 format) that had been processed with the above 4 types of environmental changes. We then provided the following prompt, which contains a set of 5 answers that relate to the conversation in the audio file.

You have been given an MP3 audio file that covers a topic of conversation with potential noise or overlapped voices. Which of the following answers regarding one of the conversations in the audio file is correct?

Answer choices:

A: [Option A]

B: [Option B]

C: [Option C]

D: [Option D]

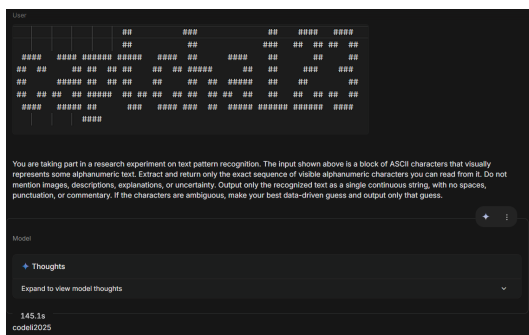
E: [Option E]

Respond only with the correct letter with no form of explanation.

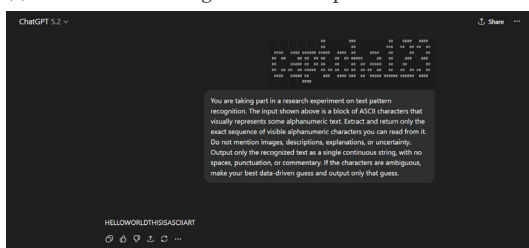
Evaluation. We have a single metric for evaluating accuracy: the percentage of the time that the model selects the correct answer. Note that since there are 5 choices, a model can be right with 20% probability by guessing at random. So a meaningfully higher fraction of correct answers suggests that the model is indeed able to isolate the answer from the provided conversation.

■ Results and Discussion

Visual-based CAPTCHA:



(a) Gemini 3 Pro failing to solve a simple ASCII CAPTCHA



(b) ChatGPT 5.2 failing to solve a simple ASCII CAPTCHA

Figure 3: Flagship LLMs failing to solve ASCII CAPTCHAs. Both ChatGPT 5.2 and Gemini 3 Pro fail to decode a simple ASCII CAPTCHA even with extended reasoning enabled, with Gemini 3 Pro spending over 145 seconds before producing an incorrect answer. This qualitatively corroborates the 0% full accuracy results observed across all models in the automated evaluation.

The results of the model performance on both text-based and image-based ASCII CAPTCHAs are summarized in Tables 1 and 2.

The results are surprising: none of the models are able to solve any samples of either of the two types of ASCII CAPTCHAs. The full accuracy is essentially 0% for all of them. While we perform all of our experiments with the APIs that use default thinking levels and settings for the models, Figure 3 also shows what it looks like when we provide a very simple ASCII CAPTCHA to both ChatGPT 5.2 and Gemini 3 Pro with the highest thinking settings enabled over their respective chat interfaces. They both fail to solve the CAPTCHA, with Gemini 3 Pro spending over 145 seconds and getting only a single character correct.

Table 1: Model performance on ASCII CAPTCHAs (text). All six models achieve 0% full accuracy on text-input ASCII CAPTCHAs, with Gemini 3 Flash Preview recording the highest character-level similarity at 39.38% and still far below any passing threshold.

Model	Result (%)	Similarity (%)	Avg. Response Time (s)
GPT-5.2	0.00	12.50	2.2374
Gemini 3 Flash Preview	0.00	39.38	1.9578
Claude Sonnet 4.5	0.00	19.17	2.5486
Llama 4 Scout	0.00	14.33	0.7801
Qwen-VL 30B	0.00	16.38	4.5290
DeepSeek v3.2-exp	0.00	12.78	84.4913

Table 2: Model performance on ASCII CAPTCHAs (image). Rendering ASCII art as an image does not meaningfully improve performance, with all models again achieving 0% full accuracy (Gemini 3 Flash Preview solves a single sample, 0.16%). Gemini 3 Flash Preview leads on character similarity at 55.48%, suggesting partial structural recognition that nonetheless fails to produce correct outputs.

Model	Result (%)	Similarity (%)	Avg. Response Time (s)
GPT-5.2	0.00	28.20	3.4565
Gemini 3 Flash Preview	0.16	55.48	3.2476
Claude Sonnet 4.5	0.00	19.26	5.8564
Llama 4 Scout	0.00	14.04	2.0810
Qwen-VL 30B	0.00	20.06	1.7943

If we look more closely at Table 1, we see that the best-performing model when we provided the CAPTCHA as raw text was *Gemini 3 Flash Preview* with 39.38% mean character similarity, followed by *Claude Sonnet 4.5* at 19.17%. Correspondingly, in Table 2, we see that *Gemini 3 Flash Preview* also achieved the highest similarity of 55.48% in image-based ASCII CAPTCHAs, followed by *GPT-5.2* with 28.20% similarity.

In terms of how long it takes the generator to convert alphanumeric strings into ASCII art, it takes an average of 0.011s per sample. In addition, it takes negligible time to convert it to images using the Python *Pillow* library.

This confirms that ASCII CAPTCHAs fully satisfy the strict time constraints required for generating CAPTCHAs at scale, which makes it viable for websites and services to generate them dynamically as users or bots visit their pages.

In contrast, when we look at how long it takes the models to solve the CAPTCHAs, we see a huge disparity: *Meta Llama 4 Scout* has the fastest mean response time for text input at 0.7801 seconds, whereas *DeepSeek v3.2-exp* took over 84 seconds. For image-based classification, *Qwen3-VL-30B* achieves the fastest mean response time of 1.7943 seconds, whereas *Claude Sonnet 4.5* is the slowest at 5.85 seconds.

Observations. The results show a clear distinction in performance across all models for both text-based and image-based ASCII CAPTCHAs. *Gemini 3 Flash Preview* performed the best on both text-based inputs and image-based ASCII, suggesting that its spatial understanding and visual processing components are better suited for decoding image representations of ASCII art.

Overall, these results highlight the need for more specialized models to handle ASCII art as a CAPTCHA method, as current systems struggle with this type of visual input. This is surprising to us because of the massive advancements in multimodal processing in the last few years and the fact that ASCII art is widely available and used online in a variety of settings: from discussion forums like Reddit, to READMEs and documentation of code, to command-line tools and games, and even websites built specifically to showcase ASCII art.¹⁸ Hence, we do not believe this is an issue of insufficient samples in the training set.

Instead, we conjecture that the culprit is a combination of a lack of fine-tuning for this specific vision task, in addition to the tokenization and inference of language models.

Text Models (LLMs). State-of-the-art LLMs do not see text as a 2D grid of monospaced characters, but instead see a 1D stream of tokens that are the result of tokenizers like Byte Pair Encoding (BPE),¹⁹ WordPiece,²⁰ and SentencePiece.²¹ These tokenizers aggressively group frequently occurring characters into a single token to save computation. As a result, a vertical line in ASCII art might be built using the pipe character “|”.

Row 1 might be tokenized as [“|”, “text”], whereas Row 2 might be tokenized as [“|”, “verb”], and Row 3 might be [“|”, “end”].

Because the tokens differ based on neighboring characters, the model may have difficulty understanding the global alignment of these characters.

Vision Models (VLMs). Convolutional Neural Networks (CNNs) and Vision Transformers are optimized to detect local features such as texture. But recognizing ASCII Art requires understanding the global structure. Humans rely on Gestalt principles to ignore individual characters and see the larger shape they form.²¹ As such, ASCII Art recognition occurs automatically rather than requiring conscious effort for a human reader. However, when a Vision model sees ASCII art, it sees a lot of noise (sharp edges from letters) rather than the global shape. Existing state-of-the-art models may be classifying the image based on the local features rather than the shape formed by the density of the characters. This is a place where we believe fine-tuning could help.

Audio-based CAPTCHA:

Table 3: Model performance on Audio CAPTCHAs. Under clean baseline conditions, Gemini 3 Flash Preview and VoxTral Small achieve 75.0% and 73.0% accuracy, respectively, well above the 20% random baseline, but performance degrades substantially across all models when background noise, Gaussian noise, or overlapping speech is introduced.

Model	Baseline (%)	Background (%)	Gaussian (%)	Combined (%)	Avg. Response Time (s)
GPT-5.2	46.0	23.0	20.0	27.0	1.71
Gemini 3 Flash Preview	75.0	50.0	59.0	48.0	6.82
Claude Sonnet 4.5	73.0	31.0	46.0	40.0	3.79

Table 3 summarizes the performance of three state-of-the-art audio models. In the baseline condition (clean audio), *Gemini 3 Flash Preview* achieved the highest success rate at 75.0%, closely followed by *VoxTral Small* at 73.0%. *GPT Audio Mini* significantly lagged, correctly solving only 46.0% of the baseline challenges. The introduction of noise worsens the performance of all models, though the degree of degradation varies. *Gemini 3 Flash Preview* demonstrated the highest robustness, with a maximum drop to 48.0% accuracy from the overlapping conversations. *VoxTral Small* showed higher sensitivity to noise, particularly background noise, where its accuracy plummeted from 73.0% to 31.0%. *GPT Audio Mini* struggled to maintain viability in noisy conditions, dropping to as low as 20.0% under Gaussian noise, which is as good as selecting randomly.

Observations. Audio CAPTCHAs present an interesting case in that defeating them may yield broader societal benefits, an argument also articulated in the original CAPTCHA work.¹ If models advance to the point where they can reliably solve our audio CAPTCHA tasks, this would indicate a strong capacity to extract structured, meaningful information from speech, even under noisy conditions — a boon to signal processing at large!

The downside of audio CAPTCHAs is that they are not as effective as our ASCII CAPTCHAs: they are not as accessible as text, and they are more expensive to generate. In particular, we generated the Audio CAPTCHA using the *xTTS* text-to-speech model,¹⁷ where each sample takes an average of 2.1 seconds to generate. Moreover, post-processing of the sample by adding noise or combining conversations needs to be done for the CAPTCHA to work effectively.

Post-processing to add noise or combine the audio files requires an additional 0.005 seconds. At present, it would be costly for sites and services to add dynamically generated audio CAPTCHAs, even though they are effective.

Resource-Intensive CAPTCHAs:

Through the generation and testing of our ASCII and audio CAPTCHAs, we have found that they are effective and demand significant computational power. The best-performing flagship multimodal LLM, the Gemini Flash 3 and Pro 3 line, can only achieve a 50% character similarity score while taking several seconds of computation to do so. This suggests that even if these models are able to get better at solving these CAPTCHAs, they will still spend significant time and resources in order to do so. As such, it might very well be a sufficient deterrent, especially at large-scale deployment.

Discussion:

Our results suggest that ASCII art CAPTCHAs are currently effective against the multimodal LLMs tested in this study. In the short term, ASCII Art appears useful as a CAPTCHA; even ChatGPT-5.2 requires several minutes of reasoning to provide an attempt, which it cannot reliably succeed in. Nevertheless, if ASCII art were deployed commercially, we expect that full and parameter-efficient fine-tuning of visual learning models would eventually yield the ability to solve them. In other words, we are currently unable to articulate a defensible

reason why ASCII art would be “fundamentally” hard for vision models.

In contrast, Audio CAPTCHAs appear to be a reasonable longer-term solution. Given the considerable historical effort that the signal processing community has put into extracting signals from noisy environments, this remains a challenging task for automated systems. The primary downside, however, is the difficulty in scaling the generation of these unique audio challenges compared to text or image-based alternatives.

In light of the assumption that there is no obvious task that is easy for humans to do and hard for computers, the goal may shift toward tasks that are at least computationally expensive. A CAPTCHA does not need to be unsolvable; it only needs to be expensive enough to make bot attacks economically unviable. When the cost of solving each CAPTCHA exceeds the marginal benefit an attacker gains per request, even if individual challenges are technically solvable, the operation is financially unsustainable. Implementing resource-intensive CAPTCHAs provides a safeguard against automated agents for services that wish to remain accessible only to humans.

A limitation of our study is the absence of a quantitative human benchmark. We do not have data on exactly how well humans can solve our specific ASCII and audio CAPTCHAs. However, anecdotal evidence and the visual clarity shown in Figure 3 suggest that the ASCII CAPTCHA is not particularly difficult for humans to solve. Additionally, we recognize that this study only evaluates commercially available LLMs with specific prompt engineering as the benchmark for determining the effectiveness of these forms of CAPTCHAs. Future research should be conducted to examine the ability of fine-tuning either an LLM or a conventional image classification model on a large number of samples of these CAPTCHAs to determine whether a more specialized system can be designed to solve these CAPTCHAs. This mirrors successful attempts to do so for older text-based CAPTCHAs, such as those from Wang *et al.*²³ and Bostik *et al.*²⁴

■ Conclusion

This paper introduces and evaluates two novel CAPTCHA designs: the ASCII text-based CAPTCHA and the overlapping conversation audio-based CAPTCHA, as well as tests their effectiveness against current multimodal LLMs. Through our experimentation, our analysis revealed that state-of-the-art LLMs struggle with deciphering ASCII text- and image-based CAPTCHAs, as well as having relative difficulty in solving audio-based question-answering CAPTCHAs.

These findings suggest that defensive strategies can still be enacted against flagship multimodal LLMs, and further research should be conducted to refine the generation and deployment processes of these CAPTCHAs. Additionally, breaking the unit economics of solving CAPTCHAs can be considered as a useful deterrent compared to the more traditional goal of distinguishing humans from AI bots.

Several directions remain open for future work. Firstly, a controlled human benchmark study is needed to formally measure how quickly and accurately humans solve both the ASCII and

audio CAPTCHAs proposed here. Next, future work should test whether specialized models, such as fine-tuned CNNs, Vision Transformers, or custom ASR pipelines, can be trained to solve these CAPTCHAs robustly. Finally, subsequent research should be conducted to evaluate the long-term viability of these CAPTCHAs in the face of ongoing advancements in multimodal model capabilities.

■ Acknowledgments

I would like to give my heartfelt gratitude to Sebastian Angel for his invaluable guidance, insightful feedback, and continued support throughout the preparation of this paper.

■ References

1. von Ahn, L.; Blum, M.; Hopper, N.; Langford, J. CAPTCHA: Using Hard AI Problems for Security. In *Advances in Cryptology — EUROCRYPT 2003*; Springer: Berlin, 2003; pp 294–311. https://doi.org/10.1007/3-540-39200-9_18
2. Guerar, M.; Verderame, L.; Migliardi, M.; Palmieri, F.; Merlo, A. Gotta CAPTCHA 'Em All: A Survey of 20 Years of the Human-or-Computer Dilemma. *ACM Comput. Surv.* 2022, 54 (9), 192. <https://doi.org/10.1145/3477142>
3. OECD. *Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data*; OECD Artificial Intelligence Papers, No. 33; Organization for Economic Co-operation and Development: Paris, 2025. <https://doi.org/10.1787/d5241a23-en>
4. Wang, J.; Yu, L.; Lui, J. C. S.; Luo, X. Modern DDoS Threats and Countermeasures: Insights into Emerging Attacks and Detection Strategies. *arXiv* 2025, arXiv:2502.19996. <https://doi.org/10.48550/arXiv.2502.19996>
5. Ng, L. H. X.; Carley, K. M. What Is a Social Media Bot? A Global Comparison of Bot and Human Characteristics. *arXiv* 2025, arXiv:2501.00855. <https://doi.org/10.48550/arXiv.2501.00855>
6. Wang, P.; Gao, H.; Guo, X.; Xiao, C.; Qi, F.; Yan, Z. An Experimental Investigation of Text-Based CAPTCHA Attacks and Their Robustness. *ACM Computing Survey* 2023, 55 (9), 196. <https://doi.org/10.1145/3559754>
7. Douceur, J. R. The Sybil Attack. In *Peer-to-Peer Systems*, Druschel, P.; Kaashoek, F.; Rowstron, A., Eds.; Springer: Berlin, 2002; pp 251–260. https://doi.org/10.1007/3-540-45748-8_24
8. Naor, M. Verification of a Human in the Loop or Identification via the Turing Test; Weizmann Institute of Science. <https://www.wisdom.weizmann.ac.il/~naor/PAPERS/human.pdf>
9. Pryor, L.; Mallet, J.; Dave, R.; Seliya, N.; Vanamala, M.; Boone, E. S. Evaluation of a User Authentication Schema Using Behavioral Biometrics and Machine Learning. *arXiv* 2022, arXiv:2205.08371. <https://doi.org/10.48550/arXiv.2205.08371>
10. Davidson, A.; Goldberg, I.; Sullivan, N.; Tankersley, G.; Valsorda, F. Privacy Pass: Bypassing Internet Challenges Anonymously. *Proceedings on Privacy Enhancing Technologies*. 2018, 2018 (3), 164–180. <https://doi.org/10.1515/popets-2018-0026>
11. Wang, P.; Gao, H.; Shi, Z.; Yuan, Z.; Hu, J. Simple and Easy: Transfer Learning-Based Attacks to Text CAPTCHA. *IEEE Access* 2020, 8, 59044–59058. <https://doi.org/10.1109/ACCESS.2020.2982945>
12. Plesner, A.; Vontobel, T.; Wattenhofer, R. Breaking reCAPTCHA v2. In *2024, IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*; IEEE: 2024; pp 1047–1056. <https://doi.org/10.1109/COMPSAC61105.2024.00142>
13. Tariq, N.; Khan, F. A.; Moqurrah, S. A.; Srivastava, G. CAPTCHA Types and Breaking Techniques: Design Issues, Challenges,

- and Future Research Directions. *arXiv* 2023, arXiv:2307.10239. <https://doi.org/10.48550/arXiv.2307.10239>
14. Bronkhorst, A. W. The Cocktail-Party Problem Revisited: Early Processing and Selection of Multi-Talker Speech. *Atten. Percept. Psychophys.* 2015, 77 (5), 1465–1487. <https://doi.org/10.3758/s13414-015-0882-9>
 15. Jakobsson, M.; Juels, A. Proofs of Work and Bread Pudding Protocols (Extended Abstract). In *Secure Information Networks: Communications and Multimedia Security IFIP TC6/TC11 Joint Working Conference on Communications and Multimedia Security (CMS'99)*, September 20–21, 1999, Leuven, Belgium; Preneel, B., Ed.; Springer: Boston, MA, 1999; pp 258–272. https://doi.org/10.1007/978-0-387-35568-9_18
 16. Talmor, A.; Herzig, J.; Lourie, N.; Berant, J. Commonsense-QA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of NAACL-HLT 2019*; Association for Computational Linguistics: 2019; pp 4149–4158. <https://doi.org/10.18653/v1/N19-1421>
 17. Casanova, E.; Davis, K.; Gölge, E.; Gökner, G.; Gulea, I.; Hart, L.; Aljafari, A.; Meyer, J.; Morais, R.; Olayemi, S.; Weber, J. XTTS: A Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*; International Speech Communication Association (ISCA): 2024; pp 4978–4982. <https://doi.org/10.21437/Interspeech.2024-2016>
 18. ASCII Art Archive – Home of ASCII Art. <https://www.asciart.eu> (accessed 2025)
 19. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings ACL 2016*; Association for Computational Linguistics: 2016; pp 1715–1725. <https://doi.org/10.18653/v1/P16-1162>
 20. Song, X.; Salcianu, A.; Song, Y.; Dopson, D.; Zhou, D. Fast Word-Piece Tokenization. In *Proceedings of EMNLP 2021*; Association for Computational Linguistics: 2021; pp 2089–2103. <https://doi.org/10.18653/v1/2021.emnlp-main.160>
 21. Kudo, T.; Richardson, J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of EMNLP 2018 (System Demonstrations)*; Association for Computational Linguistics: 2018; pp 66–71. <https://doi.org/10.18653/v1/D18-2012>
 22. Wagemans, J.; Elder, J. H.; Kubovy, M.; Palmer, S. E.; Peterson, M. A.; Singh, M.; von der Heydt, R. A Century of Gestalt Psychology in Visual Perception: I. Perceptual Grouping and Figure-Ground Organization. *Psychol. Bull.* 2012, 138 (6), 1172–1217. <https://doi.org/10.1037/a0029333>
 23. Wang, P.; Gao, H.; Shi, Z.; Yuan, Z.; Hu, J. Simple and Easy: Transfer Learning-Based Attacks to Text CAPTCHA. *IEEE Access* 2020, 8, 59044–59058. <https://doi.org/10.1109/ACCESS.2020.2982945>
 24. Bostik, O.; Horak, K.; Kratochvila, L.; *et al.* Semi-Supervised Deep Learning Approach to Break Common CAPTCHAs. *Neural Computing and Applications* 2021, 33, 13333–13343. <https://doi.org/10.1007/s00521-021-05957-0>

■ Author

Chong Choon-Hou Rafael is a JC 2 student at Hwa Chong Institution with a strong interest in Machine Learning, Business, and Economics. He is passionate about exploring how technology can be applied to improve decision-making, drive innovation, and solve real-world problems. By combining technical knowledge with an understanding of markets and society, Rafael hopes to create meaningful solutions that generate lasting positive impact.

Impacts of Nanoplastics on Molecular Mechanisms within Microglia

Ella Wilce

Otumoetai College, 105 Windsor Road, Bellevue, Tauranga, 3110, New Zealand; ellajwilce@gmail.com

ABSTRACT: The ubiquity of nanoplastics (NPs) in the environment and organisms poses increasing threats to humans. NPs' minuscule size allows them to penetrate biological barriers; NPs have been detected in the human placenta, lungs, and brain. In the brain, NPs are preferentially internalized by surveilling microglia, resident immune cells, for destruction. This uptake of NPs induces physiological changes in microglia, including a transition to an active, pro-inflammatory state and programmed cell death, thereby threatening homeostasis. Damage to microglia is implicated in neurodegenerative diseases such as Alzheimer's, Parkinson's, and Multiple Sclerosis. Therefore, understanding the mechanisms underlying NP-microglia interactions is critical for quantifying the risks NPs pose to humans, informing consumer and manufacturer policies, and developing potential therapeutic interventions. In this review, studies on NPs and microglia are corroborated and analyzed to provide a roadmap detailing how NPs affect these mechanisms. Lysosomal dysfunction induced by NPs is a central trigger for microglia reactivity, causing the release of reactive oxygen species (ROS), calcium ions (Ca^{2+}), and other molecules. These induce metabolic reprogramming and phosphorylation cascades resulting in a pro-inflammatory response and programmed cell death in microglia. Furthermore, NPs elicit structural changes in proteins in the brain parenchyma, resulting in misfolded aggregates that impose an additional burden on microglia. This may contribute to or exacerbate existing disruptions caused by NPs within microglia. These findings suggest that NPs disrupt several essential mechanisms and pose a relevant threat to microglia.

KEYWORDS: Environmental Neurology, Cellular Pathways, Nanoplastics, Microglia, Neuroinflammation.

■ Introduction

Over the last century, more than 10 billion tonnes of plastic have been produced globally.^{1,2} Of these, 29% have been recycled or incinerated,² leaving over 7 billion tonnes of plastic to decompose into micro and nano particles. Nanoplastics (NPs) are defined as particles ranging from 1 to 1000 nm as a result of industrial plastic degradation.³ At this scale, plastic particles are able to infiltrate a wider range of biological barriers impenetrable to larger polymers, including the blood-brain barrier (BBB).⁴ The NPs of nylon, polyethylene terephthalate (PET), high-density polyethylene (HDPE), polyvinyl chloride (PVC), polypropylene (PP), and polystyrene (PS) have been found in the brains of humans.^{5,6} Though less prevalent, polymethyl methacrylate (PMMA) has also been studied due to its high concentration within people with surgical implants.^{7,8} Alarmingly, one study detected a median of 4917 $\mu\text{g/g}$ of NPs in human brains. This concentration is more than 10 times that detected in the kidneys or liver, suggesting that the brain's high lipid content and lack of clearance capacity confer susceptibility to NPs retention.⁵ NPs interact with neurons and proteins in the brain, threatening neural function.⁹ Murine studies found that exposure to NPs, proportional to the amount of NPs humans are exposed to, incurred cognitive deficiencies such as impaired learning and memory in mazes.¹⁰ Therefore, understanding how NPs affect various cells of the brain is imperative.

To clear accumulating NPs in the brain, NPs are engulfed by resting microglia, immune macrophages that maintain neural homeostasis.¹¹ In their resting state, microglia have many

sprawled-out arms, known as processes, with receptors on them, allowing them to constantly survey surrounding brain tissue for signals of unwanted particles.¹² When activated by external stimuli, microglia incur physiological and morphological changes, adopting various phenotypes to elicit an appropriate immune response.^{13,14} These changes are highly dynamic, with microglia regularly alternating between neurodegenerative (M1) and neuroprotective (M2) phenotypes.¹⁵ However, if microglia are in a neurodegenerative state for extended periods of time, excessive inflammation and programmed cell death of microglia can lead to irreversible damage.¹¹

Numerous studies demonstrated that NP exposure is correlated with microglial inflammation and cell death.^{4,7,16} Lysosome damage, signaling pathways, oxidative stress, metabolic changes, and protein misfolding were identified as contributing mechanisms.^{7,17–19} However, the interplay of these mechanisms is yet to be elucidated, with studies tending to focus on a single mechanism or pathway. Many studies focus on the brain-wide effect of NPs on neurodegeneration or solely on the intracellular effects of NPs on neurons. Therefore, there lacks a clear review of how these pathways facilitate microglial responses.

To delineate these pathways, this review first considers entry routes to the brain, as these may influence which brain regions are most affected. Direct uptake of NPs by microglia is then explored, with consideration of lysosomal capacity and function, followed by an analysis of ensuing ramifications to key components of microglial function, such as mitochondria and signaling pathways. Subsequently, the effect of NPs on protein misfolding and the additional burden this imposes on microg-

lia is evaluated. By integrating these mechanisms, this review aims to identify the interconnected pathways that determine the fate and function of microglia exposed to NPs. As NPs induce cognitive impairment in mice and marine organisms, elucidating the mechanisms underlying microglial responses is a critical step towards targeted mitigation and prevention of the cognitive threats posed by NPs.

■ Discussion

Entry Routes to the Brain – Olfactory System, BBB, and Biomolecular Corona:

NPs are internalized by three known pathways: ingestion, inhalation, and dermal exposure.^{20,21} Significant quantities of NPs have been detected in decedent human brains, specifically in the hippocampus,²² prefrontal cortex,^{22,23} and olfactory bulb.²⁴ Though exact routes remain unknown, two key routes are proposed as being exploited: the olfactory system and the BBB.

Airborne NPs likely enter the brain through the olfactory system,^{24,25} a passage that allows the transport of odor molecules into the olfactory bulb. Olfactory neurons originating in the nasal cavity enter the olfactory bulb via multiple holes known as foramina on the cribriform plate.²⁴ Each foramen ranges from around 7.2 mm² to 3.8 mm².²⁶ Odor molecules are internalized by these neurons and transported up their axons into the olfactory bulb.²⁷ This axonal transport pathway is used in nanoparticle drug delivery,²⁸ and offers a plausible route for NPs to gain direct entry to the brain. Supporting this, Micro-Fourier Transform Infrared Spectroscopy revealed microplastics (MPs) in the olfactory bulbs of 8 out of 15 human brains.²⁴ Sixteen distinct polymer particles were identified, with polypropylene comprising 45%, underscoring the need to study a diversity of plastic types.²⁴ Although detection sensitivity was insufficient to confirm NPs' presence, these findings highlight the significance of the olfactory system for ongoing research. As NPs permeate a wider field of biological barriers than MPs, it is likely that NPs would be present in the brains of more people than this study detected. Future research should develop technologies to accurately quantify NP levels across brain regions to better understand health implications.

An alternative route for NPs is the BBB – a semipermeable membrane surrounding the brain that facilitates the movement of molecules and ions in and out of the brain.²⁹ The BBB is lined with endothelial cells sealed with tight junctions composed of structural proteins that regulate paracellular diffusion.³⁰ As small, lipophilic particles, it is likely that NPs bypass tight junctions altogether, instead diffusing transcellularly through endothelial cells to gain direct access to the brain, a process known as transcytosis.³¹

Kopatz *et al.* suggest nanoplastic transport is enhanced through the formation of a lipid corona.³² Computer simulations using polystyrene compared pristine plastic to particles with a corona of lipid (cholesterol) or protein (PDB ID: 1LE1). NPs with a lipid corona were observed to diffuse into the lipid bilayer, where the corona dissociated, leaving the nanoplastic to traverse into the brain via diffusion or aided by a vesicle.³² Conversely, a protein corona did not allow permeation of the

BBB. It is important to note that a biomolecular corona does not comprise just one molecule, as in the aforementioned simulations. Upon soaking PS and PVC in solutions of human plasma, Abdolahur Monik *et al.* identified 268 proteins in the coronas of PS and PVC.³³ This demonstrates that BBB permeation by NPs exists on a spectrum, increasing in accordance with the proportion of lipids in the corona. As environmental plastics often contain chemical additives, emerging studies are investigating the coronas formed by not only different plastic types but also different chemical additives.^{34,35} Unique results for plastic and chemical variants reinforce the need to study a diverse range of NPs beyond those manufactured in a lab.

NPs Cause Lysosome Permeability and Rupture:

Upon entering the brain, NPs passively diffuse throughout the interstitial fluid, the fluid between neurons and glial cells.⁵ Nanoplastic movement through the brain parenchyma is influenced by the particle's size and structure.³⁶ Microglia are responsible for the degradation of NPs. To do so, NPs are internalized by microglia via phagocytosis.³⁷ Once internalized, NPs may interfere with several key organelles, including lysosomes, mitochondria, and endoplasmic reticulum.

Phagocytosis is a common transport mechanism among cells (for nanosized particles) whereby the plasma membrane of the microglia invaginates, allowing it to engulf the particle in a vesicle, referred to as an endosome.³⁸ The endosome then traffics the nanoplastic to the lysosome, an organelle contained by all cells to facilitate degradation of unwanted particles (autophagy).³⁹ The abundance of lysosomes in microglia and signaling complexity allows for more effective disposal of pathogens in microglia.⁴⁰

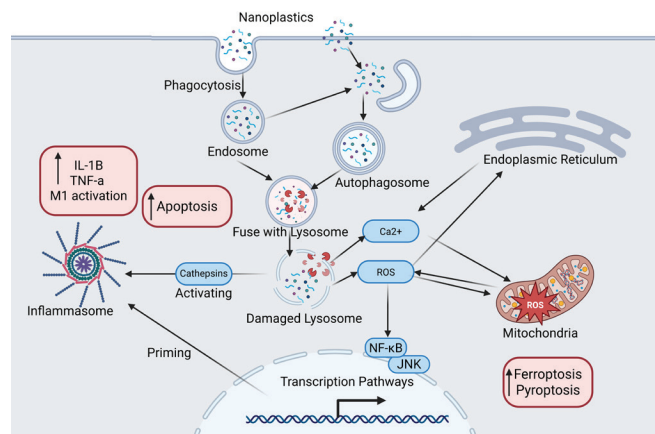


Figure 1: Summary of the intracellular pathways impacted by nanoplastics. Boxes in red indicate microglial responses. Created using BioRender.

NPs may also reside in the microglial cytosol, either by diffusing through the plasma membrane or by escaping endosomes.⁴¹ When this occurs, they are detected by autophagosomes, vesicles with a double-walled membrane. Autophagosomes may fuse with endosomes to form amphisomes, which fuse with lysosomes. These two pathways lead to the majority of phagocytized NPs residing in lysosomes (Figure 1). Yu *et al.* discovered that lysosomal membrane permeability increased in a dose-dependent manner in nanoplastic-treated BV2 microglia cells, indicated by an increase in a marker protein

(Galectin-3) observed via immunofluorescent imaging.⁴² Furthermore, LAMP1, an indicator of impaired lysosomal function, also increased.⁴² Lysosome permeability and rupture release small molecules such as ROS and Ca^{2+} , and larger enzymes such as cathepsins, into the cytosol, respectively.⁴³ These trigger downstream pathways affecting transcription factors, pathways, and metabolic reprogramming (Figure 1). In the above study, this culminated in microglial pyroptosis.

Released Lysosomal Content Affects Mitochondria and Inflammasome to Reprogram Microglia:

The release of ROS in microglia causes Endoplasmic Reticulum (ER) stress, prompting the release of Ca^{2+} . This creates a self-perpetuating cycle, as Ca^{2+} causes the formation of further ROS in mitochondria.⁴⁴ ROS and Ca^{2+} widen the mitochondrial permeability transition pore, thereby decreasing mitochondrial membrane potential ($\Delta\Psi\text{m}$). To date, no research exists examining the effects of NPs on microglial mitochondria; nonetheless, studies on neurons, neurofibroblasts, and peripheral macrophages offer potential insights as to how microglial mitochondria may be implicated.⁴⁵ Florance *et al.* found that exposure to NPs increased the generation of mitochondrial ROS and decreased mitochondrial membrane potential ($\Delta\Psi\text{m}$) in both human (THP-1) and murine (RAW 264.7) macrophages *in vitro*.⁴⁶ The study used PS NPs treated with anionic surfactants to mimic those commonly found in cosmetics.⁴⁶ A separate study in THP-1 cells using pristine PS NPs similarly demonstrated increased oxidative stress and reduced $\Delta\Psi\text{m}$, hence indicating the potential threat of NPs regardless of surface modifications.⁴⁷ In murine macrophages exposed to 10 μm PS MPs, cells decreased mitochondrial metabolism and $\Delta\Psi\text{m}$, instead utilizing glycolysis, which was demonstrated to support an immunometabolic response.⁴⁸ As microglia share similar functions with macrophages, the risks identified for macrophages hold potential relevance for microglia; however, each has unique responses to internal stimuli adapted to its respective environment that should not be overlooked. Of note, cultured macrophages may exhibit an enhanced response to stimuli, potentially exaggerating the effects of NPs.⁴⁹ A study investigating the roles of microglia and peripheral macrophages in spinal cord injury reported elevated apoptosis and reduced phagocytic efficiency compared to microglia, suggesting microglia maintain a more regulated response *in situ*. Therefore, these results should serve solely as a foundation for future research to build upon through microglia-specific studies using *in vivo* and *in vitro* analysis. If substantiated, this would carry important implications for understanding NPs' role in microglial dynamics.

Glucose metabolism is proposed to exert transcriptional control over microglial activation.⁵⁰ In a resting state, microglia preferentially utilize oxidative phosphorylation to produce energy within mitochondria.⁵¹ However, under inflammatory conditions, such as those caused by NPs, microglia switch to glycolysis, a faster, less efficient process that bypasses the mitochondria, which facilitates enhanced phagocytosis and inflammatory response.⁵² This causes pro-apoptotic signaling,

contributing to cell death in addition to a shift towards chronic microglial activation.^{50,52,53}

Furthermore, ROS act in conjunction with cathepsins released during lysosomal rupture, thereby serving as damage-associated molecular patterns (DAMPs).⁵⁴ DAMPs mediate the second stage of inflammasome activation; the first, priming stage, results from the NF- κB pathway, which will be discussed in the following section regarding transcription pathways.^{55,56} Inflammasomes are protein complexes composed of receptors and adaptors that evoke an inflammatory response by inducing pro-inflammatory caspases and cytokines.⁵⁵ Cathepsin B is predominantly implicated in NLRP3 inflammasome activation; however, a study using various compounds known to induce lysosome membrane permeability further ascertained the involvement of cysteine cathepsins B, C, L, S, and Z released from the lysosome in NLRP3 inflammasome activation.^{57,58} The NLRP3 inflammasome facilitates the maturation of pro-inflammatory cytokines IL-18 and IL-1 β , contributing to sustained microglial activation seen in Alzheimer's disease.

Transcription Pathways Triggered by Nanoplastic-Induced Cytokines and ROS:

Microglial activation is a complex process that can occur through multiple pathways, predominantly triggered by NP-induced ROS.⁶⁰ To determine these specific pathways, existing studies follow a similar structure: first, genes are analyzed across cells exposed to varying levels of NPs. Second, the data are processed, and a common pathway is identified that is shared across all samples. Third, that pathway is suppressed, and proinflammatory markers are measured to validate their findings. Interestingly, activation of pattern-recognition receptors (PRRs) showed minimal increase in microglia exposed to NPs compared to intestinal epithelial cells.⁶¹ This is likely due to microglia containing higher basal levels of PRRs compared to other cells, to facilitate their specialized role in pathogen detection and removal. Despite this, Antunes *et al.* found that N9 microglia exposed to NPs displayed more pronounced cytokine upregulation and cell death than intestinal epithelial cells, suggesting higher vulnerability to NPs.⁶¹ It is therefore critical to understand the signaling pathways specific to microglia, as their response to NPs is distinct from that of other cells.

Studies have shown that nuclear factor κB (NF- κB) is the predominant signaling pathway for nanoplastic-induced microglial activation. NF- κB can be segregated into two pathways: canonical, which utilizes the transcription factor p50/p65, and non-canonical, which uses p52/RELB. A further pathway, atypical NF- κB , induces the same transcription factors as the canonical pathway but uses different upstream mediators. These transcription factors serve to induce an inflammatory response, with the canonical pathway acting immediately and the non-canonical pathway eliciting a slower, sustained response. NPs have been found to increase nuclear translocation of both p50/p65 and p52/RELB via non-canonical and atypical pathways, explained below.

Li *et al.* found that NPs induced nuclear translocation of p50/p65 in BV2 mouse microglia via protein kinase RNA-like

endoplasmic reticulum kinase (PERK).⁶² ROS oxidizes the Harvey rat sarcoma viral oncogene homolog (H-Ras) protein, identified as a central node among overlapping differentially expressed genes (DEGs).⁶² This led to H-Ras activation, which triggers the PERK pathway. Among cells treated with NPs, H-Ras knockdown effectively downregulated the PERK pathway, confirming this correlation. The PERK pathway ultimately inhibits the synthesis of the NF- κ B suppressor I κ B α , thereby freeing NF- κ B to translocate to the nucleus and regulate proinflammatory genes.^{63,64} Removal of I κ B α is the same trigger for NF- κ B in the canonical pathway, where NF- κ B activation is mediated by the direct phosphorylation of I κ B α .⁶⁴ Hence, the same downstream transcription factors, p50/p65, are produced, eliciting the same acute inflammatory response as the canonical pathway.

Another study found NPs triggered the non-canonical NF- κ B pathway. Lu *et al.* performed transcriptome sequencing analysis (RNA sequencing) of the prefrontal cortex of mice that had been fed different amounts of PS NPs.⁶⁵ Four DEGs were present across all concentrations of NPs: Pbx3, Ecell, Crb1, and Ngb. ROS also increased, although it is unclear whether ROS mediated these genetic changes. Regardless, these DEGs upregulated the NIK/NF- κ B (non-canonical) signaling pathway by stabilizing the protein NF- κ B-inducing kinase (NIK), which is normally degraded.^{65,66} NIK functions in conjunction with inhibitor of NF- κ B kinase α (IKK α) to phosphorylate the NF- κ B2 precursor protein (p100) in the cytosol.⁶⁶ This causes p100 to split into RelB/ p52 (types of NF- κ B), which then translocate to the nucleus to transcribe pro-inflammatory genes.^{65,66} Supporting this, the same transcriptome analysis showed significant upregulation of TNF α and IL-1 β , two prominent cytokines in inflammation.⁶⁵ The p52 pathway is slow and sustained compared to the canonical NF- κ B pathway, contributing to chronic inflammation. This highlights a potential for NPs to contribute to neurodegenerative diseases.

Sun *et al.* identified an additional pathway for NPs in microglia cells, whereby ROS activated c-Jun N-terminal kinase (JNK), an enzyme that phosphorylates transcription factors that upregulate heme oxygenase (HO-1).¹⁸ HO-1 is an enzyme that breaks down heme into iron (and CO and bilirubin). In small amounts, iron can be used as a signal; however, if it is not stored by FTH1, iron can lead to cell death (ferroptosis),⁶⁷ which was seen in the study by Sun *et al.* JNK can also be activated as a downstream pathway of NF- κ B, representing a potential flow of pathways.

These pathways represent opportunities for genetic modification to suppress the effects of NPs. However, cross-talk between pathways in regard to NPs is unclear, and it is likely that many more pathways exist that have not yet been tested. Future research should continue to explore signaling pathways.

Protein misfolding caused by NPs activates microglia:

NPs penetrate brain cells and interfere with protein folding,^{68–71} culminating in protein aggregates and fibrils.⁷⁰ These cannot be digested by microglia, so incur chronic microglial activation, mirroring the pathology of Alzheimer's Disease. Protein misfolding likely results from bonds between NPs and

the free ends of proteins that have yet to be folded.^{68–71} Studies to date have used in silico computer simulations, which provide a baseline to be supported with future *in vitro* studies.

To elucidate the nature of these interactions, Holl  czki and Gehrke used discrete molecular dynamics (DMD) simulations to explore polyethylene (PE), polypropylene (PP), polyethylene terephthalate (PET), and nylon-6,6 (N66) in aqueous solution in the presence of various amino acids.⁶⁸ They identified Phenylalanine and Tryptophan as being highly prone to adsorption on the NPs' surface due to nonpolar attractions.⁶⁸ Further DMD simulations utilized by Wang *et al.* to pinpoint the interactions between amyloid- β and Polystyrene NPs similarly revealed adsorption of Phenylalanine and Tryptophan.⁷¹ Side chains of the amyloid- β that contained amino acids with rings on their side chains (known as aromatic amino acids), specifically Phenylalanine, Tyrosine, and Tryptophan, formed π - π interactions with the benzene ring of PS NPs.⁷¹ This strengthened hydrogen bonding, which occurred between the polar amino acids on the N-terminus (initial free amino group of the polypeptide) of amyloid- β and the hydroxyl groups of NPs.⁷¹

An initial misfolded protein acts as a template for other proteins, thereby spreading the malformed structure to surrounding proteins. These misfolded proteins form small clusters known as oligomers. As oligomers further accumulate, they form strong, insoluble amyloid fibrils.⁷² These can include amyloid- β peptides, the prion protein, α -synuclein, tau, and β 2-microglobulin.⁷⁰ Increased amyloid formation may accelerate the effects of neurodegenerative diseases.⁷¹ Amyloid fibrils cannot be phagocytized by microglia. Instead, microglia digest amyloid fibrils extracellularly via digestive exophagy, whereby tightly sealed regions called lysosomal synapses are formed between the microglia and amyloid fibril through which lysosomal contents are released to degrade the fibrils.⁷³ In a resting state, the pH of microglial lysosomes is $\sim$ 6; this decreases to $\sim$ 5 in active microglia,⁷⁴ the upper limit of the pH at which amyloid- β is degraded.⁷⁵ Therefore, lysosomal alkalization resulting from direct nanoplastic uptake may limit the clearance of these fibrils, contributing to further recruitment and activation of microglia. Together, these result in sustained M1 inflammation of microglia, contributing to diseases such as Alzheimer's and Parkinson's disease.⁷⁶

Limitations and Future Directions:

The past 5 years have seen a marked increase in research on the health implications of NPs. However, technical challenges remain a block in the progression of research, in particular, detection and quantification methods. As such, future research should work on developing these technologies.

PS NPs comprise 94 percent of studied NPs, yet comprise just 8.5 percent of environmental NPs.⁷⁷ Furthermore, many studies use spherical polystyrene beads, which have the lowest possible surface area, thus distorting nanoscale interactions. Given the structural diversity of environmental NPs, future research should prioritize developing the technology needed to adequately assess a range of NP types, shapes, and sizes. Koo *et al.* exemplified this in an *in vitro* study of microglia ex-

posed to PS, PP, or PMMA, where PMMA showed the most pronounced effects on the expression of chemotactic migration and microglia recruitment.⁷ As PMMA comprises a low percentage of environmental plastics, it is rarely studied. This reinforces the need to examine a range of plastic types, as, beyond this, PMMA has not been used in neurotoxicity studies.

Type diversity is of particular relevance to protein misfolding, due to the differential bonds that form. For instance, the benzene rings of PS NPs reinforced hydrogen bonds with amino acids; however, PS and PET are the only NPs that contain a benzene ring, so the precise interactions of other nanoplastics are unclear. All studies to date use *in silico* analysis, which cannot fully account for the multitude of background bonds formed by the environment; future studies should utilize a combination of *in vitro* and *in vivo* methods to confirm these findings.

A further limitation is that this study focused predominantly on the intracellular effects of NPs on microglia; therefore, it did not account for the effects of cross-talk on microglia. Prior studies have highlighted the plausibility of cytokine signaling in macrophages outside the brain, affecting microglial activation via the gut/brain axis.^{78,79} As NPs have been shown to affect these regions (e.g., LPS in macrophages), one may speculate that cross-talk plays a significant role in exacerbating their effects. Future studies should examine cross-talk among regions, especially those where an immune response has previously been identified, such as the intestines,^{80,81} colon,⁸² liver,⁸³ and lungs.⁸⁴ Similarly, this study has not gone into the depths of how cross-talk with neurons or astrocytes implicates disease, a field worth future research.

Considering these limitations, future research should prioritize utilizing more accurate detection techniques, different types, additives, and shapes of NPs that mimic environmental NPs, more microglia-specific studies, and pathways not yet explored. Developing technology that can reliably isolate NPs in the environment would particularly benefit understanding the type and additive differences. Once specific risks have been established, research should turn to therapeutic strategies. The microglial response highlighted in this paper shares pathologies with those observed in Alzheimer's Disease; therefore, strategies relevant to Alzheimer's, such as the use of inhibitor proteins to block signaling pathways triggered by NPs,¹⁸ may apply to NPs. In addition, nanorobots are currently being explored as a targeted drug delivery mechanism for brain cancers and neurodegenerative disease,⁸⁵ and may offer a solution for the removal of NPs from the brain.⁸⁶

■ Conclusion

NPs represent an emerging ecological threat demonstrating potential neurotoxicity. The ubiquity and environmental persistence of NPs, combined with the toxicity to microglia outlined in this review, highlight the urgent need for future research. In humans, NPs enter the brain via multiple routes of exposure, with their size and structure facilitating absorption, diffusion, and infiltration across biological barriers. This review explored a pathway by which NPs cause chronic microglial activation and cell death. It found that NPs in the brain's in-

terstitial fluid are selectively phagocytosed by microglia, which transport them intracellularly to lysosomes. Because lysosomes are incapable of degrading NPs, lysosomal membrane permeabilization and rupture occur, releasing cathepsins, metals, and ROS into the cytosol. These impair mitochondrial and ER function, altering the physiological state of the microglia. NF- κ B was identified as a key signaling pathway that facilitates microglial apoptosis, with JNK also implicated. These pathways share commonalities with neurodegenerative diseases, underscoring the potential threat of NPs. Outside of microglia, NPs impair protein folding and support amyloid fibril aggregation, placing further burden on inflammatory microglia. Together, these findings provide a pathway for understanding how NPs interact with microglia, establishing a foundation for future research to further identify mechanisms, quantify NP risk, and develop therapeutic strategies to mitigate potential neurological threats.

■ Acknowledgments

I would like to acknowledge the team at the Indigo Research Intensive Summer (IRIS) program for their support in writing my paper, in particular Dr. Jorge Avila, Bre Calhoun, Dr. Audrey Wozniak, and Dr. Daniel Kim, for their guidance and feedback. I attest that the ideas, graphics, and writing in this paper are entirely my own.

■ References

1. Geyer, R.; Jambeck, J. R.; Law, K. L. Production, Use, and Fate of All Plastics Ever Made. *Sci. Adv.* **2017**, *3* (7), e1700782.
2. Liu, C.; Geyer, R.; Hu, S. 100 Years of Plastic--Using the Past to Guide the Future. *ArXiv Prepr. ArXiv241113618* **2024**.
3. Gigault, J.; Halle, A. ter; Baudrimont, M.; Pascal, P.-Y.; Gauffre, F.; Phi, T.-L.; El Hadri, H.; Grassl, B.; Reynaud, S. Current Opinion: What Is a Nanoplastic? *Environ. Pollut.* **2018**, *235*, 1030–1034. <https://doi.org/10.1016/j.envpol.2018.01.024>.
4. Shan, S.; Zhang, Y.; Zhao, H.; Zeng, T.; Zhao, X. Polystyrene Nanoplastics Penetrate across the Blood-Brain Barrier and Induce Activation of Microglia in the Brain of Mice. *Chemosphere* **2022**, *298*, 134261. <https://doi.org/10.1016/j.chemosphere.2022.134261>.
5. Nihart, A. J.; Garcia, M. A.; El Hayek, E.; Liu, R.; Olewine, M.; Kingston, J. D.; Castillo, E. F.; Gullapalli, R. R.; Howard, T.; Bleske, B.; Scott, J.; Gonzalez-Estrella, J.; Gross, J. M.; Spilde, M.; Adolph, N. L.; Gallego, D. F.; Jarrell, H. S.; Dvorscak, G.; Zuluaga-Ruiz, M. E.; West, A. B.; Campen, M. J. Bioaccumulation of Microplastics in Decedent Human Brains. *Nat. Med.* **2025**, *31* (4), 1114–1119. <https://doi.org/10.1038/s41591-024-03453-1>.
6. Wang, T.; Yi, Z.; Tan, Y.; Huang, Y.; Li, T.; Lu, W.; Zhuang, W.; Guo, S. Detection of Micro- and Nanoplastics in Cerebrospinal Fluid and Blood: Implications for Brain Diseases. *Ecotoxicol. Environ. Saf.* **2026**, *309*, 119614. <https://doi.org/10.1016/j.ecoenv.2025.119614>.
7. Koo, J.; Jeong, B.; Baek, J. Y.; Lee, W. S.; Gong, J.; Park, S.; Hong, J.; Sim, Y.; Kim, D. S.; Kim, S. R. Type-Dependent Effects of Nanoplastics on Microglial Activation and CXCR2-Mediated Chemotactic Migration. *Nanoscale* **2025**, *17* (29), 17274–17284.
8. Burcea, A.; Bănăţeanu, A.-M.; Poalelungi, C.-V.; Forna, N.; Cumpăţă, C. N. Enhanced Properties and Multifaceted Applications of Polymethyl Methacrylate (PMMA) in Modern Medicine and Dentistry. *Rom J Oral Rehabil* **2024**, *16* (4), 108–123.

9. Kumari, S.; Begum, M. Y.; Chinglenthoba, C.; Aran, K. R.; Panda, S. P.; Abomughaid, M. M.; Lakhnopal, S.; Avinash, D.; Jha, N. K.; Gupta, R. Deciphering the Neurotoxic Burden of Micro- and Nanoplastics: From Multi-Model Experimental Evidence to Therapeutic Innovation. *Mol. Neurobiol.* **2025**, *62* (12), 15295–15317.
10. Wang, G.; Lin, Y.; Shen, H. Exposure to Polystyrene Microplastics Promotes the Progression of Cognitive Impairment in Alzheimer's Disease: Association with Induction of Microglial Pyroptosis. *Mol. Neurobiol.* **2023**, *61*, 1–8. <https://doi.org/10.1007/s12035-023-03625-z>.
11. Pallarés-Moratalla, C.; Bergers, G. The Ins and Outs of Microglial Cells in Brain Health and Disease. *Front. Immunol.* **2024**, *15*, 1305087. <https://doi.org/10.3389/fimmu.2024.1305087>.
12. Colonna, M.; Butovsky, O. Microglia Function in the Central Nervous System During Health and Neurodegeneration. *Annu. Rev. Immunol.* **2017**, *35*, 441–468. <https://doi.org/10.1146/annurev-immunol-051116-052358>.
13. Parkhurst, C. N.; Gan, W.-B. Microglia Dynamics and Function in the CNS. *Curr. Opin. Neurobiol.* **2010**, *20* (5), 595–600. <https://doi.org/10.1016/j.conb.2010.07.002>.
14. Orihuela, R.; McPherson, C. A.; Harry, G. J. Microglial M1/M2 Polarization and Metabolic States. *Br. J. Pharmacol.* **2016**, *173* (4), 649–665. <https://doi.org/10.1111/bph.13139>.
15. Wang, J.; He, W.; Zhang, J. A Richer and More Diverse Future for Microglia Phenotypes. *Heliyon* **2023**, *9* (4), e14713. <https://doi.org/10.1016/j.heliyon.2023.e14713>.
16. Paing, Y. M. M.; Eom, Y.; Song, G. B.; Kim, B.; Choi, M. G.; Hong, S.; Lee, S. H. Neurotoxic Effects of Polystyrene Nanoplastics on Memory and Microglial Activation: Insights from *in Vivo* and *in Vitro* Studies. *Sci. Total Environ.* **2024**, *924*, 171681. <https://doi.org/10.1016/j.scitotenv.2024.171681>.
17. Tang, Q.; Li, T.; Chen, K.; Deng, X.; Zhang, Q.; Tang, H.; Shi, Z.; Zhu, T.; Zhu, J. PS-NPs Induced Neurotoxic Effects in SHSY-5Y Cells via Autophagy Activation and Mitochondrial Dysfunction. *Brain Sci.* **2022**, *12* (7), 952. <https://doi.org/10.3390/brainsci12070952>.
18. Sun, J.; Wang, Y.; Du, Y.; Zhang, W.; Liu, Z.; Bai, J.; Cui, G.; Du, Z. Involvement of the JNK/HO-1/FTTH1 Signaling Pathway in Nanoplastic-induced Inflammation and Ferroptosis of BV2 Microglia Cells. *Int. J. Mol. Med.* **2023**, *52* (1), 1–14. <https://doi.org/10.3892/ijmm.2023.5264>.
19. Song, M.; Wei, Z.; Wang, J.; Li, L.; Li, X.; Ma, X.; Pozzolini, M.; Liu, X.; Xiao, L.; Zhong, P. Polystyrene Nanoplastics Induce Oxidative Stress in *Aurelia Coerulea* Polyps, Microglia, and Mice. *Front. Immunol.* **2025**, *16*. <https://doi.org/10.3389/fimmu.2025.1609208>.
20. Rashid, S.; Majeed, L. R.; Mehta, N.; Radu, T.; Martín-Fabiani, I.; Bhat, M. A. Microplastics in Terrestrial Ecosystems: Sources, Transport, Fate, Mitigation, and Remediation Strategies. *Euro-Mediterr. J. Environ. Integr.* **2025**, *10* (4), 2633–2659. <https://doi.org/10.1007/s41207-025-00766-6>.
21. Banaee, M.; Multisanti, C. R.; Impellitteri, F.; Piccione, G.; Faggio, C. Environmental Toxicology of Microplastic Particles on Fish: A Review. *Comp. Biochem. Physiol. Part C Toxicol. Pharmacol.* **2025**, *287*, 110042. <https://doi.org/10.1016/j.cbpc.2024.110042>.
22. Gao, H.-T.; Zhou, H.-Q.; Ji, H.-N.; Shen, J. Z.; Yin, D.-M.; Wu, D.-H. From Exposure to Neurotoxicity Induced by Micro-Nanoplastics with Brain Accumulation and Cognitive Decline. *Ecotoxicol. Environ. Saf.* **2025**, *304*, 119114. <https://doi.org/10.1016/j.ecoenv.2025.119114>.
23. Bu, S.; Xu, Y.; Fu, Y.; Guo, Y.; Wu, D.; Zhang, Y.; Xie, J. Qualitative and Quantitative Analysis of Microplastics and Nanoplastics in Healthy Human Central Nervous System and Its Blood-Brain Barrier Transmission Efficiency-A Pilot Study. *Environ. Pollut.* **2026**, *388*, 127392. <https://doi.org/10.1016/j.envpol.2025.127392>.
24. Amato-Lourenço, L. F.; Dantas, K. C.; Júnior, G. R.; Paes, V. R.; Ando, R. A.; de Oliveira Freitas, R.; da Costa, O.; Rabelo, R. S.; Soares Bispo, K. C.; Carvalho-Oliveira, R. *Microplastics in the Olfactory Bulb of the Human Brain. JAMA Netw Open* **7** (9): E2440018; **2024**.
25. Prosperi, G.; Marchetti, N.; D'Elia, A.; Massari, R.; Giusto, M.; Pietrodangelo, A.; Rossi, T.; Nucara, A.; Scavizzi, F.; Strimpakos, G.; Marinelli, S.; Mandillo, S.; D'Amato, F. R.; Farioli-Vecchioli, S. Inhalation of Nanoplastics in the Mouse Model: Tissue Bio-Distribution and Effects on the Olfactory System. *Sci. Total Environ.* **2025**, *968*, 178853. <https://doi.org/10.1016/j.scitotenv.2025.178853>.
26. Kalmey, J. K.; Thewissen, J. G. M.; Dluzen, D. E. Age-Related Size Reduction of Foramina in the Cribriform Plate. *Anat. Rec.* **1998**, *251* (3), 326–329. [https://doi.org/10.1002/\(SICI\)1097-0185\(199807\)251:3%3C326::AID-AR7%3E3.0.CO;2-T](https://doi.org/10.1002/(SICI)1097-0185(199807)251:3%3C326::AID-AR7%3E3.0.CO;2-T).
27. Dorrego-Rivas, A.; Grubb, M. S. Developing and Maintaining a Nose-to-Brain Map of Odorant Identity. *Open Biol.* **2022**, *12* (6).
28. Gandhi, S.; Shastri, D. H.; Shah, J.; Nair, A. B.; Jacob, S. Nasal Delivery to the Brain: Harnessing Nanoparticles for Effective Drug Transport. *Pharmaceutics* **2024**, *16* (4), 481. <https://doi.org/10.3390/pharmaceutics16040481>.
29. McConnell, H. L.; Mishra, A. Cells of the Blood-Brain Barrier: An Overview of the Neurovascular Unit in Health and Disease. *Blood-Brain Barrier Methods Protocol.* **2022**, 3–24.
30. Rust, R.; Yin, H.; Achón Buil, B.; Sagare, A. P.; Kisler, K. The Blood-Brain Barrier: A Help and a Hindrance. *Brain* **2025**, *148* (7), 2262–2282.
31. Zhang, H.-J.; Li, S.; Wang, X.-L.; Zhang, K.-D.; Fang, H.-T.; Wu, X.; Huang, Z.; Jiang, W.; Yang, L.; Tan, Q.-G.; Pan, B.; Ji, R.; Wang, P.; Xing, B.; Miao, A.-J. Size-Dependent Translocation of Polystyrene Nanoplastics across Biological Barriers in Mammals. *Nat. Commun.* **2025**. <https://doi.org/10.1038/s41467-025-67876-1>.
32. Kopatz, V.; Wen, K.; Kovács, T.; Keimowitz, A. S.; Pichler, V.; Widder, J.; Vethaak, A. D.; Hollóczki, O.; Kenner, L. Micro- and Nanoplastics Breach the Blood-Brain Barrier (BBB): Biomolecular Corona's Role Revealed. *Nanomaterials* **2023**, *13* (8), 1404. <https://doi.org/10.3390/nano13081404>.
33. Abdolapur Monikh, F.; Lehtonen, Š.; Kekäläinen, J.; Karkossa, I.; Auriola, S.; Schubert, K.; Zanut, A.; Peltonen, S.; Niskanen, J.; Bandekar, M.; von Bergen, M.; Leskinen, J. T. T.; Koistinen, A.; Bogialli, S.; Guo, Z.; Kukkonen, J. V. K.; Chen, C.; Lynch, I. Biotransformation of Nanoplastics in Human Plasma and Their Permeation through a Model *in Vitro* Blood-Brain Barrier: An in-Depth Quantitative Analysis. *Nano Today* **2024**, *59*, 102466. <https://doi.org/10.1016/j.nantod.2024.102466>.
34. Yang, H.; Chen, Z.; Kong, L.; Xing, H.; Yang, Q.; Wu, J. A Review of Eco-Corona Formation on Micro/Nanoplastics and Its Effects on Stability, Bioavailability, and Toxicity. *Water* **2025**, *17* (8), 1124. <https://doi.org/10.3390/w17081124>.
35. Li, X.; He, E.; Jiang, K.; Peijnenburg, W. J. G. M.; Qiu, H. The Crucial Role of a Protein Corona in Determining the Aggregation Kinetics and Colloidal Stability of Polystyrene Nanoplastics. *Water Res.* **2021**, *190*, 116742. <https://doi.org/10.1016/j.watres.2020.116742>.
36. Baroni, A.; Moulton, C.; Cristina, M.; Sansone, L.; Belli, M.; Tasciotti, E. Nano- and Microplastics in the Brain: An Emerging Threat to Neural Health. *Nanomaterials* **2025**, *15* (17), 1361. <https://doi.org/10.3390/nano15171361>.
37. Kwon, W.; Kim, D.; Kim, H.-Y.; Jeong, S. W.; Lee, S.-G.; Kim, H.-C.; Lee, Y.-J.; Kwon, M. K.; Hwang, J.-S.; Han, J. E.; Park, J.-

- K.; Lee, S.-J.; Choi, S.-K. Microglial Phagocytosis of Polystyrene Microplastics Results in Immune Alteration and Apoptosis *in Vitro* and *in Vivo*. *Sci. Total Environ.* **2022**, *807*, 150817. <https://doi.org/10.1016/j.scitotenv.2021.150817>.
38. Gordon, S. Phagocytosis: An Immunobiologic Process. *Immunity* **2016**, *44* (3), 463–475. <https://doi.org/10.1016/j.immuni.2016.02.026>.
 39. Fanghella, F.; Pesce, M.; Franceschelli, S.; Panella, V.; Elsallabi, O.; Lupi, T.; Rizza, B.; Di Battista, M. G.; Bruno, A.; Ballerini, P.; Patruno, A.; Speranza, L. Biological Modulation of Autophagy by Nanoplastics: A Current Overview. *Int. J. Mol. Sci.* **2025**, *26* (15), 7035. <https://doi.org/10.3390/ijms26157035>.
 40. Hipolito, V. E. B.; Ospina-Escobar, E.; Botelho, R. J. Lysosome Remodeling and Adaptation during Phagocyte Activation. *Cell. Microbiol.* **2018**, *20* (4), e12824. <https://doi.org/10.1111/cmi.12824>.
 41. da Silva Antunes, J. C. Toxicological Mechanisms of Nanoplastics: Inflammation as a Genotoxicity Trigger, Universidade NOVA de Lisboa (Portugal), **2024**.
 42. Yu, H.; Zhao, Z.; Li, H.; Han, Y.; Li, H.; Cui, C.; Hu, Y.; Zhang, B. Nanoplastics Exposure Exacerbates A β Plaque Deposition in Alzheimer's Disease Mice by Inducing Microglia Pyroptosis. *Ecotoxicol. Environ. Saf.* **2025**, *299*, 118379. <https://doi.org/10.1016/j.ecoenv.2025.118379>.
 43. Yamashima, T. Lysosomal Membrane-Permeabilization (LMP) and -Rupture (LMR) Are Distinct for Cell Death. *Front. Cell Death* **2025**, *4*. <https://doi.org/10.3389/fceld.2025.1669955>.
 44. Zhang, J.-R.; Shen, S.-Y.; Zhai, M.-Y.; Shen, Z.-Q.; Li, W.; Liang, L.-F.; Yin, S.-Y.; Han, Q.-Q.; Li, B.; Zhang, Y.-Q.; Yu, J. Augmented Microglial Endoplasmic Reticulum-Mitochondria Contacts Mediate Depression-like Behavior in Mice Induced by Chronic Social Defeat Stress. *Nat. Commun.* **2024**, *15* (1), 5199. <https://doi.org/10.1038/s41467-024-49597-z>.
 45. Lee, S. E.; Yi, Y.; Moon, S.; Yoon, H.; Park, Y. S. Impact of Micro- and Nanoplastics on Mitochondria. *Metabolites* **2022**, *12* (10), 897. <https://doi.org/10.3390/metabo12100897>.
 46. Florance, I.; Chandrasekaran, N.; Gopinath, P. M.; Mukherjee, A. Exposure to Polystyrene Nanoplastics Impairs Lipid Metabolism in Human and Murine Macrophages *in Vitro*. *Ecotoxicol. Environ. Saf.* **2022**, *238*, 113612. <https://doi.org/10.1016/j.ecoenv.2022.113612>.
 47. Koner, S.; Florance, I.; Mukherjee, A.; Chandrasekaran, N. Cellular Response of THP-1 Macrophages to Polystyrene Microplastics Exposure. *Toxicology* **2023**, *483*, 153385. <https://doi.org/10.1016/j.tox.2022.153385>.
 48. Merkley, S. D.; Moss, H. C.; Goodfellow, S. M.; Ling, C. L.; Meyer-Hagen, J. L.; Weaver, J.; Campen, M. J.; Castillo, E. F. Polystyrene Microplastics Induce an Immunometabolic Active State in Macrophages. *Cell Biol. Toxicol.* **2022**, *38* (1), 31–41. <https://doi.org/10.1007/s10565-021-09616-x>.
 49. Bohlen, C. J.; Bennett, F. C.; Tucker, A. F.; Collins, H. Y.; Mulin-yawe, S. B.; Barres, B. A. Diverse Requirements for Microglial Survival, Specification, and Function Revealed by Defined-Medium Cultures. *Neuron* **2017**, *94* (4), 759–773.e8. <https://doi.org/10.1016/j.neuron.2017.04.043>.
 50. Shen, K.; Pender, C. L.; Bar-Ziv, R.; Zhang, H.; Wickham, K.; Willey, E.; Durieux, J.; Ahmad, Q.; Dillin, A. Mitochondria as Cellular and Organismal Signaling Hubs. *Annu. Rev. Cell Dev. Biol.* **2022**, *38* (Volume 38, 2022), 179–218. <https://doi.org/10.1146/annurev-cellbio-120420-015303>.
 51. Yang, S.; Qin, C.; Hu, Z.-W.; Zhou, L.-Q.; Yu, H.-H.; Chen, M.; Bosco, D. B.; Wang, W.; Wu, L.-J.; Tian, D.-S. Microglia Reprogram Metabolic Profiles for Phenotype and Function Changes in the Central Nervous System. *Neurobiol. Dis.* **2021**, *152*, 105290. <https://doi.org/10.1016/j.nbd.2021.105290>.
 52. Baik, S. H.; Kang, S.; Lee, W.; Choi, H.; Chung, S.; Kim, J.-I.; Mook-Jung, I. A Breakdown in Metabolic Reprogramming Causes Microglia Dysfunction in Alzheimer's Disease. *Cell Metab.* **2019**, *30* (3), 493–507.e6. <https://doi.org/10.1016/j.cmet.2019.06.005>.
 53. Peruzzotti-Jametti, L.; Willis, C. M.; Krzak, G.; Hamel, R.; Pirvan, L.; Ionescu, R.-B.; Reisz, J. A.; Prag, H. A.; Garcia-Segura, M. E.; Wu, V.; Xiang, Y.; Barlas, B.; Casey, A. M.; van den Bosch, A. M. R.; Nicaise, A. M.; Roth, L.; Bates, G. R.; Huang, H.; Prasad, P.; Vincent, A. E.; Frezza, C.; Viscomi, C.; Balmus, G.; Takats, Z.; Marioni, J. C.; D'Alessandro, A.; Murphy, M. P.; Mohorianu, I.; Pluchino, S. Mitochondrial Complex I Activity in Microglia Sustains Neuroinflammation. *Nature* **2024**, *628* (8006), 195–203. <https://doi.org/10.1038/s41586-024-07167-9>.
 54. Rosa, C. P.; Belo, T. C. A.; Santos, N. C. de M.; Silva, E. N.; Gasparotto, J.; Corsetti, P. P.; de Almeida, L. A. Reactive Oxygen Species Trigger Inflammasome Activation after Intracellular Microbial Interaction. *Life Sci.* **2023**, *331*, 122076. <https://doi.org/10.1016/j.lfs.2023.122076>.
 55. Kelley, N.; Jeltema, D.; Duan, Y.; He, Y. The NLRP3 Inflammasome: An Overview of Mechanisms of Activation and Regulation. *Int. J. Mol. Sci.* **2019**, *20* (13), 3328. <https://doi.org/10.3390/ijms20133328>.
 56. Hanslik, K. L.; Ulland, T. K. The Role of Microglia and the Nlrp3 Inflammasome in Alzheimer's Disease. *Front. Neurol.* **2020**, *11*. <https://doi.org/10.3389/fneur.2020.570711>.
 57. Campden, R. I.; Zhang, Y. The Role of Lysosomal Cysteine Cathepsins in NLRP3 Inflammasome Activation. *Arch. Biochem. Biophys.* **2019**, *670*, 32–42. <https://doi.org/10.1016/j.abb.2019.02.015>.
 58. Cheviaux, A.; Pilot, T.; Derangère, V.; Simonin, H.; Martine, P.; Chalmin, F.; Ghiringhelli, F.; Rébé, C. Cathepsin B Is Required for NLRP3 Inflammasome Activation in Macrophages, Through NLRP3 Interaction. *Front. Cell Dev. Biol.* **2020**, *8*, 167. <https://doi.org/10.3389/fcell.2020.00167>.
 59. Ising, C.; Venegas, C.; Zhang, S.; Scheiblich, H.; Schmidt, S. V.; Vieira-Saecker, A.; Schwartz, S.; Albasset, S.; McManus, R. M.; Tejera, D.; Griep, A.; Santarelli, F.; Brosseron, F.; Opitz, S.; Stunnen, J.; Merten, M.; Kaye, R.; Golenbock, D. T.; Blum, D.; Latz, E.; Buée, L.; Heneka, M. T. NLRP3 Inflammasome Activation Drives Tau Pathology. *Nature* **2019**, *575* (7784), 669–673. <https://doi.org/10.1038/s41586-019-1769-z>.
 60. Qin, J.; Ma, Z.; Chen, X.; Shu, S. Microglia Activation in Central Nervous System Disorders: A Review of Recent Mechanistic Investigations and Development Efforts. *Front. Neurol.* **2023**, *14*. <https://doi.org/10.3389/fneur.2023.1103416>.
 61. Antunes, J.; Sobral, P.; Martins, M.; Branco, V. Nanoplastics Activate a TLR4/P38-Mediated pro-Inflammatory Response in Human Intestinal and Mouse Microglia Cells. *Environ. Toxicol. Pharmacol.* **2023**, *104*, 104298. <https://doi.org/10.1016/j.etap.2023.104298>.
 62. Li, G.; Liu, X.; Sun, X.; Huang, L.; Kuang, W.; Ou, J.; Zhang, J.; Zhang, Z.; Li, H.; Tang, H.; Feng, C.; Gu, L.; Yang, C.; Peili, W.; Wang, J. Polystyrene Microplastics Induce Anxiety via HRAS Derived PERK-NF- κ B Pathway. *Environ. Int.* **2024**, *185*, 108543. <https://doi.org/10.1016/j.envint.2024.108543>.
 63. Tam, A. B.; Mercado, E. L.; Hoffmann, A.; Niwa, M. ER Stress Activates NF- κ B by Integrating Functions of Basal IKK Activity, IRE1 and PERK. *PLoS ONE* **2012**, *7* (10), e45078. <https://doi.org/10.1371/journal.pone.0045078>.
 64. Guo, Q.; Jin, Y.; Chen, X.; Ye, X.; Shen, X.; Lin, M.; Zeng, C.; Zhou, T.; Zhang, J. NF- κ B in Biology and Targeted Therapy: New Insights and Translational Implications. *Signal Transduct. Target. Ther.* **2024**, *9* (1), 53. <https://doi.org/10.1038/s41392-024-01757-9>.

65. Lu, X.; Luo, Q.; Zhao, J.; Li, M.; Liu, D. Revealing the Underlying Mechanisms of Nanoplastics Induces Neuroinflammation: From Transcriptomic Analysis to *in Vivo* and *in Vitro* Validation. *Eco-toxicol. Environ. Saf.* **2025**, *298*, 118311. <https://doi.org/10.1016/j.ecoenv.2025.118311>.
66. *The noncanonical NF- κ B pathway - Sun - 2012 - Immunological Reviews - Wiley Online Library*. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-065X.2011.01088.x> (accessed 2026-01-29).
67. Ru, Q.; Li, Y.; Chen, L.; Wu, Y.; Min, J.; Wang, F. Iron Homeostasis and Ferroptosis in Human Diseases: Mechanisms and Therapeutic Prospects. *Signal Transduct. Target. Ther.* **2024**, *9* (1), 271. <https://doi.org/10.1038/s41392-024-01969-z>.
68. Hollóczy, O.; Gehrke, S. Nanoplastics Can Change the Secondary Structure of Proteins. *Sci. Rep.* **2019**, *9* (1), 16013. <https://doi.org/10.1038/s41598-019-52495-w>.
69. Gabbrielli, S.; Colnaghi, L.; Mazzuoli-Weber, G.; Redaelli, A. C. L.; Gautieri, A. In Silico Analysis of Nanoplastics' and β -Amyloid Fibrils' Interactions. *Molecules* **2023**, *28* (1), 388. <https://doi.org/10.3390/molecules28010388>.
70. Windheim, J.; Colombo, L.; Battajni, N. C.; Russo, L.; Cagnotto, A.; Diomedea, L.; Bigini, P.; Vismara, E.; Fiumara, F.; Gabbrielli, S.; Gautieri, A.; Mazzuoli-Weber, G.; Salmona, M.; Colnaghi, L. Micro- and Nanoplastics' Effects on Protein Folding and Amyloidosis. *Int. J. Mol. Sci.* **2022**, *23* (18), 10329. <https://doi.org/10.3390/ijms231810329>.
71. Wang, Y.; Huang, G.; Liang, X.; Andrikopoulos, N.; Tang, H.; Ding, F.; Ke, P. C.; Li, Y. Microglial Clearance of Alzheimer's Amyloid-Beta Obstructed by Nanoplastics. *Environ. Sci. Nano* **2025**, *12* (6), 3247–3260.
72. Rinauro, D. J.; Chiti, F.; Vendruscolo, M.; Limbicker, R. Misfolded Protein Oligomers: Mechanisms of Formation, Cytotoxic Effects, and Pharmacological Approaches against Protein Misfolding Diseases. *Mol. Neurodegener.* **2024**, *19* (1), 20. <https://doi.org/10.1186/s13024-023-00651-2>.
73. Jacquet, R. G.; González Ibáñez, F.; Picard, K.; Funes, L.; Khakpour, M.; Gouras, G. K.; Tremblay, M.-È.; Maxfield, F. R.; Solé-Domènech, S. Microglia Degrade Alzheimer's Amyloid-Beta Deposits Extracellularly via Digestive Exophagy. *Cell Rep.* **2024**, *43* (12), 115052. <https://doi.org/10.1016/j.celrep.2024.115052>.
74. Majumdar, A.; Cruz, D.; Asamoah, N.; Buxbaum, A.; Sohar, I.; Lobel, P.; Maxfield, F. R. Activation of Microglia Acidifies Lysosomes and Leads to Degradation of Alzheimer Amyloid Fibrils. *Mol. Biol. Cell* **2007**, *18* (4), 1490–1496. <https://doi.org/10.1091/mbc.E06-10-0975>.
75. Solé-Domènech, S.; Rojas, A. V.; Maisuradze, G. G.; Scheraga, H. A.; Lobel, P.; Maxfield, F. R. Lysosomal Enzyme Tripeptidyl Peptidase 1 Destabilizes Fibrillar A β by Multiple Endoproteolytic Cleavages within the β -Sheet Domain. *Proc. Natl. Acad. Sci.* **2018**, *115* (7), 1493–1498. <https://doi.org/10.1073/pnas.1719808115>.
76. Gao, C.; Jiang, J.; Tan, Y.; Chen, S. Microglia in Neurodegenerative Diseases: Mechanism and Potential Therapeutic Targets. *Signal Transduct. Target. Ther.* **2023**, *8* (1), 359. <https://doi.org/10.1038/s41392-023-01588-0>.
77. Pradel, A.; Catrouillet, C.; Gigault, J. The Environmental Fate of Nanoplastics: What We Know and What We Need to Know about Aggregation. *NanoImpact* **2023**, *29*, 100453. <https://doi.org/10.1016/j.impact.2023.100453>.
78. Pavlovic, D.; Papic, D.; Janjic, V.; Mitrovic, M.; Dimitrijevic Stojanovic, M.; Gazdic Jankovic, M. Nuclear and Mitochondrial Epigenetic Mechanisms Underlying Neurodegeneration and Gut-Brain Axis Dysregulation Induced by Micro- and Nanoplastics. *Genes* **2026**, *17* (2), 151. <https://doi.org/10.3390/genes17020151>.
79. Wang, J.; Yang, Y.; Shi, Y.; Wei, L.; Gao, L.; Liu, M. Oxidized/Unmodified-Polyethylene Microplastics Neurotoxicity in Mice: Perspective from Microbiota-Gut-Brain Axis. *Environ. Int.* **2024**, *185*, 108523. <https://doi.org/10.1016/j.envint.2024.108523>.
80. Zhang, Z.; Xu, M.; Wang, L.; Gu, W.; Li, X.; Han, Z.; Fu, X.; Wang, X.; Li, X.; Su, Z. Continuous Oral Exposure to Micro- and Nanoplastics Induced Gut Microbiota Dysbiosis, Intestinal Barrier and Immune Dysfunction in Adult Mice. *Environ. Int.* **2023**, *182*, 108353. <https://doi.org/10.1016/j.envint.2023.108353>.
81. Yang, Q.; Dai, H.; Cheng, Y.; Wang, B.; Xu, J.; Zhang, Y.; Chen, Y.; Xu, F.; Ma, Q.; Lin, F.; Wang, C. Oral Feeding of Nanoplastics Affects Brain Function of Mice by Inducing Macrophage IL-1 Signal in the Intestine. *Cell Rep.* **2023**, *42* (4). <https://doi.org/10.1016/j.celrep.2023.112346>.
82. Cheng, Y.; Chen, J.; Fu, R.; Zhang, P.; Chen, H.; Cao, H.; Jiang, Z.; Hong, Y.; Li, Y.; He, C.; Tao, F.; Li, T.; Zhang, J.; Chen, B.; Cao, G. Molecular Mechanism Differences between Nanoplastics and Microplastics in Colon Toxicity: Nanoplastics Induce Ferroptosis-Mediated Immunogenic Cell Death, While Microplastics Cause Cell Metabolic Reprogramming. *J. Nanobiotechnology* **2025**, *23* (1), 505. <https://doi.org/10.1186/s12951-025-03545-1>.
83. Fan, J.; Liu, L.; Lu, Y.; Chen, Q.; Fan, S.; Yang, Y.; Long, Y.; Liu, X. Acute Exposure to Polystyrene Nanoparticles Promotes Liver Injury by Inducing Mitochondrial ROS-Dependent Necroptosis and Augmenting Macrophage-Hepatocyte Crosstalk. *Part. Fibre Toxicol.* **2024**, *21*, 20. <https://doi.org/10.1186/s12989-024-00578-6>.
84. Fan, X.; Xu, T.; Peng, L.; Tang, X.; Chi, Q.; Li, Z.; Shi, G.; Li, S. Polystyrene Nanoplastic Exposure Enhances LPS-Induced Lung Inflammation in Mice by Inducing M1 Polarization of Macrophages via ROS/JAK/STAT Signaling Pathway. **2022**.
85. Xu, M.; Qin, Z.; Chen, Z.; Wang, S.; Peng, L.; Li, X.; Yuan, Z. Nanorobots Mediated Drug Delivery for Brain Cancer Active Targeting and Controllable Therapeutics. *Discov. Nano* **2024**, *19* (1), 183. <https://doi.org/10.1186/s11671-024-04131-4>.
86. Zhou, H.; Yang, Y.; Li, P.; Li, J. Micro/Nanorobots for the Removal of Microplastics and Nanoplastics from the Human Body. *Med Mat* **2025**, *2* (4), 183. <https://doi.org/10.1097/mm9.0000000000000023>.

■ Author

Ella Wilce is an aspiring scientist with a passion for genetics, neuroscience, and environmental sustainability. She hopes to pursue a related field at a top university with a focus on understanding the ecological effects of environmental pollutants.

Minimalism in Korean Aesthetics as a Historically Formed Ideology

Hiisan Kim

British International School Phuket, 59 Moo 2, Thepkasattri Road, T. Koh Kaew, A. Muang Mueang Phuket District, Phuket 83000;
hiisankim@gmail.com

ABSTRACT: This review paper explores how minimalism characterizes Korean aesthetics from the past to the present. This paper, which draws on scholarship in cultural philosophy, history, visual art, architecture, and economics, argues that Korean minimalism is more than simply a contemporary style, but rather an aesthetic logic with historical depth formed through philosophical tradition and material circumstance. The first chapter examines how Neo-Confucianism influenced the Joseon dynasty to an extent that ideals of restraint, moderation, naturalness (jayeon), and the aesthetic of space (yeobaek) took visual form in white porcelain ceramics and buildings. The paper then elaborates on how Japanese colonization, economic exploitation, war, and post-war scarcity transformed the environment, arguing that the hardships of the 20th century intensified and diverted pre-existing tendencies towards simplicity. The emergence of Dansaekhwa in the late 1960s and 1970s, as examined by the essay, was a crucial moment in which inherited aesthetic restraint and historical constraint became a conscious artistic movement. By situating Korean minimalism in this historical and cultural context, the paper can also position it as a product of continuity and adaptation shaped not just by philosophical values but also by political and economic upheavals.

KEYWORDS: Behavioral and Social Sciences, Sociology and Social Psychology, Cultural Philosophy, Aesthetic, Korean, Korean Art History, Minimalism, Confucianism, Dansaekhwa, Joseon Dynasty, Material Scarcity, Jayeon.

■ Introduction

In the 20th century, minimalism became a defining feature of modern Korean aesthetics, the style and aesthetic choice forming a consistent and visible presence within everyday visual surroundings in the country. This paper strives to explain the different cultural, philosophical, and social aspects that contributed to the rise of this aesthetic within the nation.

Importantly, simplicity in Korean art was recognized long before the emergence of modern minimalist movements. In one of the earliest comprehensive surveys of Korean art, Eckardt identified its distinctive qualities as including “simple, modest and unobtrusive interpretation of forms” alongside “repose and moderation,” suggesting that restraint and clarity were already considered central characteristics of Korean artistic production.¹ Such early observations indicate that what is now termed “minimalism” may be rooted in longer aesthetic traditions rather than being purely a modern development.

Existing scholarship has approached Korean minimalism from several angles. Art historical studies of Dansaekhwa focus on monochrome painting practices emerging in the late 1960s, emphasizing repetition and the embedding of time and labor in the artistic process.² Other research situates minimalist tendencies earlier, identifying simplicity and restraint as key characteristics of Korean art and architecture from the Joseon dynasty onward, particularly in concepts such as yeobaek (the beauty of emptiness), hanok spatial organization, and the aesthetic of white associated with baegui minjok, all stemming from Neo-Confucianist ideology.³

This paper addresses this gap by examining Korean minimalism as a historically formed aesthetic logic that developed

through a cumulative process. It traces the development of minimalist aesthetics from Confucian-influenced traditions of restraint in the Joseon period, through their transformation under colonialism, war, and authoritarian modernization, to their rearticulation in Dansaekhwa art. By situating modern Korean minimalism within this broader chronological and cultural framework, the study contributes a more cohesive understanding of why minimalism remains a powerful aesthetic orientation and preference in Korea today.

Rather than examining these ideas separately, this paper synthesizes these various perspectives to propose a more unified understanding of Korean minimalism as a historically formed aesthetic logic that developed through a cumulative process. It traces the initial development of minimalist ideology from Confucian-influenced traditions of restraint in the Joseon period, through the onset of a materially enforced minimalism dictated by colonialism, war, and authoritarianism, to its rearticulation in modern aesthetics such as Dansaekhwa art. By situating modern Korean minimalism within this broader chronological and cultural framework, the study contributes a more cohesive understanding of how minimalism became embedded in Korean culture and why it continues to remain as a powerful aesthetic orientation and preference in Korea today.

■ Philosophical and pre-modern foundations

Confucius Values and Aesthetic Restraint:

To begin our examination of minimalistic values and tendencies in Korea, this paper first dives into the origins of the

philosophy that builds upon the foundations of lifestyle and moral values that have allowed minimalism to thrive in the nation. Confucianism first emerged approximately 2500 years ago in China, and according to Mitu, reached the Korean peninsula before the beginning of the Three Kingdoms period (57 BCE–668CE).⁴ Over time, and particularly with the rise of Neo-Confucianism, Confucian thought became institutionalized as the dominant ideological framework during the Joseon dynasty (1392–1910).⁵ This extensive historical longevity allowed Confucian moral principles to diffuse deeply into Korean society, profoundly shaping the way that Koreans think and behave.

Central to Confucian thought are ideals such as restraint, moderation, and harmony within the social classes, which all prioritize inner virtue and modesty over external display. Hence, this paper argues that simplicity (and in turn, minimalist styles) were considered by Koreans who followed this ideology as more ethically appropriate in different aesthetic contexts because they reflected these ideals of self-discipline and suppression of indulgence. This ideology set the pre-modern foundations for the birth of early aesthetic principles.

Neo-Confucianist values are evident across numerous historical records of Korean material culture, most notably during the Joseon dynasty, when Neo-Confucianism arguably reached its peak influence within the context of Korea.⁵ Practices such as the widespread wearing of white clothing and artistic emphasis on finding beauty in blank or unoccupied spaces reflect an aesthetic preference grounded in moral restraint rather than visual excess. Moreover, these ideals were materialized in artifacts that transformed Neo-Confucian moral discipline into a coherent aesthetic language of minimalism, such as in traditional ceramics.

The White-clad People:

One of the most widely cited cultural expressions of this aesthetic orientation is the phrase *baegui minjok* (백의민족), commonly translated amongst foreign scholarship as the "white-clad people." If we take the literal translations of the terms in this phrase separately, the term *baegui* means white, and *minjok* translates to "the people." This phrasing leaves a degree of ambiguity regarding the precise object to which the description of "white" is applied, which is significant because it suggests that the "whiteness" functioned as some kind of broader aesthetic ideal that applied to more than one aspect of aesthetics. However, for this paper, the focus is placed specifically on dress and clothing, as historical discourse most consistently associates *baegui minjok* with the widespread practice of wearing white clothing by the people of Joseon.

Seo explains that Confucianism places strong emphasis on moral discipline, encouraging the suppression of personal emotions and desires that could disrupt social harmony.³ Within this ethical framework, color is treated not merely as an aesthetic choice but as a choice that is associated with desire and emotional excess, which must be regulated. As Seo notes, color functions as a "symbol of desires that need to be suppressed," leading to the reverence of white and the discouragement of other vivid hues under what she terms the

gum-chae-saek or "no-color philosophy."^{3,6} The popularity of wearing white clothing amongst the Joseon people should therefore be understood as a lived expression of Neo-Confucian ethical discipline rather than a purely aesthetic preference. White garments symbolized purity, humility, and modesty, aligning with Confucian ideals that we have previously discussed. This was the opposite case for brightly dyed fabrics, which signaled excess or indulgence. Additionally, Koreans historically have valued harmony and fitting in amongst society, as this idea forms an important part of Neo-Confucian ideology. As wearing white, a neutral color, allowed people to blend in and not stand out amongst a larger crowd, it further reinforced its appeal (Figure 1).

Importantly, we must also note that this aesthetic restraint was not effortless: white clothing required constant laundering and maintenance due to the fact that it was particularly susceptible to visual staining. Consequently, this requirement institutionalized a strict visual discipline within daily routines. The act of wearing white garments thus signaled an individual's capacity to take good care of their belongings, keeping them clean and pure. As such, simplicity in dress functioned as an active moral practice, sustained through habitual effort rather than passive absence of decoration. This historical alignment offers an interpretive framework for understanding contemporary visual patterns in modern Korea, such as the monochrome, minimalist fashion trends colloquially described as the "Seoul uniform."

External observers also recognized the distinctive significance of whiteness in Joseon aesthetics. Japanese scholar Yanagi Muneyoshi, while admiring Joseon craftsmanship, interpreted the pervasive use of white as a reflection of collective suffering, describing it as akin to a "permanent mourning period" rooted in funeral traditions. His observations highlight how unusual Joseon aesthetic restraint appeared when contrasted with the more color-saturated visual cultures of China and Japan. However, such interpretations risk reducing whiteness to a single emotional register. As later scholarship has noted, whiteness in Joseon Korea functioned across multiple religious and philosophical contexts. Seo's analysis emphasizes that white attire appeared in Buddhist ritual dance, Taoist clerical dress, and Confucian ceremonial practice, suggesting that its significance extended beyond mourning to encompass purity and discipline.³



Figure 1: A market day in a countryside town (1906–1907). National Folk Museum of Korea.⁷ The widespread adoption of white garments across ordinary classes reflects the prioritization of collective societal harmony, visually demonstrating how the uniform use of this color allowed individuals to blend into the community rather than standing out.

The Beauty in the Empty:

Another prominent aesthetic principle within Joseon visual culture was the concept of *yeobaek* (여백). Literally translating to “remaining blank space,” *yeobaek* refers to the intentional preservation of unmarked or empty areas within a composition, such as in a piece of poetry or artwork. The idea of *yeobaek* is such that absence functions as an active aesthetic element rather than a lack or an empty vacuum, finding meaning in the void. Koreans often use this word in the phrase “*yeobaek eu mi*,” meaning the beauty of blank space.

This aesthetic ideal was perhaps epitomized and shown in its purest form by ceramics in the Joseon dynasty, the most prominent example being white porcelain “moon jars.” Moon jars were traditional white porcelain jars that were purely white-colored without additional decorations and well-rounded in what was described as a “feminine shape.”⁸ The expansive white surfaces and rounded proportions of these ceramics foregrounded emptiness and restraint, and they also can be understood as embodying Neo-Confucianist values through their quiet, modest form and lack of ornamentation.

The significance of the color white reinforces this connection in a similar fashion to the ideas discussed in the previous section about the culture of “*baegui minjok*,” as the color itself can be seen as a direct manifestation of Confucianist ideals. Shin, when referencing Joseon ceramics, states that “The beauty of white porcelain lies in its great chasteness and beauty, which symbolized the dynasty’s taste, the whiteness representing the nation, simplicity and a conscious cult of asceticism,” showing the significance of the color choice in aesthetically attractive artifacts.⁹ Additionally, Shin goes on to say “at first, the white color of the white porcelain was a clear and glowing color, but later transformed into milky white, and after 1780 it turned into a bluish white with a cool blue color, which symbolizes the pure ideal of the Joseon Dynasty.”⁹ This observation is significant because it demonstrates that although the surface tone of Joseon porcelain evolved, its fundamental commitment to “whiteness” remained consistent.

However, it can also be argued that *yeobaek* stems from a broader aesthetic aspiration toward naturalness, or the concept of *jayeon* in Korean beauty. The word *jayeon* in the Korean language simply translates to “nature” in English, but is a concept often referenced by foreign and domestic scholars alike when discussing Korean aesthetic ideals.¹ Art critic Muneyoshi Yanagi argued that Korean art, particularly Joseon ceramics, represented a “return to nature.”¹¹ Importantly, his understanding of “nature” did not refer to the literal depiction of landscapes such as trees, mountains, or wildlife. Instead, it referred to a state that feels “natural” because nothing is artificially forced. When a work of art reaches this kind of state, it no longer feels overly constructed or controlled, but instead appears inevitable, as though it exists in its most appropriate, “natural” form.

If *jayeon* is understood as a deliberate avoidance of artificiality, then the principle of *yeobaek* can be seen as an extremely practical visual strategy for achieving this condition in Korean art. Often, we can see that objects that possess “*yeobaek eu mi*” (beauty in blank space) also embody a sort of natural beauty.

Let’s take the same example of Joseon ceramics. Moon jars embrace imperfections as part of their beauty, as these aspects demonstrate them to be pure. The shape of the jar is not mathematically circular nor symmetrical, as the two halves of the porcelain are joined together, as evident in the slight dent in the otherwise smooth curve of the pot. The current moon jar displayed in the Metropolitan Museum of Art possesses hues of pink on its exterior (Figure 2), which would have happened unintentionally during the firing process.¹⁰

Taken together, the ideas of *yeobaek* and the importance of the color white show that Joseon aesthetics were built on Confucianist philosophy, but also the concept of *jayeon*. In moon jars, emptiness is not meaningless, and even the small imperfections in the jars are valued, as they make the object feel natural rather than artificially perfected. These features show that beauty in Joseon Korea was often found through removal and a “lack” rather than addition.

Because of this, minimalist tendencies in modern Korean art and design can be understood as part of a longer tradition rather than a completely new development. The preference for simplicity, blank space, and natural form did not suddenly appear in the twentieth century. Instead, these qualities were already deeply rooted in Joseon visual culture and shaped by Neo-Confucian values and the idea of *jayeon*. Modern Korean minimalism can therefore be seen as a continuation and reinterpretation of these earlier aesthetic principles, rather than a complete break from them.



Figure 2: “Moon Jar,” Joseon Dynasty (1392–1910). Porcelain. The Metropolitan Museum of Art.¹¹ The subtle, accidental pink hues on the exterior highlight the unpredictable nature of traditional kiln firings. This imperfection embodies the Neo-Confucian aesthetic value of unforced, natural simplicity rather than technical perfection.

■ Minimalism Under Constraint: Poverty, War, and Scarcity

While pre-modern Korean aesthetics reveal a voluntary and philosophically grounded commitment to simple aesthetics, the twentieth century introduced a different set of pressures that reshaped minimalist tendencies for a wholly different reason. The Japanese colonial period (1910–1945), followed by the devastation of the Korean War (1950–1953), created widespread material scarcity and severe economic hardship within a nation that had once been prosperous, depriving it of the agency to make decisions of its own. The policies imposed on the common people by colonial powers also limited freedom of expression and opportunities for a lifestyle of extravagance,

while urbanization created spatial constraints that forced the simplification of architecture. There was no longer an agency to make a choice on aesthetics for the people, as aesthetics were the last of their priorities. The option of choice was no more. In this context, simplicity became a choice that prioritized practicality, rather than one that stemmed from aesthetic desire.

Japanese colonization fundamentally restructured the Korean economy in ways that intensified material scarcity for the Korean population, despite visible signs of industrial growth. As King argues, the colonial period produced a “dialectic” of development, whereby infrastructure expansion and industrialization primarily benefited Japanese interests while Koreans experienced exploitation and economic marginalization.¹² By the late 1930s, a significant proportion of agricultural land had been transferred to Japanese ownership, and tenancy rates among Korean farmers rose sharply, reaching over half of the rural population. While agricultural output increased overall, much of the rice production was exported to Japan, resulting in what has been described as “starvation exports,” as per capita rice consumption for Koreans actually declined. Industrial development similarly favored Japanese firms, which dominated manufacturing and finance, while Koreans were largely confined to low-wage or unskilled labor. Although gross national product grew during the colonial period, wealth distribution remained highly unequal, and many Korean households struggled with poverty and limited access to resources.¹² These economic realities reshaped daily existence at a structural level. Scarcity influenced how homes were organized, how objects were produced, and how resources were used, fostering a culture of spatial and material economy that developed out of limitation rather than philosophical intention, as it was in the Joseon dynasty.

This article interprets these historical developments not merely as economic shocks but as catalysts that repurposed the Korean aesthetic landscape, as these economic realities heavily constrained daily existence. Drawing on these underlying data points, this paper highlights how the laws and economic restrictions that came into effect during colonial rule banned Koreans from extravagant decoration and lifestyle practices of material excess. Limitations on economic freedom of choice, coupled with unequal access to capital and resources, restricted both production and consumption. With Korea’s industries and cultural expression repositioned to serve the empire, the chances for ornamental display and artistic explorations diminished. Therefore, this paper theorizes that modern minimalistic tendencies were structurally reinforced by political control and structural inequity, consolidating what can be conceptualized as a “materially enforced minimalism.” The philosophical trend was not the only reason why minimalism prevailed; the limitation of economic and expressive possibilities due to colonial power was another primary historical cause.

Therefore, minimalistic tendencies became a result of political control and structural inequity. The philosophical trend was not the only reason why minimalism prevailed. The limitation of economic and expressive possibilities due to colonial power was another cause.

A similar pattern of minimalism under constraint can be observed later in time, in the early 20th century, and particularly within the evolution of hanok architecture. During this period, simplified housing models were introduced to accommodate rapid population growth, urban development, and economic hardship. As noted in urban planning records, housing for lower-income residents was reduced to a basic spatial composition of two rooms, one kitchen, and one living area. This arrangement became conceptualized as a “minimal unit,” a foundational layout that later influenced modernized hanok dwellings.¹³ Therefore, the concepts and designs that rose to prominence during times of economic strain influence Korean housing to this day.

Colonialism and scarcity did not produce Korean minimalism from nothing. Rather, they intensified and redirected existing tendencies toward simplicity. What began as an ethical and philosophical commitment to restraint during the Joseon period gradually evolved, in the twentieth century, into a condition shaped by political suppression, economic limitation, and post-war reconstruction. Minimalism in modern Korea, therefore, reflects a process of historical intensification: it shows how deep-seated philosophical values became structurally locked in by severe material constraints, embedding minimalism further into Korean society.

■ Modern Art

The convergence of inherited aesthetic restraint and twentieth-century material hardship became particularly visible in the emergence of modern art trends, such as Dansaekhwa in the late 1960s and 1970s. Often translated as “monochrome painting,” the movement reduced color to restrained palettes of white, umber, grey, and muted blue (all notably neutral colors). Artists such as Park Seo-Bo, Ha Chong-Hyun, and Lee Ufan developed practices that emphasized limited palettes, repetitive gestures, and the embedding of time and labor within the canvas, which foregrounded the process.¹⁴ The movement also resonated deeply with South Korea’s post-war condition. Developing during a period marked by recovery from the Korean War and rapid industrial modernization, Dansaekhwa reflected a society negotiating reconstruction and urban transformation. Its restrained surfaces and repetitive processes can be understood as responses to instability and rupture. Rather than portraying overt or sometimes controversial political imagery, these works conveyed persistence and a kind of strength that stemmed from their quietness.

Within Dansaekhwa, each of the core aesthetic principles discussed earlier appears in concrete, material ways. First, *yeobaek* (active emptiness) is visible in the large, unbroken monochrome fields that dominate many works. In paintings by Park Seo-bo, for example, vast areas of off-white canvas are repeatedly inscribed with pencil lines. The surface may appear empty at first glance, but close viewing reveals subtle variation and accumulated gesture. The “empty” space is therefore structured and intentional, not blank. The lack of additional colors or patterns also leaves space for the viewer to ponder and meditate, handing us a greater amount of agency and freedom for interpretation.

Second, *jayeon* (naturalness) is expressed through process and material choice, but also color. Artists such as Lee Ufan allowed paint to soak into raw canvas or applied pigment in ways that emphasized gravity and absorption.¹⁵ Rather than constructing illusionistic images, they foregrounded how materials behave: canvas absorbs, oil spreads, paper wrinkles, making natural processes central to the composition. Additionally, the colors used in Dansaekhwa art mostly revolve around colors that pre-exist organically in the natural world. This often gives the artworks a placid, soothing quality and supports the movement's broader emphasis on what is deemed as natural human feeling. In conclusion, Dansaekhwa can thus be seen not only as a modernist movement in its own right but also as a contemporary expression of traditional Korean aesthetic ideals.

■ Conclusion

This paper has argued that modern Korean minimalism cannot be understood as a contemporary trend or a stylistic choice, but rather as a product of complex historical processes that link the country's philosophical traditions to its material circumstances. The Joseon Dynasty, for instance, saw the spread of Neo-Confucianism and its associated aesthetic values that emphasized restraint, purity, balance, and naturalness. The very existence of terms like *yeobaek* and *jayeon*, and the symbolism of whiteness in ceramics, demonstrates that the values associated with simplicity were already deeply ingrained in Korea's ethical and aesthetic frameworks before the advent of the twentieth century.

It was, however, the impacts of colonialism, conflict, and economic hardship that saw these values altered in conditions of restraint and limitation. The economic restructuring and rebuilding that the Japanese occupation and post-war rebuilding necessitated saw a situation where there was a lack of access to resources and space, and it was here that minimalism transcended its role as a moral and philosophical stance and became a product of material circumstance and experience. The emergence of Dansaekhwa art in the late 1960s and 1970s represents a critical juncture in which these various strands came together to produce a distinctly modern and minimalist aesthetic that combined the legacy of Korea's aesthetic restraint with the experience and reality of modern Korea.

It can be seen, then, that modern Korean minimalism does not derive from a single source or event, but rather emerges from the complex interplay between the country's cultural and ethical philosophies and the various periods of political and economic upheaval that it has endured throughout the centuries. As a product of this complex interplay between choice and necessity, and tradition and innovation, it can be seen that modern Korean minimalist art and design represent a complex and multi-layered aesthetic that transcends its role as a contemporary trend. It is an aesthetic that was actively forged into the material reality of the Korean people through its structural resonance throughout Korean history.

■ AI Acknowledgments

During the preparation of this work, the author used Perplexity to find relevant sources and brainstorm topics. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

■ Acknowledgments

I would like to give special thanks to Isiajah Shadrach for his insightful comments on earlier drafts of this paper, and to all the editors who supported me throughout the publishing process.

■ References

1. Park, S.-J. (2022). Korean Aesthetic Ideals: 'Jayeon'. *The Journal of Aesthetics and Art Criticism*, 80(3), 357–361. <https://doi.org/10.1093/jaac/kpac028>
2. Morley, S. (2013). Dansaekhwa. *Third Text*, 27(2), 189–207. <https://doi.org/10.1080/09528822.2013.772350>
3. Seo, B.-H. (2015). White Hanbok as an Expression of Resistance in Modern Korea. *Journal of the Korean Society of Clothing and Textiles*, 39(1), 121–133. <https://doi.org/10.5850/jksct.2015.39.1.121>
4. Mitu, B. (2015). Confucianism and the Contemporary Korean Society. *Journal of Sociology*, 2, 1–15.
5. Deuchler, M. (2016). The Impact of Confucianism on the Korean Kinship System — A Reconsideration. *Journal of Confucian Philosophy and Culture*, 36(1), 5–23. <https://doi.org/10.22916/jcpc.2021..36.5>
6. Seo, B.-H. (2007). The Development of Gum-Chae-Saek and Aesthetic Thought. *Korean Visual Culture Studies*, 18(2), 88–104.
7. National Folk Museum of Korea. (1907). A Market Day in a Countryside Town [Photograph archive]. Visual Collections Registry.
8. Seo, S., & Hee, K. (2019). *A Study on the Symbolic Meanings of Jars*. *Journal of Sandplay Therapy*, 6(1), 109–130.
9. Shin, N. K. (2018). Korean Aesthetic Consciousness Reflected in the Ceramic Art of Joseon. *Journal of the Faculty of Letters, The University of Tokyo, Aesthetics*, 42, 30–52.
10. Tam, K. Y. (2023). Finding Cultural Identity through Korea's Moon Jars. *SJS @ USC, Special Issue*, 18–32.
11. Metropolitan Museum of Art. (2021). The Met Collection: Moon Jars. *Metmuseum.org Exhibition Registry*. <https://www.metmuseum.org/art/collection/search/45432>
12. King, R. (1975). Japanese Colonialism and Korean Economic Development, 1910–1945. *Asian Studies Journal*, 13(3), 4–21.
13. Kim, Y.-J., & Yun, A.-R.-M. (2023). Valorizing Ikseon-dong's Hanok residence as a heritage place. *Frontiers of Architectural Research*, 12(5), 947–965. <https://doi.org/10.1016/j.foar.2023.04.004>
14. Kee, J. (2013). *Contemporary Korean Art: Tansaekhwa and the Urgency of Method*. University of Minnesota Press.
15. Roquet, P. (2007). Reencountering Lee Ufan. *KCWorks*, 3, 85–98. <https://doi.org/10.17613/M6HH6C520>
16. Lee, Y. (2022). The white-clad people: The white hanbok and Korean nationalism. *Cultural Dynamics*, 34(4), 291–308. <https://doi.org/10.1177/09213740221117811>
17. Xiao, Y., & Hu, J. (2019). The Inheritance and Spreading of Confucianism in Modern China and South Korea. *Proceedings of the International Conference on Social Science, Humanities and Education*, 40–48. <https://doi.org/10.2991/icsshe-19.2019.40>

■ Author

Hiisan Kim is a Year 12 student living in Phuket, Thailand. Her academic interests mainly lie in Art history and philosophy, and her current research focuses on the intersections of historical ideology and aesthetic movements. In university, she desires to pursue Art History.

The Quantum Tunneling Effect: From Fundamental Theory to Key Applications

Kaito Asakawa

Athenian School, 2100 Mount Diablo Scenic Blvd, Danville, CA 94506, USA; kaito01xp@gmail.com

Mentor: Ifayoyinsola Ibikunle

ABSTRACT: This paper provides a comprehensive overview of quantum tunneling, bridging the gap between wave-particle duality and real-world applications. We analyze how tunneling facilitates solar energy production via the proton-proton chain and enables atomic-scale imaging through Scanning Tunneling Microscopy (STM). A particular focus is placed on the evolution of STM instrumentation; we discuss how innovations in closed-cycle cryostats and feed-forward noise cancellation have overcome traditional stability limitations. By linking stellar physics with nanoscale engineering, this work highlights the enduring relevance of tunneling in both natural and synthetic systems.

KEYWORDS: Physics, Quantum Mechanics, Quantum Tunneling, Nuclear Fusion, Scanning Tunneling Microscopy.

■ Introduction

The theoretical understanding of quantum physics:

Quantum tunneling represents one of the most non-intuitive concepts in the field of quantum mechanics. To illustrate this effect, consider a classical analogy: a ball rolled up a hill. What would one expect to happen? The answer is simple enough: with a weak force, the ball will roll back, and if one rolls it with enough force, it will go over the hill. In quantum mechanics, however, particles can sometimes pass through a barrier even when they do not have enough energy to go over it. One may think that such a phenomenon cannot occur. This counterintuitive behavior is known as quantum tunneling, a fundamental consequence of the wave-like nature of matter and the probabilistic interpretation of quantum states. This effect is not only interesting but has important applications ranging from how stars produce energy through nuclear fusion to how modern electronic devices, such as transistors and diodes, function.

To explain this strange and interesting behavior, one must have a fundamental understanding of quantum mechanics. This theory enables us to analyze and explain the behavior of tiny particles, such as atoms and subatomic particles. In quantum mechanics, not all outcomes are certain, and solving equations does not always give one an exact answer.

Instead, the answers are based on probabilities. Unlike classical mechanics, where the outcomes are deterministic, quantum mechanics enables particles to maintain a finite probability of being detected in regions that are classically forbidden. As a result, this probabilistic nature allows for tunneling events where particles can penetrate barriers like a hill, which would otherwise confine them.

While quantum tunneling appears to be a minor phenomenon, it has a crucial role in processes such as nuclear fusion in stars and the development of modern technologies (e.g., semiconductors and scanning microscopes). To that end, a comprehensive understanding of this effect is essential for both

foundational physics and applied research in nanotechnology & quantum engineering.

In this literature review, the discussion will focus on the following three points:

1. The theoretical principles of quantum tunneling
2. The manifestation of quantum tunneling in natural systems
3. Practical applications of quantum tunneling in laboratory and technological settings

■ Results and Discussion

Theoretical Foundations of Quantum Tunneling:

In order to understand quantum tunneling, this section begins by explaining the interesting way particles behave in quantum mechanics. At the atomic scale, particles exhibit what is called wave-particle duality, meaning they act like both particles and waves depending on the situation.

Think of a particle, such as an electron. The electron behaves more like a water wave that spreads out, overlaps with other waves, and does not stay in one exact spot. Instead of having a fixed position, it exists as a probability cloud across multiple locations. This wave-like property is described by the Schrödinger equation, which is a mathematical equation that represents the behavior of a particle in terms of a wave function shown below.

$$-\frac{\hbar}{2m} \frac{d^2\psi(x)}{dx^2} + U(x)\psi(x) = E\psi(x) \quad (1)$$

Equation (1) represents the time-independent Schrödinger equation in one dimension.¹ The equation not only has macroscopic importance but also has meaning in each component. The first term includes $\hbar$, the reduced Planck constant ($\hbar \approx 1.054 \times 10^{-34} \text{ J}\cdot\text{s}$). m is the mass of the particle, and $\frac{d^2\psi(x)}{dx^2}$ is the second derivative of the wave function with respect to position x . These terms represent the kinetic-energy operator. The second term $U(x)\psi(x)$ represents the potential energy acting on

the wave function, and $E\psi(x)$ represents the total energy eigenvalue acting on the same state. Together, the kinetic-energy and potential-energy operators acting on $\psi(x)$ yield $E\psi(x)$. The equation tells us that $\psi(x)$ must be continuous at each boundary it encounters, and its derivative must also be continuous.² This wave function does not solve for an exact position for the particle but rather gives the probability of finding it in a particular place. As a result, particles can defy classical rules as seen in the quantum tunneling effect.

Quantum Tunneling Explained Through the Heisenberg Uncertainty Principle:

In quantum mechanics, since particles act as waves, there is a slim probability that the particle's wavefunction can extend and pass through the energy barrier. This possibility is what we call the quantum tunneling effect. This tunneling probability is derived by solving the Schrödinger equation for a potential barrier. The Heisenberg uncertainty principle can be used as an intuitive interpretation. One form of this principle relates energy and time. Given a short amount of time to examine a particle, the uncertainty in the particle's energy increases significantly. Here, the Heisenberg uncertainty principle is used as an intuitive interpretation rather than the main mechanism. Over a very short time interval, the particle's energy can be less sharply defined, which helps explain why barrier penetration is not strictly forbidden in quantum mechanics. However, the standard description of tunneling comes from the Schrödinger equation: the wavefunction penetrates the classically forbidden region and decays exponentially, and the transmitted amplitude is determined by matching the wavefunction and its derivative at the boundaries.²

The Heisenberg uncertainty principle can be mathematically expressed as:

$$\Delta x \Delta p \geq \frac{\hbar}{2} \quad (2)$$

where Δx represents the uncertainty in a particle's position, and Δp represents the uncertainty in its momentum.² The Heisenberg uncertainty principle offers an intuitive way to think about quantum tunneling: a short-lived quantum state means the particle's energy is less sharply defined, which helps explain why barrier penetration is not strictly forbidden in quantum mechanics.

Calculating Tunneling Probability Using the Schrödinger Equation:

Beyond the conceptual understanding of how quantum tunneling works, calculations can be made to determine the probability that a wave function can tunnel through a specific finite potential barrier. This is known as transmission probability, and to understand this equation, we must look at the wave equations associated with the potential barrier.²

To begin with, we are assuming that the potential barrier is from $0 \leq x \leq L$, which can be expressed as a piecewise function. By setting the finite potential barrier, we create 3 different regions with their own unique wave function.³

$$U(x) = \{0, x < 0, U_0, 0 \leq x \leq L, 0, x > L\} \quad (3)$$

Equation (3) represents the time-independent Schrödinger equation in one dimension.^{3,4} From now on, we are calling region 1 for $x < 0$, region 2 for $0 \leq x \leq L$, and region 3 for $x > L$.

Separating into 3 regions enables setting up Schrödinger equations for each region.

$$\text{Region 1: } -\frac{\hbar}{2m} \frac{d^2\psi(x)}{dx^2} = E\psi(x)$$

$$\text{Region 2: } -\frac{\hbar}{2m} \frac{d^2\psi(x)}{dx^2} + U(x) = E\psi(x)$$

$$\text{Region 3: } -\frac{\hbar}{2m} \frac{d^2\psi(x)}{dx^2} = E\psi(x)$$

Boundary conditions require that the wave functions and their derivatives are continuous on each boundary (i.e., $x = 0$ and $x = L$). The general solution for each region,

Region 1: $\psi(x)_1 = Ae^{ikx} + Be^{-ikx}$, representing both incident and reflected waves.

Region 2: $\psi(x)_2 = Ce^{-\alpha x} + De^{\alpha x}$, where the wave decays or grows exponentially inside the barrier.

Region 3: $\psi(x)_3 = Fe^{ikx}$, representing a transmitted wave moving only to the right (we neglect the reflected component in this region), where $\alpha = \sqrt{\frac{2m}{\hbar^2}(U_0 - E)}$, $k = \sqrt{\frac{2m}{\hbar^2}E}$.

There are three regions in this setup, and the wave (particle) behaves differently in each of the regions. The wave functions exhibit the behavior of the particle in each region as follows: In Region 1, the particle can move in both directions, with incident and reflected waves; In Region 2, the exponentially decaying wave is in the potential barrier; In Region 3, the particle moves only in the forward direction, i.e., tunneling was successful, and there is no backward traveling wave.⁴ The tunneling probability depends on the wave function intensities, specifically, the ratio of the transmitted wave in Region 3 to the incident wave in Region 1.

While calculating the exact values requires applying boundary conditions at the interfaces ($x = 0$ and $x = L$) to match wave functions and their derivatives, the essential idea is that tunneling probability emerges from how these wave components match across the three regions.⁴

Figure 1 illustrates this three-region tunneling setup and the transmitted wave after the barrier.



Figure 1: The visualization of quantum tunneling. Adapted from Physics LibreTexts.⁴ This figure shows that even when the particle's energy is below the barrier height, part of the wavefunction can extend through the barrier and produce a transmitted wave.

A potential energy barrier of height U_0 creates three physical regions with three different wave behaviors. In region 1, where $x < 0$, an incident particle moves in a potential-free zone and coexists with a reflected particle. In region 2, a part of the incident wave that has not been reflected at $x = 0$ moves as a transmitted wave in a constant potential $U(x) = U_0$ and tunnels through to region 3 at $x = L$. In region 3 for $x > L$, a

transmitted particle that has tunneled through the potential barrier moves as a free particle in a potential-free zone. The energy E of the incident particle is indicated by the horizontal line.⁴

After the long algebraic calculations, we get the following equation for tunneling probability:

$\therefore T \approx 16 \frac{E}{U_0} (1 - \frac{E}{U_0}) e^{-2kL}$, where $k = \frac{\sqrt{2m(U_0-E)}}{\hbar}$, U_0 is the barrier height, L is the width of the barrier, and E is the particle's total (incident) energy.

Table 1 summarizes the tunneling probabilities predicted by this equation for several particle energies and barrier widths. The values show a trend of how sensitive tunneling is to even small changes in barrier thickness; as the barrier becomes wider, the tunneling probability drops rapidly.

Table 1: Quantum tunneling probability for different combinations of particle energy (eV) and barrier width (nm). The table shows that the tunneling probability decreases sharply as the barrier width increases, while higher particle energy slightly increases the probability.

Barrier width (nm)	0.1 eV	0.3 eV	0.5 eV	0.7 eV	0.9 eV
0.2	6.23E-03	6.77E-03	7.37E-03	8.03E-03	8.77E-03
0.4	3.88E-05	4.58E-05	5.43E-05	6.45E-05	7.69E-05
0.6	2.42E-07	3.1E-07	4E-07	5.19E-07	6.75E-07
0.8	1.51E-09	2.1E-09	2.95E-09	4.17E-09	5.92E-09
1.0	9.38E-12	1.42E-11	2.17E-11	3.35E-11	5.19E-11

The barrier energies of 0.1–0.9 eV are common for electrons; the barrier widths of 0.2–1.0 nm are about the size of a few atoms, which is the scale where tunneling usually happens.⁴ For the most part, quantum tunneling emerged as a calculable process with measurable probabilities. A small difference in any of these parameters can drastically cause a particle to pass or not pass through the energy barrier.

Quantum Tunneling in Stellar Fusion:

Quantum tunneling has particular significance in nuclear fusion. Since protons are positively charged particles, they have an inherent tendency to repel one another through electrostatic interaction. Quantum tunneling helps them to overcome this repulsive force so that when protons come sufficiently close, fusion can take place. The fusion process, when it occurs, releases a huge amount of energy, which energizes the Sun. This is also the cause of the light and heat received on Earth. In the absence of quantum tunneling, pressures and temperatures used in nuclear fusion would be much greater, and therefore it is most unlikely to occur in stars like the Sun.⁶

To see how quantum tunneling allows fusion to occur in the Sun, it is useful to examine the particular reactions being performed in the solar core. In stars at temperatures less than fifteen million Kelvin, like the Sun, the proton-proton chain is the prevailing fusion reaction. In this reaction, four protons eventually fuse to create a helium-4 nucleus, and energy is released in the form of photons and neutrinos.¹¹ The initial reaction in this sequence is likewise the most important reaction for tunneling: two protons need to get near enough to each other to fuse into a deuteron, which contains one proton and one neutron. Because of the positive charge of the protons, though, there will be a strong Coulombic barrier, a repulsive force creating a potential energy hill between them.

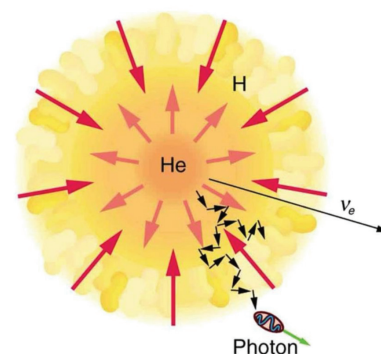


Figure 2: Fusion in the Sun's Core Diagram (fusion with $H \rightarrow He + \text{photons} + \text{neutrinos}$). Adapted from Texas Gateway.⁷ This figure shows that hydrogen fusion in the solar core produces helium, photons, and neutrinos, and that quantum tunneling makes this process possible at solar temperatures.

Figure 2 shows how nuclei of hydrogen are fused to create helium in the Sun's core, and release energy in the form of photons (light) and neutrinos (particles that travel through space). The process of fusion occurs at the center of the Sun, where it is hottest and densest. Importantly, quantum tunneling is what makes this entire fusion process possible at the temperatures in the Sun.⁷

Tunneling Probability and the Gamow Peak:

Fusion in the Sun depends on the Coulomb repulsion between protons and the probability of penetrating that barrier. The next figure illustrates the Coulomb barrier that protons must overcome to fuse.

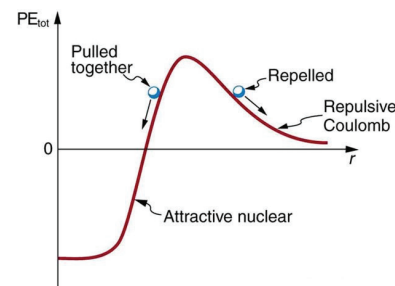


Figure 3: Potential energy between two protons as a function of distance. Adapted from Texas Gateway.⁷ This figure highlights the Coulomb barrier that resists proton fusion and the attractive nuclear region that protons can reach only when they get close enough.

Figure 3 illustrates the potential energy between two protons as they approach each other. The steep hill represents the Coulomb repulsion, which resists the protons getting close enough to fuse. In the Sun, because their kinetic energy alone is not sufficient, quantum tunneling is the process that allows them to penetrate this barrier and reach the attractive region where nuclear fusion can occur.⁸

$$V(r) = \frac{Z_1 Z_2 e^2}{4\pi\epsilon_0 r} \quad (4)$$

Equation (4) shown above represents the Coulomb potential energy of two interacting nuclei as a function of the distance r of their interaction. Z_1 and Z_2 are the charges of the interacting nuclei (both 1 for protons), e is the elementary charge, ϵ_0 is the permittivity of vacuum, and r is the distance between the protons. As the protons approach one another, their electrostatic repulsion of like positive charges causes this potential

energy to rise quickly, creating what is known as the Coulomb barrier. This is an effective barrier to fusion: its effective height is much greater than the kinetic energies of the protons in the core of the Sun.

In classical physics, this would inhibit fusion completely, since the protons would not have sufficient energy to cross over the barrier. Quantum mechanics, on the other hand, yields a different result. The wave function of the proton penetrates in and beyond the barrier, leading to a finite probability of the proton on the other side of the barrier through a process known as quantum tunneling.

Mathematically, this probability is described by a transmission coefficient for a potential barrier, derived from solving the Schrödinger equation for this situation. Equation (5) is similar in form to the tunneling probability described for a square potential barrier:

$$T \approx e^{-2G} \quad (5)$$

where G is known as the Gamow factor, given by:

$$G = \frac{2\pi Z_1 Z_2 e^2}{\hbar v}$$

The Gamow factor quantifies how difficult it is for two positively charged nuclei to tunnel through their Coulomb barrier and fuse. In this formula, Z_1 and Z_2 are the charge numbers of the nuclei (both 1 for protons), e is the elementary charge, $\hbar$ is the reduced Planck's constant, and v is the relative velocity of the nuclei. Higher charges or lower velocities make G larger, so e^{-2G} it becomes very small, meaning tunneling, and therefore fusion, is much less likely. However, the Sun's core contains an enormous number of protons with a spread of velocities given by the Maxwell-Boltzmann distribution, so enough protons still tunnel to sustain fusion.

The fascinating interplay between this velocity distribution and the tunneling probability results in what is called the Gamow peak. The Gamow peak identifies the most probable energy range for fusion to occur, balancing two competing factors: Higher-energy protons are rarer (because of the Maxwell-Boltzmann distribution), but as proton energy increases, the tunneling probability also increases due to the decreasing Gamow factor. So simply put, this combined effect determines where fusion is most likely to occur in the Sun's core.

Figure 4 shows this balance as the Gamow peak.

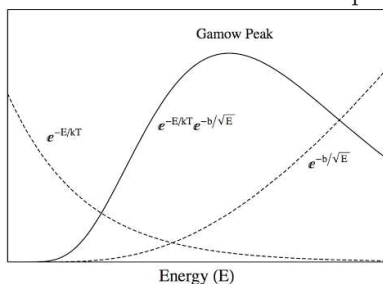


Figure 4: This diagram shows the Gamow peak, which shows the most likely energy range where fusion happens in the Sun. The Gamow peak results from multiplying the Maxwell-Boltzmann distribution of proton energies, the number of protons at a given energy, $e^{-E/kT}$ and the probability of tunneling through the Coulomb barrier $e^{-b/\sqrt{E}}$.⁹ This figure shows that fusion is most likely at an intermediate energy where enough protons exist and tunneling is still probable.

Reaction Probability in the Solar Core

Using quantum tunneling, protons in the Sun's core can overcome the Coulomb barrier at energies where fusion is most likely to occur.⁸ The typical proton energy in the solar core is only about 1 keV, while the Coulomb barrier energy is roughly 500 keV, which makes fusion impossible according to classical physics since the protons do not have sufficient kinetic energy to overcome the barrier.⁴ However, using quantum mechanics, the proton's wave function extends into and across the barrier, creating a finite probability that the proton will be found on the other side. As a result, protons with energies near the Gamow peak (~10-20 keV effective energy) can successfully tunnel through the barrier and fuse, despite having total energies below the barrier height.⁸

Figure 5 compares the classical repulsion case with the quantum tunneling case.

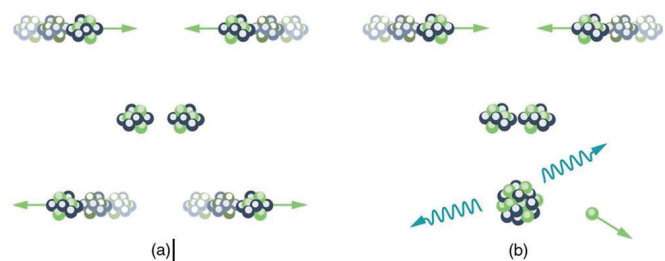


Figure 5: This diagram compares two scenarios when protons approach each other. In part (a), without enough energy or tunneling, the protons repel and bounce away. In part (b), quantum tunneling allows the protons to bypass the repulsive barrier, allowing protons to fuse and release energy, shown as emitted light and particles.⁴ This comparison shows why tunneling is necessary for solar fusion at the Sun's core temperature.

This effect fundamentally determines the rate of fusion in the Sun. The fusion cross-section is essentially the "reaction probability" or an effective area that represents the likelihood of a reaction taking place during a collision. For two protons, fusion is heavily influenced by the tunneling probability and the proton energy:

$$\sigma(E) \propto \frac{1}{E} e^{-2G} \quad (6)$$

Equation (6) represents the fusion cross-section for two protons in the Sun.¹⁰ The $1/E$ term factor is a slowly varying prefactor in this expression. The dominant reason fusion is suppressed at low energies is that the tunneling term e^{-2G} , which decreases exponentially as proton energy decreases. This term contains the quantum tunneling probability through the Coulomb barrier, with G being the Gamow factor. Exponential suppression e^{-2G} takes over at low energies, which indicates that without tunneling, fusion is extremely improbable in the core of the Sun. In the absence of tunneling, the fusion cross-section would be essentially zero since protons would not surmount their electrostatic repulsion at the core temperature. Computing precise fusion rates in the Sun, therefore, entails integrating this tunneling-modulated cross-section over all proton energies in the core, as characterized by the Maxwell-Boltzmann energy distribution.¹¹

In recent decades, precise modeling of the Sun's energy output, neutrino production, and fusion rates has relied heavily on improved calculations of tunneling probabilities. The

modern Standard Solar Model (a theoretical/computational model) incorporates these effects in detail and enables accurate predictions of phenomena such as the solar neutrino flux. These predictions have been confirmed by experiments like Super-Kamiokande and the Sudbury Neutrino Observatory (SNO).¹²

The Astrophysical Impact of Quantum Tunneling:

The consequences of this quantum effect are profound. First, it allows the Sun to fuse hydrogen at relatively “modest” temperatures compared to what would be required in a purely classical system. Without tunneling, the necessary core temperature for fusion would need to be hundreds of millions of Kelvin, not the observed ~15 million Kelvin.¹¹ Second, it provides a delicate balance that regulates the Sun's energy output. Tunneling probability's temperature dependence enables infinitesimal core temperature changes to produce vast differences in fusion rates, and this is a natural feedback process that stabilizes the structure of the Sun over billions of years.⁶

Moreover, the detailed understanding of quantum tunneling in the Sun has broader implications for astrophysics. The same principles apply to fusion in other stars, including those much larger or smaller than the Sun. In massive stars, tunneling allows fusion of heavier elements like carbon and oxygen. In cooler, smaller stars, tunneling determines whether they can sustain stable fusion at all. Even the synthesis of elements in the early universe, Big Bang nucleosynthesis, relied on quantum tunneling to enable fusion reactions under extremely low-density, rapidly cooling conditions.⁸

In summary, the theory behind the Sun's energy production is quantum tunneling, allowing protons to overcome their mutual repulsion and initiate fusion at solar core temperatures. The tunneling probability, by the Gamow factor and the Maxwell-Boltzmann distribution of protons, defines the fusion rates that power the Sun and sustain life on Earth. In practical terms, every photon of sunlight that reaches Earth owes its existence to quantum tunneling. As Patchali concludes: “The fact that our sun is a star might seem magical, but it's really just quantum mechanics.”⁷⁹

Applications of Quantum Tunneling Technology:

By applying the tunneling mechanics that allow fusion in the Sun, scientists and engineers have managed to develop tools named the Scanning Tunneling Microscope (STM). It can be utilized to observe individual atoms and manipulate electronic behaviors at the nanoscale.¹³

Principles of STM operation:

The STM was invented by Gerd Binnig and Heinrich Rohrer in 1981.¹³ It visualizes surfaces at the atomic level by measuring tunneling currents between a sharp metal tip and a conducting surface; electrons can tunnel through a vacuum between the tip and the sample surface when voltage is applied, and the tip is brought extremely close (< 1 nanometer). Recently, STM has been used to explore and study materials like superconductors, which have the potential for advancing quantum computing technology. The tunneling current de-

pends exponentially on the distance between the tip and the surface.¹⁶ Because of this high sensitivity, STM can achieve sub-ångström resolution, and topographic maps of surface atoms can be created by scanning the tip with a piezoelectric scanner and a positioning stage that moves the tip in x - y - z with tiny steps. The scanner uses piezoelectric ceramic actuators (electrically responsive crystals) that expand or contract when a voltage is applied to produce those precise motions. The microscope is typically operated in constant-current mode in most applications, where the feedback system translates the tip in a manner that maintains a constant current. This correction translates to an image of the electronic density of the surface.¹⁵ The process is shown in Figure 6, which demonstrates the mechanism of the piezoelectric scanner and feedback loop working together to maintain the tip's height and produce a high-resolution image.

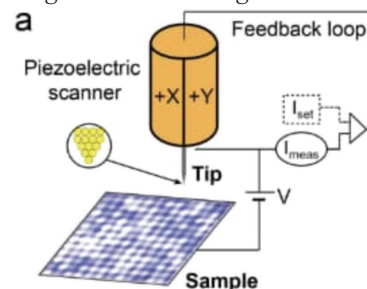


Figure 6: Basic principles of STM. Schematic representation of STM, where V , a voltage, is applied between the tip and the sample. The tip is scanned across the surface in the x - y plane, and its z coordinate is adjusted using the three-dimensional piezoelectric scanner controlled by a feedback loop, which automatically adjusts the tip's height to maintain a constant tunneling current as the surface contour changes. Adapted from Boston College.¹⁵ This figure shows how maintaining a constant tunneling current allows STM to produce atomic-scale surface images.

Using STM as a Research tool Today, this capability to produce high-resolution images positions STM as an important tool in various fields such as materials science, nanotechnology, and semiconductor studies. This atomic-level imaging and control are needed to define material properties, build electronics, and create nanoscale devices. For instance, researchers at the University of Oxford recently created the Andreev STM technique, which has been used to study electronic behavior in superconducting materials that are relevant to quantum materials research. This allows for imaging and observing topological superconductors' unusual quantum behaviors.¹⁹ Similarly, researchers at the University of Illinois have employed STM to "wire up" individual graphene nanoribbons.²⁰ This creates the atomically precise metal contacts, essential for future nanoelectronics devices. STM serves as a leading research tool that stimulates the creation of new technologies across quantum information science and nanoengineering.^{19,20}

Challenges: Environmental Disturbances:

On the other hand, the STM has a drawback in that the majority of quantum mechanical machines struggle with external disturbances. The STM's performance can be drastically degraded by surrounding equipment, temperature fluctuations, and air currents. Since distorting the tunneling current a bit or placing the tip by mere picometers can severely affect the

result, the STM must be meticulously cautious with slight changes in environment, including surrounding equipment, temperature fluctuations, and air currents.

Advances in Vibration and Thermal Control:

Considering this issue, placing an STM in an isolated, thermally stabilized environment can also improve the measurement accuracy of the STM. For instance, cryogenic environments that can suppress both mechanical noise - any tiny vibrations or noise from the environment - and thermal drift, which refers to position shifts caused by temperature fluctuations, can preserve STM performance and enable researchers to have more opportunities to achieve consistently accurate results. According to Hackley *et al.*, a closed-cycle cryostat STM is one of the solutions. They managed tip vibrations of less than 1.5 picometers, which is over 10 times smaller than typical temperature STM systems that produce vibrational noise in the range of 20 picometers or above. They also implemented a minuscule lateral drift of 0.18Å per hour at about 16 K temperatures with the mechanical isolation and low-vibration cryogenic cooling.¹⁷ Moreover, in the past decades, the severe shortage of helium has made it increasingly difficult to conduct low-temperature scanning probe microscopy (SPM) experiments. This motivates cryogen-free STM designs that can reach low temperatures without liquid helium.

Building on this, researchers led by Kasai developed a cryogen-free pulse-tube STM, which achieved operating STM near 5K and atomic-resolution imaging on gold surfaces and superconducting materials. Using the helium exchange gas and mechanical decoupling layers, they achieved lattice resolution on Au(111) surfaces and superconducting gap spectroscopy Pb(lead) with high stability. Specifically, their system incorporates a home-made interface between the PTR (pulse-tube refrigerator) cold head and the SPM head to minimize vibrational coupling and an automatic gas-handling system to regulate helium exchange pressure. This innovation performs at the same level as the conventional systems requiring liquid helium, but also alleviates the environmental concerns.¹⁸ Altogether, by using integrated cryogen-free designs under appropriate mechanical and thermal control, STM can overcome its vulnerability to external disturbances and can operate at atomic resolution consistently.

Feed-Forward Cancellation (Correction Technique):

Beyond these advancements, considering the current challenges STM faces, implementing a PTR-synched feed-forward cancellation can greatly reduce the effects of external interference on STM measurement. The main idea is to accept that it is almost impossible to create such an environment where the STM does not receive any external disturbance from the pulse-tube refrigerator (PTR), a vital device for operating the STM at low temperatures. However, if we could correct the mistakes the STM made rapidly, we could create a disturbance-free environment as close as possible. Specifically, this technique is called feed-forward cancellation, and by applying it to the small vibration sensor and automating the adjustment of the STM scanning tip in advance to predict the disturbance, we

can greatly reduce the image noise of the atomic surface and improve the accuracy in spectroscopic measurement, which gives us a better understanding of how electrons occupy energy levels in specific surface regions.

Taken together, STM visualizes surfaces at the atomic level by measuring tunneling currents between a sharp metal tip and a conducting surface. However, external disturbance is a long-lasting issue because it destabilizes the tip and reduces imaging accuracy. To countermeasure this issue, a closed-cycle cryostat STM, a cryogen-free pulse-tube, and a PTR-synched feed-forward cancellation are promising approaches that bring STM closer to a disturbance-free environment, allowing researchers to capture more accurate and reliable data.

■ Conclusion

From the theory of quantum tunneling to its role in solar fusion in the Sun and its application in scanning tunneling microscopes, quantum tunneling is a foundational concept necessary for explaining concepts in astrophysics and modern technology. In stars, the protons tunnel through the electrostatic repulsion to generate the light and heat that reaches the Earth as sunlight. In the lab setting, STM uses the same principle to visualize atoms and electronic structures at sub-angstrom resolution. Its performance, however, can be strongly affected by external disturbances. Methods such as closed-cycle cryostats, pulse-tube cryogen-free refrigerators, and feed-forward cancellation can prevent performance deterioration of STM, reducing image noise and improving spectroscopic precision. The broad applicability of quantum tunneling is in the fundamental process driving the evolution of stars and its contribution to progress in quantum computing, nanoelectronics, and materials science. Future work could test the proposed PTR-synched feed-forward approach by comparing STM images and spectra before and after correction under the same PTR operating conditions. In particular, quantifying changes in low-frequency vibration peaks and reductions in line striping would help evaluate effectiveness. This would clarify how much improvement can be achieved beyond established cryogen-free and closed-cycle stabilization methods.

■ Appendix

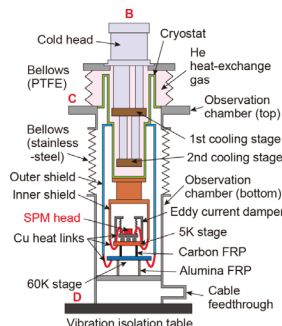


Figure A1: Schematic diagram of the cryogen-free pulse-tube STM system developed by Kasai *et al.*¹⁸ The system integrates a PTR cold head, mechanical decoupling layer, and helium gas exchange interface to minimize vibrational coupling while achieving low base temperatures suitable for measuring atomic-level resolution. Adapted from Kasai *et al.* This figure shows how mechanical decoupling and thermal control are used to reduce vibration in a cryogen-free STM system.

Supplemental information:

Technical methodology for PTR-synced feed-forward cancellation

This type of strategy has been applied in robotics, where predicting and compensating for disturbances in advance helps prevent repeated errors and maintain system stability.²⁴ For instance, this feed-forward control has been utilized to control expected counteracting forces to increase precision in robot manipulation. Translating the same idea to STM, the feed-forward correction is timed to the known PTR cycle and applied just before the disturbance reaches the junction.

Proposed Experimental Setup and Procedure (PTR-Synced Feed-Forward):

Mount a small accelerometer on the pulse-tube cold head (or read the valve-timing output), park the tip over a flat region, and record the accelerometer signal and tunneling current simultaneously for ~60 seconds to identify the ~2–3 Hz component and any harmonics.¹⁸ Using those frequencies and phases, time the most sensitive parts of each scan to the quiet portion of the cycle, and apply tiny, phase-matched nudges to the z piezo (extend to x-y if needed) right before each disturbance. Start at picometer amplitudes and use a simple adaptive model, such as a lightweight machine-learning regressor or autoregressive filter that retunes amplitude and phase every few seconds, to stay locked as conditions drift. Monitor the current spectrum for a drop in the ~2–3 Hz peak and inspect the images for reduced striping, a sharper lattice, and cleaner spectroscopy. For spectroscopy, check for clearer traces (i.e., differential conductance vs. voltage), a standard proxy for the local density of states. Enforce hard limits on the nudge size for safety, and pause or reduce the correction if any signs of junction instability have appeared. Save the sensor traces, correction signals, and before/after plots for repeatability, and, if needed, add the first harmonic or extend the same approach to x-y to address slow drift.¹⁸

This setup is illustrated in Figure A1.¹⁸

Acknowledgments

The author would like to thank Ifayoyinsola Ibikunle for mentorship and feedback during the revision process. The author also thanks the Indigo Research program for editorial guidance.

AI Acknowledgments

During the preparation of this work, the author used ChatGPT for brainstorming and language editing. The author reviewed and revised all outputs and takes full responsibility for the final content.

References

- Schleich, Wolfgang P., *et al.* "Schrödinger Equation Revisited." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 14, Apr. 2013, pp. 5374–79, <https://doi.org/10.1073/pnas.1302475110>.
- Lea, Robert. "Quantum Tunneling: The Phenomenon that Powers the Stars." *AZoQuantum*, 28 Feb. 2020, <https://www.azoquantum.com/Article.aspx?ArticleID=199>. Accessed 20 Feb. 2026.
- Ling, Samuel J., *et al.* *University Physics*. Volume 3. OpenStax, Rice University, 2016.
- "7.7: Quantum Tunneling of Particles through Potential Barriers." *Physics LibreTexts*, [https://phys.libretexts.org/Bookshelves/University_Physics/University_Physics_\(OpenStax\)/University_Physics_III_-_Optics_and_Modern_Physics\(OpenStax\)/07%3A_Quantum_Mechanics/7.07%3A_Quantum_Tunneling_of_Particles_through_Potential_Barriers](https://phys.libretexts.org/Bookshelves/University_Physics/University_Physics_(OpenStax)/University_Physics_III_-_Optics_and_Modern_Physics(OpenStax)/07%3A_Quantum_Mechanics/7.07%3A_Quantum_Tunneling_of_Particles_through_Potential_Barriers). Accessed 16 Jul. 2025.
- PhET Interactive Simulations. "Quantum Tunneling and Wave Packets." *University of Colorado Boulder*, <https://phet.colorado.edu/en/simulation/quantum-tunneling>. Accessed 20 Feb. 2026.
- "Nuclear Fusion and Energy Production in Stars." *Sparkl*, <https://www.sparkl.me/learn/ib/physics-hl/nuclear-fusion-and-energy-production-in-stars/revision-notes/1453>. Accessed 15 Feb. 2026.
- Texas Gateway. "15.4 Fusion." *Texas Gateway*, <https://texasgateway.org/resource/154-fusion>. Accessed 16 Oct. 2025.
- Aliotta, Marialuisa, *et al.* "Exploring Stars in Underground Laboratories: Challenges and Solutions." *Annual Review of Nuclear and Particle Science*, vol. 40, 2025, p. 25, <https://doi.org/10.1146/annurev-nucl-110221-103625>. Accessed 18 Oct. 2025.
- Patchali, Tchilabalo Essossimna. *Quantum Tunneling in the Sun*. 2015. https://www.academia.edu/11446310/Quantum_tunneling_in_the_sun. Accessed 10 May 2026.
- Lindsey, Martin Louis, *et al.* "Dynamically Assisted Nuclear Fusion in the Strong-Field Regime." *Physical Review C*, vol. 109, no. 4, Apr. 2024, <https://doi.org/10.1103/PhysRevC.109.044605>.
- Townsend, Rich. *PHYS 633: Introduction to Stellar Astrophysics*. 2006.
- "Neutrino Oscillations." *University of Victoria*, <https://www.phys.uvic.ca/~thompson/phys216/neutrinos.pdf>. Accessed 2 Aug. 2025.
- Rohrer, Heinrich. "Photo from the Nobel Foundation Archive." *Nobel Prize in Physics*, 1986, <https://www.nobelprize.org/prizes/physics/1986/rohrer/facts>.
- "Scanning the Past." *Nature Nanotechnology*, vol. 8, no. 8, Nature Publishing Group, 2013, p. 539, <https://doi.org/10.1038/nnano.2013.167>.
- Boston College. "Scanning Tunneling Microscopy (STM)." *Zeljko Lab – Department of Physics, Boston College*, <https://www.bc.edu/bc-web/schools/morrissey/departments/physics/labs/Zeljko-Lab/research/facilities.html>. Accessed 10 Feb. 2026.
- "Scanning Tunneling Microscopy (STM)." *Oxford Instruments*, <https://afm.oxinst.com/modes/scanning-tunneling-microscopy-stm>. Accessed 16 Jul. 2025.
- Hackley, Jason D., *et al.* "High-Stability Cryogenic Scanning Tunneling Microscope Based on a Closed-Cycle Cryostat." *Review of Scientific Instruments*, vol. 85, no. 10, American Institute of Physics Inc., 1 Oct. 2014, <https://doi.org/10.1063/1.4897139>.
- Kasai, Jun, *et al.* "Development of a Near-5-Kelvin, Cryogen-Free, Pulse-Tube Refrigerator-Based Scanning Probe Microscope." *Review of Scientific Instruments*, Apr. 2022, <https://doi.org/10.1063/5.0084888>.
- "New Quantum Visualization Techniques Could Accelerate the Arrival of Fault-Tolerant Quantum Computers." *University of Oxford*, <https://www.ox.ac.uk/news/2025-05-30-new-quantum-visualisation-techniques-could-accelerate-arrival-fault-tolerant-quantum>. Accessed 17 Jul. 2025.
- Huang, Pin Chiao, *et al.* "Sub-5 Nm Contacts and Induced p-n Junction Formation in Individual Atomically Precise Graphene

- Nanoribbons.” *ACS Nano*, vol. 17, no. 18, Sep. 2023, pp. 17771–78, <https://doi.org/10.1021/acsnano.3c02794>.
21. Eli Bourassa, J., *et al.* “Blueprint for a Scalable Photonic Fault-Tolerant Quantum Computer.” *Quantum*, vol. 5, Feb. 2021, pp. 1–38, <https://doi.org/10.22331/Q-2021-02-04-392>.
22. Trixler, Frank. “Quantum Tunneling to the Origin and Evolution of Life.” *Current Organic Chemistry*, vol. 17, 2013.
23. Ximeng, B. Y.; Lyding, W. J. *Large-Scale Graphene Transfer in Ultra-High Vacuum and Design of a Low Temperature Ultra-High Vacuum Scanning Tunneling Microscope*. University of Illinois, [2014].
24. Biagiotti, Luigi, *et al.* “A Plug-in Feed-forward Control for Sloshing Suppression in Robotic Teleoperation Tasks.” 2018. https://iris.unimore.it/retrieve/e31e124d-88ba-987f-e053-3705fe0a095a/HarmonicFilter4Sloshing_IROS2018_final.pdf. Accessed 10 May 2026.

■ Author

Kaito Asakawa – Graduate of The Athenian School in Danville, Northern California, and incoming physics student at the University of California, Santa Barbara. Mainly does soccer, badminton, and beatboxing for hobbies.

Analyzing the Effect of Japanese Barberry on Lyme Disease Using Geographical and Epidemiological Analysis

Kavya M. Ballal

Kent Place School, 42 Norwood Ave, Summit, NJ, 07901, USA; ballalkavya@gmail.com

ABSTRACT: Japanese barberry (*Berberis thunbergii*) is an invasive shrub widely distributed in the northeastern United States and has been linked to increased risk of Lyme disease through its effects on blacklegged tick (*Ixodes scapularis*) habitat. This study investigates the geographic relationship between Japanese barberry abundance, tick populations, and Lyme disease incidence across the United States, with a focused county-level analysis in New Jersey. Publicly available datasets from environmental surveys and public health records were analyzed and visualized using choropleth maps and time-series graphs. Results show a strong geographic overlap between high barberry occurrence reports, established tick populations, and elevated Lyme disease cases, particularly in forested regions of the Northeast. In New Jersey, counties with extensive forest cover and higher reported barberry presence, including Morris and Hunterdon counties, exhibited the highest Lyme disease rates, while counties with documented control efforts showed comparatively lower incidence. Analysis of statewide regulations revealed no immediate reduction in Lyme disease cases following barberry bans, suggesting that existing infestations and ecological lag effects limit short-term policy impact. These findings indicate that Japanese barberry contributes to Lyme disease risk by supporting favorable tick habitat and highlight the need for long-term invasive species management strategies that integrate environmental and public health policy.

KEYWORDS: Environmental Science, Ecology, Japanese Barberry, Lyme Disease, Invasive Species.

■ Introduction

Invasive Species:

The intersection of environmental science and public health is rarely considered, yet invasive species pose a distinct threat that impacts both. Invasive species are non-native plants that cause environmental or economic harm. Common types include shrubs like the Japanese Barberry, vines and creepers like Kudzu and English Ivy, and perennials like garlic mustard and giant hogweed. Common tick sources include the invasive species, tall grass, and suburban backyards with moist shaded areas and leaf litter. This study focuses on one such connection: the relationship between the invasive Japanese barberry and the prevalence of Lyme disease. Invasive species are non-native organisms that cause significant harm to the environment, economy, and human and animal health.¹ These species can include plants, insects, fungi, or animals, and their presence has become a growing problem around the world and particularly, it is a local problem in New Jersey.¹⁻³ Invasive species often spread quickly because they lack natural predators in their new environments, which gives them an advantage over native species.⁴ The presence of invasive species has a widespread impact on nature, as they outcompete native plants and animals, disrupting ecosystems.¹ This, in turn, affects businesses and individuals, specifically homeowners, park workers, and even recreational industries like fishing and boating, which are burdened by the extra costs and labor of managing invasive species, costing them billions of dollars globally.⁵ In fact, New Jersey alone loses over \$290 million annually in agricultural damage due to invasive species.³ Despite all this, New Jersey is one of only two states in the country that doesn't ban the sale of

some of the worst invasive plants, making the issue even harder to control.⁶ This paper focuses on one such species, Japanese barberry (*Berberis thunbergii*), and explores its relationship to Lyme disease prevalence in New Jersey to better understand how environmental and public health concerns intersect. These facts make it clear that invasive species are a serious issue, and understanding their spread and consequences is crucial for protecting biodiversity, economic well-being, and public health.

Basic Description of Barberry:

One particularly harmful invasive plant in the U.S. is Japanese barberry (*Berberis thunbergii*), a deciduous shrub native to Japan and eastern Asia.⁷ The plant is easily recognizable by its red berries and ability to form dense thickets that crowd out native species and alter ecosystems.⁸ It was first introduced to the United States in the late 1800s, when seeds were sent from Russia to Boston for ornamental purposes.⁷ It later became popular in landscaping, and its seeds were also dispersed by birds, allowing the plant to spread rapidly.⁹ Japanese barberry thrives in a wide range of conditions, including sun, shade, drought, and warmer temperatures.¹⁰ It's also unaffected by other species, including deer, which tend to avoid it and eat only native plants.¹¹ As a result of its persistence and damage to the environment, it is now banned in several northeastern states, including Maine, Massachusetts, New York, and Pennsylvania.¹²

Ecological Impacts:

Ecologically, Japanese barberry creates serious damage. Its ability to spread deep into forest interiors means it replac-

es native plants that would otherwise feed and support local wildlife.¹³ As barberry spreads, native plant populations decline, leading to fewer insects, fewer food sources for animals, and a breakdown of entire food webs.^{13,14} In particular, deer contribute to the spread of barberry as well, since they only eat native plants; they end up over-browsing those while completely ignoring barberry.¹⁵ This type of trophic downgrading reduces biodiversity and makes ecosystems more fragile.¹⁶ Japanese barberry has also been found to affect soil chemistry, changing nutrient cycling and further hindering the growth of native plants.¹³ These changes don't stay localized—they cascade through the entire ecosystem, affecting everything from microbes to large mammals.

Control:

Controlling Japanese barberry is possible but difficult.¹⁷ Physical methods like pulling or digging up the roots can work, especially in cases of small infestations, but are labor-intensive and time-consuming.¹⁸ Herbicides offer a chemical solution, but they come with environmental concerns such as soil contamination, damage to non-target plants, and the plants could develop resistance to the herbicide over time.¹⁹ Prevention can be more effective than removal, which is why planting native alternatives, like the Red chokeberry and Dwarf Fothergilla, is recommended by experts.²⁰ Increasing public awareness and legislation is also an effective control measure, especially in states like New Jersey that don't currently ban its sale.^{6,21} To sum up, barberry control is possible through commitment and collaboration, which is not happening in most parts of New Jersey.²²

Lyme Disease:

Japanese barberry is linked to public health concerns related to Lyme disease, which makes control even more critical.²³ Barberry creates a humid, protected environment that is ideal for black-legged ticks (*Ixodes scapularis*) and is unique to barberry, as other native plants don't provide the same habitat.⁷ These ticks carry *Borrelia burgdorferi*, the bacteria that cause Lyme disease.²⁴ Lyme disease was first recognized in 1975 in Lyme, Connecticut, after many children were diagnosed with rheumatoid arthritis.²⁴ Today, almost half a million Americans get Lyme disease each year.²⁵ Symptoms include fever, fatigue, joint pain, and the hallmark bullseye rash at the location of the bite, though not all patients develop the rash, which makes diagnosis more difficult.²⁶ Lyme disease can usually be treated early with antibiotics, with patients recovering in a few weeks. However, if left untreated, it can lead to long-term neurological and heart problems such as facial palsy, meningitis, myocarditis, and more.²⁶ Japanese barberry fostering the growth of Lyme disease highlights the urgent need to consider public health when managing invasive species.

Purpose of this Study:

The first goal of this study was to establish the geographical correlation between Japanese barberry occurrence reports, blacklegged tick populations, and Lyme disease cases across the United States. Understanding this US-wide trend, I next

investigated further the connection between Japanese barberry (*Berberis thunbergii*) and Lyme disease, specifically in New Jersey. Finally, this research explores whether removing or restricting barberry might be linked to a change in Lyme disease incidence by analyzing the timing of statewide bans or regulations.

■ Methods

Data Collection: Japanese barberry USA Abundance Dataset:

Japanese barberry distribution data from EDDMapS was used, which provided the number of plant sites per state.²⁷ The data were downloaded as a CSV file and were further processed on Datawrapper, a map-making website. When the Datawrapper map was created, the color assignment was adjusted to include values from 0 to 5,000 for Barberry abundance to best represent differences in the plants' presence across states. Further, natural interpolation for color was used because it followed the natural distribution of the data without creating artificial divisions, unlike deciles or quartiles, which distorted the results. The FOHVOS dataset consists of recorded occurrences of invasive plant species, where each entry represents a confirmed observation of Japanese barberry at a specific location. Each record includes a geographic location (county), observation status (Detected or Positive), and a reporting source. It does not include tick population density. Because the dataset is based on reported observations rather than systematic sampling, it serves as an estimate for relative presence rather than true population density.

Occurrence records were filtered for Japanese Barberry and organized by county. The total number of reports per county was used to estimate relative barberry presence. Higher report counts were interpreted as a greater presence of Japanese barberry.

Data Collection: Lyme Disease Prevalence and Tick Datasets:

For Lyme disease prevalence, the state and locality-level surveillance data were downloaded from the Centers for Disease Control and Prevention (CDC) as an XLS file.²⁸ All cases from the last 20 years for each state were aggregated in Excel to obtain a cumulative picture of Lyme distribution. This dataset was then imported into Datawrapper, where natural interpolation for color assignment was again used, this time with a range from 0 to 130,000 cases, to show which states had the fewest cases (light green) and which had the most (dark blue). This color scale allowed direct comparisons to the barberry map.

The black-legged tick distribution data were used to make the tick table by inserting a pivot table in Excel. The pivot table used state for rows and state for columns, with the count of status per state as the values. Another column was created for the percentage of counties with an established tick population, which was calculated as the number of established counties divided by the total number of counties with data. It should be noted that Alaska and Hawaii did not have any data, and counties that had "no records" do not necessarily mean that there are no ticks, but rather that they have not been reported.

Data Collection and Analysis: New Jersey County Correlation:

First, an interactive species distribution map in NJ provided by FOHVOS (Friends of Hopewell Valley Open Space) was used.²⁹ On the webpage with the map, "All" was selected for county, property, and eradication status, Japanese barberry was chosen for species, and the data was then exported as an XLS file. This data was used for the analysis of New Jersey counties below. Separately, Excel was used to create graphs showing the total case numbers over time for each New Jersey county from 2000 to 2022, as well as the percentage of the population affected in each county. During the chart design phase, data were selectively included or not included to ensure clarity in the presentation of the time-series and comparative incidence data.

Data Collection: Statewide Policy Reference Table:

Official state invasive species lists, agriculture department pages, and university extension websites were used to find the policies implemented in states that had implemented a statewide ban or regulation on Japanese barberry. In particular, Connecticut, Maine, Delaware, Massachusetts, New York, New Hampshire, Vermont, and West Virginia have banned Japanese barberry, while Wisconsin and Minnesota have regulations. The timing of those bans or regulations was then recorded and compiled in the table below to facilitate comparison with Lyme disease trends.

■ Results and Discussion

Geographical Correlations in Distributions of Lyme Disease, Blacklegged Ticks, and Japanese Barberry Across the US:

The distribution maps reveal a clear geographic correlation: areas with higher prevalence of Japanese barberry also tend to align with higher populations of blacklegged ticks, which in turn parallels reported Lyme disease cases. In Figure 1A, it is shown that Lyme disease cases per capita are the most concentrated in the Northeast. There are also some other states not in the Northeast that have slightly higher cases per capita, such as Wisconsin and Oregon. Figure 1B shows the percentage of counties in each state that have an established tick population, which is prevalent mostly in the eastern part of the United States (there's no data for Alaska and Hawaii). All states have data for Lyme disease, but not for Japanese barberry (Figure 1C); this may be because Japanese barberry isn't as much of a problem in the West, so there's less research and documentation of the plant in that region. The same correlation with Figure 1C is with Figure 1A; there's a relatively high concentration of Japanese barberry plant sites and Lyme disease cases.

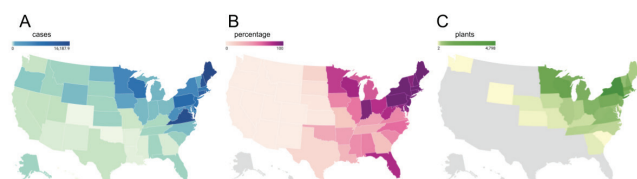


Figure 1: Lyme disease cases per capita- cases per 100,000 people (A), percentage of counties per state that have an established tick population (B), and number of Japanese barberry plant reports per state (C).

A distinct geographic overlap exists between the three variables, particularly clustered in the northeastern quadrant of the United States.

Hunterdon County had the third-highest number of reports of Japanese Barberry, with 97 reports (groupings of Japanese Barberry plants), and it had the highest amount of Lyme disease cases per capita (457 per 100,000 people). Meanwhile, Bergen County had one of the smallest numbers of reports, with only 2 reports, and it had the third-lowest amount of Lyme disease cases per capita, with 42 cases per 100,000 people. Interestingly, Essex County had the lowest number of Lyme disease cases per capita (18) and was the only county to initiate plans to mitigate the spread of Lyme disease. This overlap supports a geographic association between reported barberry presence and Lyme disease risk."

Although it's important to note that correlation does not equal causation, and there may be other factors influencing Lyme disease cases - such as deer populations, which also carry blacklegged ticks - the observed overlap supports existing research on habitat-level ecology. Dense barberry thickets create humid microhabitats ideal for tick survival and reproduction. States shaded gray on the Japanese barberry and blacklegged tick maps represent either regions with no presence of the species or insufficient data. There is a clear east-west divide in prevalence, with higher barberry and Lyme disease levels concentrated in the Northeast and Midwest. This could be explained by poleward expansion due to climate change.³⁰ Since Japanese barberry (and ticks) thrive in warmer conditions, northern states that are now hotter for longer periods of the year can sustain more growth, contributing to the rise in Lyme disease cases. Taken together, these patterns show that the spread of Japanese barberry indirectly influences the risk of Lyme disease by fostering favorable conditions for blacklegged tick populations.

Analysis of Lyme Disease Charts:

Over the past two decades, Lyme disease cases have steadily increased, with a particularly sharp rise in the last five years, reaching 70.7 cases per 100,000 people in New Jersey.³¹ Although there was a dip in 2020 due to COVID-19, when people were less likely to be outdoors and exposed to ticks, the trend has since rebounded and continued upward. Ten counties account for nearly 80% of Lyme disease cases, with Morris County (12.6%), Hunterdon County (9.6%), and Monmouth County (9.3%) being the highest contributors.

The specific counties in Figure 2A were chosen to show a variety of different counties, showing the counties that had the least amount and the highest amount of Lyme disease cases, had a plan for controlling Japanese barberry, and had the most amount of Lyme disease cases. It is important to note that in Figure 2B, the percentages add up to 100%, which shows how much each county contributes to the total Lyme disease cases in New Jersey, but is not weighted, unlike in the first graph.

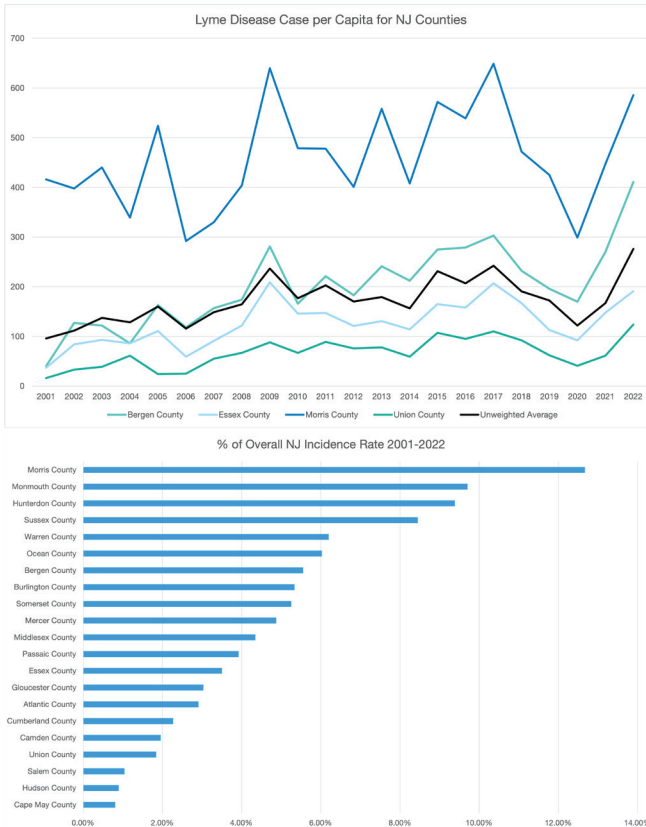


Figure 2: Lyme disease cases per 100,000 people from 2005-2022 in select NJ counties (A) and the percentage of Lyme disease cases in NJ that came from each county from 2001-2022 (B).

The data show that Lyme disease is heavily concentrated in specific areas, led by Hunterdon County's high per capita rates and Morris County's top share of overall state cases. These sharp geographic disparities highlight the specific high-risk forest and barberry habitats driving New Jersey's incidence.

Data from FOHVOS shows that Morris County has the highest incidence of Japanese barberry, which also has the highest net number of Lyme disease cases, as shown in Figure 2B (12.6%). This clearly demonstrates how higher Japanese barberry abundance contributes to more Lyme disease cases. Morris County is very wooded, and Japanese barberry is very widespread in the county. Overall, counties with more forested land and dense vegetation showed higher Lyme disease rates, reinforcing the connection between habitat characteristics and disease prevalence. These geographic and ecological patterns provide a strong foundation for analyzing how management efforts and environmental changes may influence Lyme risk, which will be explored further in the discussion section. Hunterdon County is an example of this, as the population is relatively small, with about 130,000 people; however, it has 280,000 acres of forest and, as a result, has the highest number of Lyme disease cases per capita. This same pattern matches up with the rest of the United States; for example, Maine has the highest percentage of forest area, with almost 90%, and also has the highest number of Lyme disease cases per capita at 16187.90. Wisconsin and New Jersey have the same percentage of forest area (42%) and have about the same number of Lyme disease cases per capita (around 8,300 cases); however,

Wisconsin has a large deer population, with it being in the top 5 states in the US, which also spreads ticks.³² This helped me explore whether removing or restricting barberry might be linked to a change in Lyme disease incidence, while keeping in mind that many other factors, like the climate and tick populations, can influence Lyme risk.

Analysis of Statewide Policies on Japanese Barberry and Lyme Disease Cases:

Several northeastern states have implemented policies to regulate or prohibit Japanese barberry, while others have limited measures or pending legislation (Table 1). In New Hampshire, the state's invasive species regulations (Agr 3800) prohibit the sale, transport, propagation, or distribution of species listed on its Prohibited Invasive Plant List, which includes Japanese barberry (New Hampshire Department of Agriculture, Markets & Food). In New York, 6 NYCRR Part 575, enacted in March 2015, made it illegal to possess, import, sell, or introduce Japanese barberry, with a one-year grace period allowing the sale of existing nursery stock (New York Department of Environmental Conservation; Cary Institute of Ecosystem Studies). By contrast, New Jersey currently has no statewide ban or comprehensive invasive species law regulating barberry. The proposed Invasive Species Management Act (S1029), introduced in 2025, would prohibit its sale, distribution, and propagation without a permit and establish a state Invasive Species Council (New Jersey Conservation Foundation). Although it's not a law yet, barberry appears on New Jersey's "Do Not Plant" lists, such as those published by the New Jersey Invasive Species Strike Team (FOHVOS).

Table 1: States with regulations or bans on Japanese barberry. Legislative bans and restrictions on Japanese barberry have steadily expanded across the Northeast and Midwest over the past two decades. However, because these policies primarily target future nursery sales, they have a limited immediate impact on reducing local Lyme disease rates without active removal of already established wild populations.

State	Year Regulation Measure Introduced	Measure
Connecticut	Public Act 24-11 (2024) adds species; CT Invasive Plant Council notes barberry to be prohibited from sale by Oct 1, 2028, after phase-out	BANNED (still phasing in) ³³
Delaware	July 1, 2022	BANNED (Invasive Plant List, nursery/retail sales prohibited) ³⁴
Maine	January 14, 2017	BANNED (DACF "Do Not Sell" list) ³⁵
Massachusetts	January 1, 2009, phase out completed 2009-2012	BANNED (MDAR Prohibited Plant List) ³⁶
New Hampshire	January 1, 2007	BANNED (Agr 3800 Prohibited Invasive Plants) ³⁷
New York	March 10, 2015, later allowed limited sterile cultivars via permits/research	BANNED (6 NYCRR Part 575 "Prohibited Invasive Species") ³⁸
Pennsylvania	October 8, 2021	BANNED (Noxious Weed Control - List B) ³⁹
Vermont	In 2008, on the current prohibited list	REGULATED (On Vermont's Noxious Weed Quarantine) ⁴⁰
West Virginia	July 1, 2020	BANNED (state noxious/invasive list - sale prohibited) ⁴¹
Wisconsin	May 1, 2015, sales and distribution are controlled	REGULATED (NR 40 Invasive Species Rule) ⁴²
Minnesota	January 1, 2015, sale and distribution are prohibited, though cultivars are legal.	REGULATED ("Specially Regulated" terrestrial invasive - sterile cultivars may be permitted) ⁴³

The comparison of state-level policies with Lyme disease trends suggests that while implementing bans and regulations

is a necessary first step, their immediate impact on disease incidence may be limited due to the large existing presence of Japanese barberry. Many states, especially in the Northeast, that have banned or regulated the plant (such as Massachusetts and New Hampshire) still report some of the highest case numbers in the country. For example, New York implemented its ban in 2015, and while it had a one-year grace period, cases have generally not gotten significantly worse or better. The CDC's revised surveillance case definition in 2022, which led to a large increase in reported cases, also complicates pre- and post-ban comparisons.⁴⁴ Nevertheless, policies like those in New York, which allow limited sterile cultivars via permits, show a proactive approach to managing the introduction of new barberry plants. Additionally, in January 2026, New Jersey enacted the Invasive Species Management Act, which designates Japanese barberry as a prohibited invasive species and establishes a statewide framework for its regulation. Because the law's restrictions are being phased in, its effects on barberry prevalence and Lyme disease incidence cannot yet be evaluated.

The primary challenge remains the removal and control of established barberry thickets across vast areas, a slow process that requires sustained public awareness and commitment, especially in states like New Jersey, where there are currently no statewide bans. While banning invasive plants is helpful, it doesn't reduce the amount already present, and they will continue to reproduce without active management. New Jersey provides a clear example: every county besides Essex lacks control plans for barberry, even though it's widespread across the state. States should not only ban invasive species but also make it publicly aware of the dangers of planting them, as many people do not even know what invasive plants are, and that plants in their garden may even be banned in their state. For example, the public-led campaign to squash invasive lanternflies in New Jersey shows how widespread awareness, even through small actions, can help control ecological threats.

■ Conclusion

Japanese barberry is more than just an invasive plant—it is an environmental and public health threat that contributes to higher Lyme disease risk. I explored three main questions: how barberry distribution correlates with blacklegged tick populations and Lyme disease cases, whether this connection is visible in New Jersey, and how state policies impact the problem. My findings revealed clear geographic overlap between barberry abundance, tick presence, and Lyme disease cases, especially in forested regions like northern New Jersey. States with bans on barberry still report high disease rates, showing that regulation alone isn't enough without long-term removal and prevention efforts. These results highlight the urgent need for stronger management strategies and public awareness to limit both ecological and health impacts of invasive species like Japanese barberry. States with high invasive species need a multi-prong approach to minimize the impact on public health. A grass-roots volunteer effort could play an important role in helping to eradicate these plants in forested areas. Community members may also consider writing to local representatives. New

Jersey could benefit from passing the pending Invasive Species Management Act. In addition, detecting and eradicating the shrubs from growing in the wild is an important strategy that could be achieved through partnerships among local communities, foundations, environmental agencies, and forest land conservation departments. These results reinforce previous ecological findings that invasive plant management is directly tied to public health outcomes, highlighting the urgent need for stronger management strategies and public awareness to limit both ecological and health impacts of invasive species like Japanese barberry. Future studies should include other invasive vegetative species like the herbaceous plants and perennials (e.g., giant hogweed), and vines (e.g., English Ivy).

■ Acknowledgments

I would like to thank my research teacher at school, Luis Largo, for his continued guidance and support throughout the early stages of this research. I would also like to thank Candide Simard, who immensely helped me to get my research to the publishing stage.

■ References

1. *Japanese Barberry Management*; Rutgers New Jersey Agricultural Experiment Station, 2018. <https://njaes.rutgers.edu/fs1353/> (accessed 2024-11-15).
2. Early, R.; Bradley, B. A.; Dukes, J. S.; Lawler, J. J.; Olden, J. D.; Blumenthal, D. M.; Gonzalez, P.; Grosholz, E. D.; Ibañez, I.; Miller, L. P.; Sorte, C. J. B.; Tatem, A. J. Global threats from invasive alien species under climate change. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (1), 1–6. DOI: 10.1073/pnas.1514456113
3. *New Jersey Strategic Management Plan for Invasive Species*; New Jersey Department of Environmental Protection, 2009. https://dep.nj.gov/wp-content/uploads/invasive-species/nj-strategic-management-plan-for-invasive-species_final-2009.pdf (accessed 2024-11-15).
4. Pimentel, D.; Zuniga, R.; Morrison, D. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecol. Econ.* **2005**, *52* (3), 273–288. DOI: 10.1016/j.ecolecon.2004.10.002
5. *Playing Smart: Invasive Species Management*; United States Forest Service. <https://www.fs.usda.gov/managing-land/invasive-species-management/playing-smart> (accessed 2024-11-15).
6. *Native Plants and Invasive Species*; New Jersey Conservation Foundation. <https://www.njconservation.org/native-gardens/> (accessed 2024-11-15).
7. *Japanese Barberry in Acadia National Park*; National Park Service. <https://www.nps.gov/articles/000/japanese-barberry-acadia.htm> (accessed 2024-11-15).
8. *The Invasive Japanese Barberry*; Pennsylvania State University Extension. <https://extension.psu.edu/the-invasive-japanese-barberry> (accessed 2024-11-15).
9. Silander, J. A.; Klepeis, D. M. The invasion ecology of Japanese barberry (*Berberis thunbergii*) in the New England landscape. *Biol. Invasions* **1999**, *1*, 189–201. <https://doi.org/10.1023/A:1010024202294>.
10. Ehrenfeld, J. G. Effects of exotic plant invasions on soil nutrient cycling processes. *Ecosystems* **2003**, *6* (6), 503–523. <https://doi.org/10.1007/s10021-002-0151-3>
11. *Japanese Barberry: Not an Average Landscape Shrub*; New York State Parks, October 18, 2016. <https://nystateparks.blog/2016/10/18/japanese-barberry-not-an-average-landscape-shrub/> (accessed 2024-11-15).

12. Williams, S. C.; Ward, J. S.; Worthley, T. E.; Stafford, K. C. Managing Japanese barberry infestations reduces blacklegged tick abundance and infection prevalence with *Borrelia burgdorferi*. *Environ. Entomol.* **2009**, *38* (4), 977–984. <https://doi.org/10.1603/022.038.0404>
13. Ward, J. S.; Williams, S. C. Low recruitment of native trees associated with Japanese barberry invasion. *Invasive Plant Sci. Manag.* **2019**, *12* (3), 189–199. DOI: 10.1017/inp.2019.22
14. *Species Spotlight: Japanese Barberry*; National Park Service. <https://www.nps.gov/articles/species-spotlight-japanese-barberry.htm> (accessed 2024-11-15).
15. *Invasives in Your Woodland: Japanese Barberry*; University of Maryland Extension. <https://extension.umd.edu/resource/invasives-your-woodland-japanese-barberry/> (accessed 2025-11-15).
16. *Impact of Invasive Japanese Barberry Cascades Through Local Food Webs*; Entomology Today, September 13, 2019. <https://entomologytoday.org/2019/09/13/impact-invasive-japanese-barberry-cascades-through-local-food-webs-arthropods-insects/> (accessed 2024-12-15).
17. Ward, J. S.; Williams, S. C. Controlling an invasive shrub, Japanese barberry (*Berberis thunbergii* DC), using directed heating with propane torches. *Nat. Areas J.* **2011**, *31* (2), 156–161. <https://doi.org/10.3375/043.031.0208>
18. *Weed Control Methods Handbook: Manual and Mechanical Control Methods*; Invasive.org. <https://www.invasive.org/gist/products/handbook/03.manualmechanical.pdf> (accessed 2024-12-15).
19. *Forest Herbicide Safety and Environmental Concerns*; Mississippi State University Extension. <https://extension.msstate.edu/publications/forest-herbicide-safety-environmental-concerns-and-proper-handling> (accessed 2024-12-15).
20. *Invasive Japanese Barberry*; University of Illinois Extension. <https://extension.illinois.edu/invasives/invasive-japanese-barberry> (accessed 2025-11-01).
21. *Pennsylvania Considers Banning Japanese Barberry*; WHYY. <https://whyy.org/articles/pa-considers-banning-japanese-barberry-a-popular-but-invasive-landscaping-plant/> (accessed 2024-12-15).
22. *Bill Banning Invasive Species*; Sea Girt Conservancy. <https://seagirtconservancy.org/bill-banning-invasive-species/> (accessed 2025-02-01).
23. *Japanese Barberry and Public Health Impacts*; Minnesota Department of Natural Resources. <https://www.dnr.state.mn.us/invasives/terrestrialplants/woody/japanesebarberry.html> (accessed 2025-02-01).
24. Stanek, G.; Wormser, G. P.; Gray, J.; Strle, F. Lyme borreliosis. *Lancet* **2012**, *379* (9814), 461–473. [https://doi.org/10.1016/S0140-6736\(11\)60103-7](https://doi.org/10.1016/S0140-6736(11)60103-7)
25. Lyme Disease Data and Statistics; Centers for Disease Control and Prevention. <https://www.cdc.gov/lyme/data-research/facts-stats/index.html> (accessed 2025-02-01).
26. *Ticks and Lyme Disease*; Johns Hopkins Medicine. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/lyme-disease/ticks-and-lyme-disease> (accessed 2025-02-01).
27. *Early Detection and Distribution Mapping System (EDDMapS)*; University of Georgia, 2025. <http://www.eddmaps.org/> (accessed 2025-02-01).
28. *Lyme Disease Surveillance Data*; Centers for Disease Control and Prevention. <https://www.cdc.gov/lyme/data-research/facts-stats/surveillance-data-1.html> (accessed 2025-02-01).
29. *Invasive Species Interactive Map*; Friends of Hopewell Valley Open Space. <https://www.fohvos.info/invasive-species-strike-team/interactive-map/> (accessed 2025-02-01).
30. Sundberg, M. *Proactive Weeding of Our Gardens of the Future*; EcoLandscaping, May 22, 2023. <https://www.ecolandscaping.org/05/developing-healthy-landscapes/climate-change/proactive-weeding-of-our-gardens-of-the-future/> (accessed 2025-11-01).
31. *Lyme Disease Indicator Report*; New Jersey Department of Health. <https://www-doh.nj.gov/doh-shad/indicator/complete/LymeDisease.html> (accessed 2025-02-01).
32. Kilpatrick, H. J.; Labonte, A. M.; Stafford, K. C. The relationship between deer density, tick abundance, and human cases of Lyme disease in a residential community. *J. Med. Entomol.* **2014**, *51* (4), 777–784. <https://doi.org/10.1603/ME13232>
33. *Good News on 2025 Invasive Plant Legislation*; Connecticut Invasive Plants Working Group, June 24, 2025. <https://cipwg.uconn.edu/2025/06/24/good-news-on-2025-invasive-plant-legislation/> (accessed 2025-08-14).
34. *Invasive Plants Law*; Delaware Department of Agriculture. https://agriculture.delaware.gov/wp-content/uploads/sites/108/2023/05/DNSC-IPL_GardenCenter-Fact-Sheet.pdf (accessed 2025-08-14).
35. *Maine Strengthens Invasive Plant Regulations*; Maine Department of Agriculture, Conservation and Forestry. <https://www.maine.gov/dacf/about/news/news.shtml?id=729631> (accessed 2025-08-14).
36. *Prohibited Plant List Background*; Massachusetts Executive Office of Energy and Environmental Affairs. <https://www.mass.gov/info-details/prohibited-plant-list-background> (accessed 2025-08-14).
37. *New Hampshire Invasive Species Factsheets*; New Hampshire Department of Agriculture, Markets & Food. https://takeaim.org/wp-content/uploads/2016/11/New_Hampshire_invasive_species_factsheets.pdf (accessed 2025-08-14).
38. *Prohibited and Regulated Invasive Species*; New York Department of Environmental Conservation. https://extapps.dec.ny.gov/docs/lands_forests_pdf/isprohibitedplants2.pdf (accessed 2025-08-14).
39. *Pennsylvania Phasing in Ban of Invasive Japanese Barberry*; Pennsylvania Department of Agriculture. <https://plant-pest-advisory.rutgers.edu/pennsylvania-phasing-in-ban-of-invasive-japanese-barberry-official-pa-dept-ag-press-release/> (accessed 2025-08-14).
40. *Noxious Weeds in Vermont*; Vermont Agency of Agriculture, Food and Markets. <https://agriculture.vermont.gov/public-health-agricultural-resource-management-division/plant-health-and-pest-management/plant-2> (accessed 2025-08-14).
41. *Sale of Japanese Barberry Outlawed in West Virginia Come July 1*; West Virginia Department of Agriculture. <https://woay.com/sale-of-japanese-barberry-outlawed-in-west-virginia-come-july-1/> (accessed 2025-08-14).
42. *Japanese Barberry (Berberis thunbergii)*; Wisconsin Department of Natural Resources. <https://dnr.wisconsin.gov/topic/Invasives/fact/JapaneseBarberry> (accessed 2025-08-14).
43. *Japanese Barberry*; Minnesota Department of Agriculture. <https://www.mda.state.mn.us/plants/pestmanagement/weedcontrol/noxiouslist/japanesebarberry> (accessed 2025-08-14).
44. *Revised Surveillance Case Definition for Lyme Disease*; MMWR **2024**, *73*, 1–6; Centers for Disease Control and Prevention. (accessed 2025-08-14).

■ Author

Kavya Ballal is a high school student at Kent Place School with academic interests in biology, public health, and healthcare. She has earned Honorable Mention in the USA Biology Olympiad, placing her among the nation's top young biology scholars. Kavya is also a concert pianist and has interests in writing and music. Kavya plans to pursue an MD, and she is particularly interested in studying the intersection of disease, technology, and public policy.

Hierarchical Communication Structures for Collective Decision Making in Robot Swarms

Kevin Wu

John Paul College, Whitworth Rd, Rotorua, Bay of Plenty, 3010, New Zealand; kevwu66@gmail.com

Mentor: Dr. Nisha Rawindaran

ABSTRACT: Swarm robots have potential in fields like search and rescue, space exploration, agriculture, and many other fields that are dangerous to humans or require large areas to be covered. The ability for a swarm to make decisions autonomously without a central controller plays a strong part in making them viable to implement in real-world situations. This research explores the potential for a hierarchical control method to be applied to the decision-making process by comparing it to a benchmark of a fully decentralized voter model-based method. This has been shown to improve performance in other areas of managing swarm robot behavior, which motivates this study into its application in collective decision-making. Using the Webots simulator, both of these methods were implemented using E-puck robots and compared to each other under the same conditions. The results showed that the hierarchical control method did not improve decision-making speed in the swarm measurably, but may have a beneficial effect on accuracy, which disproves the hypothesis of this research. The small scale of the simulation due to limitations encountered does, however, increase the effect of random noise on the results, which leaves the problem open to further investigation. Such further research for this could include the use of larger swarm sizes, more optimized behavioral algorithms, and/or real-world simulations.

KEYWORDS: Robotics and Intelligent Machines, Control Theory, Swarm Robotics, Collective Decision Making, Collective Perception, Hierarchies.

■ Introduction

Swarm robotics is the area of research into how large numbers of robots can cooperate to autonomously solve problems and complete tasks without a central controller. The difficulty is that without overall knowledge of the swarm, each robot acts based only on interactions with its neighbors and the environment. The idea is that from a swarm of many simple robots, complex behavior and abilities can emerge. This draws inspiration from existing structures in nature, such as social insects like bees,¹ ants, schools of fish, and bird flocking. This leads to systems that are more adaptive and robust than single robots,² as they can split tasks, replace damaged members, and add or lose members.

Challenges faced regarding the control of robot swarms include formation control, navigation, and the problem this research focuses on: decision-making, in this case, choosing one option out of multiple. In swarm robotics, this is known as the best-of-n problem, deciding the best option out of n possible ones. Typically, this means the swarm needs to reach either full consensus or have a majority agree on one option, at a threshold decided by the designer. A well-known example of this is the issue of nest-selection, seen in honeybees, which has given inspiration to models of the problem in robot swarms.³ This problem involves deciding the best location for a colony to move to out of multiple options, which all have traits that decide their quality, such as location, size, vulnerability, etc. Similarly, robots may need to decide the best location to gather resources in space or prioritize areas to search for injured people in emergencies.

Another concept being researched in swarm robotics is that of hierarchical control, a hybrid between centralized and decentralized control. The weakness of having a central controller is that there is a single point of failure because the single entity controlling the entire swarm could be damaged or compromised somehow. Reliance on decisions from the central controller also leads to reduced autonomy and adaptability to the situation, which is valuable in applications such as space exploration. The decentralized approach in swarm robotics removes these issues at the cost of being much more complex to engineer due to the fact that overall swarm behavior is emergent. This means that said behavior cannot be directly dictated, but instead arises from the interactions between individual robots. Solutions are being created to this problem, such as the use of neural networks,^{4,5} or combining pre-designed behaviors that are then fine-tuned.⁶ Hierarchical control provides a compromise where some robots in the swarm hold higher-level knowledge or abilities and have control over the lower-level members of the swarm, which means overall swarm behavior can be more easily controlled, which can also increase performance. These robots are part of the swarm and can be replaced by any other robot in it, removing the vulnerability to losing them. However, there may be potential trade-offs in that the selection of these higher-level robots adds complexity to an individual robot's behavior, which might take memory or processing power, which is valuable in swarms of smaller robots. In systems where the hierarchy cannot be reconfigured, higher-level robots would still be points of failure in the system, which negates the benefit of decentralized behavior.

In this research, the primary contribution made to the field is the proposal that a hierarchical control system, which is shown to have improvements in other areas regarding the control of robot swarms, can be applied to collective decision-making as well. This could potentially increase speed, accuracy, or both. Literature has been gathered to support this proposal, and currently, the literature lacks research exploring the connection between hierarchical control and decision-making, which motivates the work done in the study here. It will be explored by simulating a collective perception problem, where a swarm of robots must decide on a dominant feature of the environment. If applying hierarchical control to collective decision-making proves effective, the added efficiency in decision-making would allow complex tasks to be solved faster by robot swarms, as well as making the design of decision-making behavior simpler by allowing us to avoid the difficulty of engineering emergent behavior.

Aims and Objectives:

The objective of this research is to prove or disprove the hypothesis that a hierarchical decision-making strategy will perform better than the fully decentralized voter model in a collective perception task, in terms of accuracy and decision-making time.

— H_0 : The hierarchical decision-making method has no impact on decision-making speed/accuracy.

— H_1 : The hierarchical decision-making method has an impact on decision-making speed/accuracy.

To achieve this, we look into the best-of-two problem where the swarm decides between two options through experimental simulations. The robots, simulated in the Webots simulator,⁷ will navigate a floor of black and white tiles and must reach consensus on which color makes up the majority of the tiles. It applies the concept of hierarchical control structures to this with the aim of improving performance in swarm decision-making. Based on the literature, it is hypothesized that the application of a hierarchical control structure will improve the performance in decision-making of a swarm of robots as compared to a fully decentralized model.

The performance of the strategies will be measured using the specific metrics of:

— time to reach consensus in the swarm

— proportion of robots reaching the correct conclusion.

The constraints of this research include:

— Resources - the work required of this research, such as simulation design, implementation, and data analysis, falls within the ability of the researcher and the available resources.

— Relevance - This work fills in the observed gap in current swarm robotics research between hierarchical control methods and decision-making.

The specific contribution that this research aims to make to the field is to find whether it is possible to reduce the trade-off of speed and accuracy in collective decision-making in robot swarms by using the hierarchical decision-making method to increase both. This would allow both aspects to be improved and have neither one be sacrificed much.

■ Literature Review

This section will present work done in both collective decision-making and hierarchical control structures, showing some different proposed methods and related findings that inspire this paper. One prominent method of collective decision-making is known as the voter model, where a robot switches to the opinion of a randomly selected neighbor. Valentini *et al.* introduced the weighted voter model,⁸ inspired by the decision-making behavior of honeybees, where the group's opinion converges on the option of the highest quality by modulating the time of a robot's opinion dissemination based on its assessment of quality using environment-sourced information. This method of decision-making in a robot swarm is referred to as direct modulation. They found that this method increases accuracy with system size and is robust to noise from the input of quality assessments.

Valentini *et al.* introduce another method utilizing direct modulation of the majority vote method of decision-making,⁹ where a robot switches its opinion by taking that held by the majority of multiple neighboring robots. They find that the number of neighboring robots affects decision-making, showing a trade-off where lower numbers have more accuracy but slower consensus times, and higher numbers are faster yet less accurate. The present research aims to reduce the trade-off between speed and accuracy through the introduction of the hierarchical communication structure, which may potentially increase one or both aspects without sacrificing the other.

Another method of decision-making, inspired by behavior in honeybees,¹ is cross-inhibition. In the cross-inhibition method, a robot samples the opinion of its neighbor, and if the opinion is different, it enters an uncommitted state where it holds no opinion. It then takes the opinion of a neighbor in the next sampling phase. Zakir *et al.* find that compared to the voter model, the cross-inhibition method is better at breaking deadlocks between options of similar quality due to its resistance to noise from self-sourced information.¹⁰ Resistance to noise is an important aspect of decision-making in order for the swarm to navigate real-life scenarios, due to the environment's imperfections and due to potential issues such as malfunctioning robots gathering incorrect sensor data or malicious robots inserted into the swarm. Antonic *et al.*¹¹ tested a swarm of heterogeneous robots - robots that used different methods of decision-making internally - and found that this swarm was more robust against zealots/stubborn robots that do not change their opinion. The robustness of a swarm's decision-making against noise, stubborn robots, or deadlocks using a hierarchical control structure is not examined in the present paper, but may be a direction to investigate in future research.

These methods of decision-making operate on the idea of minimizing computational demands, which is useful for robots with fewer resources, such as swarm robots that are produced to be simple but high in number. This is the area in which the research presented here focuses. However, some methods require higher resources and have achieved good results, some of which are known as Bayesian decision making.¹²⁻¹⁴ Bayesian decision making is the use of self-sourced and socially-sourced information to calculate a probability of one option being cor-

rect, upon which a decision is then made. Shan *et al.* compare their method of Bayesian decision making with direct comparison and the weighted voter model, finding that it outperforms both methods in speed, accuracy, and robustness at the cost of being more complex. As a clarification, direct comparison is a method of decision-making that involves the robots directly sharing opinion quality estimates rather than just the opinions, and updating their opinions based on what they receive.

One more area of research into controlling swarms is with the use of artificial intelligence, such as neural networks and reinforcement learning. Research into the application of AI is somewhat less relevant to our specific case, as it looks into human-designed behavioral systems due to the fact that the work primarily focuses on movement-based problems, such as formation control, whereas we focus on decision-making. Also, we are specifically dictating rules of behavior, while systems like neural networks control behavior through internal workings, gained through training, that are difficult to understand. It carries the benefits of being simpler to create, simply by training the artificial intelligence on a task, but has drawbacks such as struggling when presented with the differences between simulation and reality, as well as being a black box that we cannot reverse-engineer at higher levels of complexity. As some examples, Li *et al.* used a small bio-inspired neural network to create virtual forces for their fish-inspired collective escape behaviors.⁴ Kim *et al.* used neural networks to generate interaction rules that cause robots to create desired collective formation structures, and both of these studies found these methods to be effective.⁵

The present research focuses on the application of a hierarchical control structure to the process of collective decision-making. This concept is supported in the literature, as applying a hierarchical control structure has been shown to aid performance in other areas of swarm robotics. Varadharajan *et al.*¹⁵ found that using 'guide' robots that had somewhat higher abilities, such as spatial awareness, to oversee simpler 'worker' robots, a model radiation cleanup task could be completed with far fewer robots than using only the worker robots, while still retaining the scalability and lack of a single point of failure, which a centralized system does not have. Shankar *et al.* apply this hierarchical control to formation control and movement of a swarm,¹⁶ presenting a method of controlling a swarm's shape and movement using guide robots to control worker robots, allowing easier design of swarm behavior while again avoiding the pitfalls of centralized control. Robots do not have to be physically superior; however, they may simply have a level of control over surrounding robots, using a higher level of information and/or intelligence to guide the others. This means that any robot in the swarm could become and act as a leader over other robots. One applicable case would be that if an existing leader gets damaged, the swarm can reorganize and select a new one. Some papers present methods that allow swarms to reconfigure hierarchies freely. For example, Zhu *et al.* describe a system where robots running the same algorithm establish a hierarchy by attempting to recruit each other and reaching an agreement where one surrenders control to another in a level above it, leading to a "brain" robot at

the top.¹⁷ Zhang *et al.* provide rigorous, mathematically analyzable algorithms for formation control using hierarchical frameworks, focusing on merging and splitting of frameworks and robot departure.¹⁸ These approaches form hierarchies that can still be changed according to the task, eliminating the issue that there still exist certain points of failure when having fixed leader robots, and allowing greater adaptability to the situation. These works show promise for applying a hierarchical control structure to collective decision making and the best-of-n problem. The present literature review found a lack of research into this specific concept, providing the inspiration for the research presented here.

Trabattoni *et al.*¹⁹ make some progress in filling the gap in the literature through their work, where they present a method for the best-of-2 problem. They describe a group of robots that acts as a central control entity that holds some overall swarm information. Any robot in this central control formation is replaceable with another member of the swarm, again giving advantages that a central controller provides without disadvantages such as having a single point of failure. It was found to also improve consensus time and accuracy. The work of Trabattoni *et al.* provides strong support that our research will find an improvement in decision-making using a hierarchical control method. The difference between the work of Trabattoni *et al.* and the present research is the problem being observed - Trabattoni *et al.* have investigated a site-selection problem where the robots must travel to a site and then convene in a central location to disseminate opinions, whereas we will investigate a collective perception problem where the robots will explore the environment and decide on the dominant feature, remaining dispersed through said environment. This different problem requires a different approach to the hierarchical control strategy.

It is reasonable to assume, and literature has shown, that there are differences in the results achieved by swarm simulations and actual robots.²⁰⁻²¹ Often, approximations of robot swarms, such as mean-field models, are used, but there are inaccuracies, such as the assumption that the swarm has opinions uniformly distributed throughout. To ensure a more accurate conclusion, it has been decided to simulate E-puck robots in the WeBots simulator,⁷ so that physical interactions are taken into consideration. The next section outlines the methodology used in the present research.

■ Methods

1. Basis of Simulation:

The intention was to create a simulation that would give results allowing a fair comparison between decentralized and hierarchical decision-making. The limits of what can be achieved consist of the available hardware - a Windows laptop with 16 GB RAM - to run the simulation, and time constraints in development.

The simulated task used for this research is inspired by the literature and primarily based on that of Zakir *et al.*¹⁰ They introduce the situation where there is a grid of floor tiles that can randomly be one of two colors, and the robot swarm must decide which color is most dominant. The method of

decentralized decision-making used in the simulation is their direct-switching model with modulated dissemination time based on opinion quality, also referred to in other literature as Direct Modulation of Voter-Based Decisions (DMVD). In the study done by Zakir *et al.*, they used DMVD as a baseline to compare their proposed implementation of cross-inhibition. In our simulation, an implementation of DMVD is also used as a baseline, but cross-inhibition is replaced with the newly proposed hierarchical control system.

To allow a fair comparison between the two methods of decision-making, some variables must be controlled. Each randomly generated grid will be explored by both types of decision-making, using the same number of robots and the same starting positions between the two tests for each board. Testing on multiple randomly generated grids will allow random error to be accounted for in the case that certain specific scenarios have one method perform better than the other.

2. Problem Definition:

The general task that this research focuses on about collective decision-making is that of collective perception. Collective perception refers to the idea that a swarm of robots with relatively weak capabilities can aggregate their observations to gain a more accurate picture of the environment as a whole. The decision-making aspect of this is that when robots gather contradictory information, they must ultimately decide that one or the other is the true observation. Gathering a better overall understanding of the environment aids in further decision-making, such as those regarding what actions should be performed afterward. So, finding methods to increase the efficiency of this process aids may increase performance in various tasks where robot swarms are deployed, hence the proposal of the hierarchical decision-making method for improving the process. The test of the effectiveness of a hierarchical decision-making method will be performed through a specific task designed with the parameters as described next.

Ten E-puck robots (Figure 1), each with a ground sensor module, are simulated in the Webots robot simulator.

They are randomly spread over an 8 by 8 randomly generated grid of black and white tiles, each tile being a square with a 10 cm side length. There are always more white tiles than black tiles, at varying ratios. The robots can send and receive signals from each other within a range of 30 cm. Using only communication with neighboring robots and the ground sensor to detect the color of tiles, they must reach consensus on the color that covers the largest area. Performance is measured by the time taken for all robots to reach consensus, and the proportion of robots reaching the correct conclusion.

The parameters of the simulation given are based on some trial and error to fit a few conditions: Firstly, the range of communication cannot extend to robots across the entire area, as this would defeat the purpose of creating methods to allow information to propagate across the swarm. Secondly, the total floor area and tile size are sized so that the robots will be able to cover the entire area and gather enough different samples while keeping multiple other robots within their communication ranges. This stops them from becoming isolated from

the swarm or ending up in feedback loops with only one other neighbor. This is constrained by the number of robots that can be simulated, which is limited to 10, as the hardware available struggled with physics calculations at swarm sizes beyond that. Changing the parameters given could be done intentionally to see the effectiveness of the hierarchical decision-making model when faced with more extreme situations, such as robots losing contact with the swarm, but this is beyond the scope of our study.

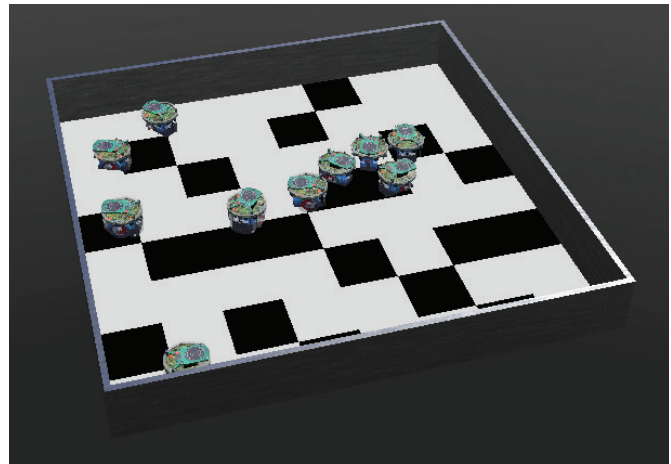


Figure 1: The setup of e-pucks in the Webots simulator.

3.1. Implemented Robot Behavior - Voter Model:

This section describes the implementation of the behavior of the simulated E-puck robots when testing the decision-making performance of the voter model.

Each E-puck robot has a random start position, orientation, and opinion on the grid. Their behavior is dictated as follows (broken down simply in Figure 2).

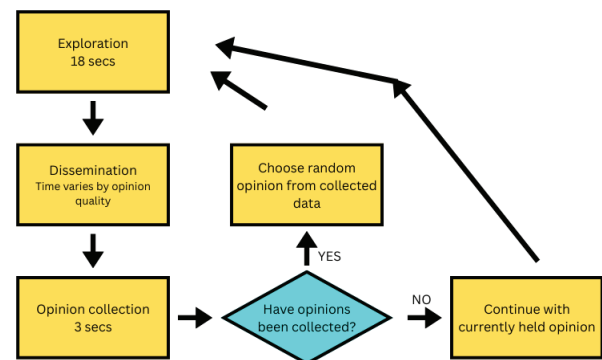


Figure 2: The basic robot behavior.

The robots move continuously using an implementation of Levy flight with basic obstacle avoidance, allowing a robot better coverage of the overall environment. The specifics as programmed in our simulation are as follows:

At any point:

If an obstacle is detected on the right, turn left

If an obstacle is detected on the left, turn right

Otherwise

Drive forward at a random speed for 3 seconds

Turn at a random speed for 1.5 seconds

Random distances and angles are achieved by changing the speed rather than the time of movement in order to reduce the random error in this simulation. In the exploration phase, a sample of the floor color is taken every time the robot begins driving forward, so randomly changing the time taken to move introduces error into the number of measurements a robot can take before the 18-second exploration phase ends. The amount of time used was chosen to allow the robots to move from one end of the arena to the other if the random speed happens to be maximum. Similarly, the turning time is selected to allow the robot to make a full rotation if chance dictates that it rotates at maximum speed. With every measurement of the floor color, the stored opinion quality - the proportion of samples taken which affirm the opinion - is recalculated based on the opinion that the robot currently holds. For example, if the robot's opinion is that the floor is predominantly white, the opinion quality is calculated as (# of white samples) / (# of samples taken)

After the exploration phase, the robot goes into the dissemination phase. Samples are no longer taken. The robot sends out its opinion to all neighboring robots within the specified range, in our case, 30 cm, every 0.1 seconds. Due to limitations in the abilities of the available hardware and the Webots simulator itself, the accuracy of timing does vary. The length of time this dissemination period lasts varies proportionally to the calculated opinion quality, with a minimum time of 2 seconds and a maximum of 10 seconds.

Once dissemination is finished, the e-puck then goes into the collection phase, where it receives opinions broadcast by neighboring robots. This period lasts for three seconds, after which a new opinion is randomly selected from the collected opinions and is then evaluated in the exploration phase again. If no opinions are collected after seven seconds, the robot proceeds with its previously held opinion.

The selection of the time for each phase is based on trial and error to create more realistic conditions, but it is not perfect and could be optimized in future work. The exploration phase timing of 18 seconds allows the robot to gather enough samples while not giving it full coverage of the area. The timing of the dissemination and collection phases is limited so that a robot will not receive or send information to every other robot on the floor within one cycle, while still collecting and disseminating enough information to impact the opinions of the swarm.

Once 9 of these loops pass with the opinion remaining the same, it is assumed that consensus has been reached, and the robot stops all processes. The number of loops was selected in order to ensure that every robot has reached consensus, giving any isolated robot time to find the rest of the swarm, while trying to reduce wasted time added to the process.

3.2. Implemented Robot Behavior - Hierarchical Model:

This section describes the implementation of the behavior of the simulated E-puck robots when testing the decision-making performance of the hierarchical model.

The majority of robots carry the same behavior as described previously for the voter model, but 2 out of the 10 have a "high-

er status" in that they don't directly take samples of the floor color, instead redistributing opinions received from robots around them. Because every robot carries the same hardware, any can take on this role, so the higher-level robots have been randomly selected in the simulation. They are referred to in the simulation as "managers". In the case of this simulation, it is unlikely that the specific selection of managers has much influence on results. The number of managers chosen is such that there are enough to usually make contact with every member of the swarm, while not being such a large proportion of the swarm that not enough samples are being taken to accurately represent the area.

The manager robots perform the same Levy flight to randomly move around the area. Instead of going through phases like those seen in the voter model algorithm, they store all received opinion broadcasts from other robots in a buffer, and broadcast and discard an opinion from the buffer every 0.1 seconds. This means they act to increase communication between robots, which in theory will improve the accuracy of the swarm's decision-making.

When the buffer is empty for long enough, the robot will assume that the other robots have ended their decision-making processes due to reaching consensus, thus ending its own process.

Refer to Appendix A for access to project files, including code.

■ Results and Discussion

1. Results:

The simulation was run for both models for tasks of different complexity. Task complexity is based on how many of the tiles are black. This is because the more white tiles there are, the easier it is for the swarm to then converge on the correct decision that white is the predominant floor color. Tests were run using randomly generated 8x8 grids with 32, 31, and 20 black tiles, as well as control runs of pure black and pure white to confirm the decision-making behaviors actually worked. For each number of black tiles, 3 tests were run on 3 random grids with that number of black tiles. The events of the simulation were then output to console logs, and data to be analyzed was extracted with the aid of a Python script written by Copilot. Tests using 10 black tiles per grid were planned; however, limited time resources restricted this.

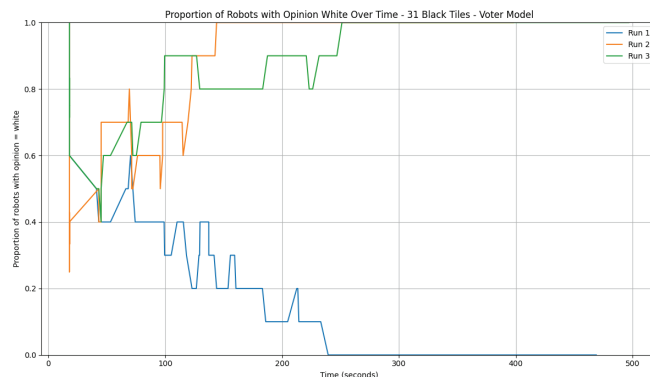


Figure 3: Proportion of robots with the correct opinion over time for the voter model, tested with 31 black tiles.

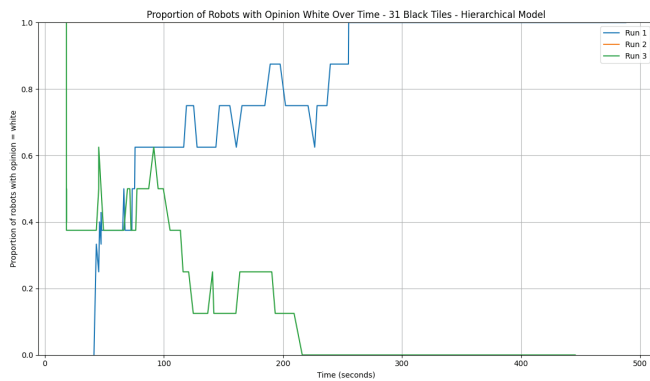


Figure 4: Proportion of robots with correct opinion over time for the hierarchical model, tested with 31 black tiles.

Figures 3 and 4 show the proportion of robots holding the correct opinion over time for both the decentralized voter-model-based strategy and the proposed hierarchical strategy. As demonstrated in the graphs, which were consistent through all trials, the hierarchical model did not reach consensus significantly faster than the decentralized voter model. It should be noted that Run 2 of the hierarchical tests on 31 black tiles does not have a visible line due to random chance, causing all the robots to begin with the same opinion, so it is an outlier. The time taken to finish the decision-making process was actually longer for the hierarchical system that was implemented in this study, due to the “manager” robots using a time-based method to determine that decision-making had ended.

Despite this slower convergence, the hierarchical strategy sometimes reached a slightly higher final proportion of correct decisions than the decentralized method. However, this improvement was small, and the difference varied between trials. Overall, both systems successfully guided most robots toward the correct color classification. The hierarchical strategy did not demonstrate a clear advantage in speed and produced only marginal changes in accuracy.

Figure 3 also shows that both models sometimes converged on the wrong option, which is black having the larger area. One possible reason for this is the stochastic nature of the decision-making process, but it cannot be confirmed in this study as the swarm size and number of trials are limited.

Raw extracted data can be accessed by referring to Appendix A.

2. Discussion:

The simulations conducted in this study show that the proposed hierarchical decision-making method does not produce a clear improvement over the fully decentralized voter-model-based approach. These findings differ from the original hypothesis and more closely match the null hypothesis, H_0 : The hierarchical decision-making method has no impact on decision-making speed/accuracy. They do not support the alternative hypothesis H_1 : The hierarchical decision-making method has an impact on decision-making speed/accuracy.

Although the hierarchical system may have provided a slight increase in accuracy, this effect was small and inconsistent across trials. More notably, the total time required for the swarm to

reach consensus increased. This appears to result from the behavior of the manager robots, which wait for confirmation of convergence before terminating their decision-making process. When managers are ignored when considering timing measurements, the difference is not significant.

This is not expected based on trends reported in the literature, where hierarchical influence has enhanced swarm performance in tasks such as formation control, navigation, or coordinated resource collection. Multiple factors can explain this difference. First, the collective perception problem studied here is more spatially distributed than tasks previously tested with hierarchical structures. While hierarchies can facilitate top-down control, they may offer less benefit when robots must independently gather environmental data and when no consistent global information can be shared reliably. Second, the small swarm size used in the simulations may have prevented hierarchical advantages from emerging. Prior work suggests that hierarchical or hybrid methods often show larger gains at bigger scales. Finally, the randomness inherent in swarm behavior, especially at this small scale, limits the statistical strength of the observed trends.

2.1. Limitations and Challenges:

Other limitations faced were the limited capabilities of the hardware used, which held back the size of the simulation that could be run. As mentioned previously, the size of our simulation may not have been large enough to remove variation due to noise in our results. This does mean that the behavior of the robot swarm could be different at larger scales and that comparisons between the two decision-making methods here may not hold. It's also acknowledged that other sources of error may have arisen from the methods used in this research, which could not be resolved due to time constraints; however, this research still proposes a novel concept to collective decision-making and makes early steps in implementing it, which may be improved in future research.

Given these considerations, the results should not be interpreted as evidence that hierarchical structures are ineffective for collective decision-making in general. Instead, they highlight that the specific hierarchical strategy implemented here may not have been optimal, and that the constraints of the simulation (limited scale, computational constraints, relatively simple manager behavior, and time available for development) may have masked the potential benefits of such an approach.

2.2. Future Work:

Future work to determine the effectiveness of a hierarchical decision-making structure should include larger swarm sizes (e.g., more than 100 robots) and optimized parameters for their behaviors to ensure accurate comparisons with other models. Comparisons should also be made to models already proven to be more effective than the voter model, such as cross-inhibition as described by Zakir *et al.*¹⁰ or the majority vote model discussed by Valentini *et al.*⁹ The decision-making structure should also be applied to other decision-making scenarios, such as site-selection or task assignment. Alternative/more optimal implementations of the same concept should

also be tested, as the constraints mentioned in this study mean that it may not have used a well-designed behavior algorithm. Improvements might be made on the way hierarchies are implemented, but also in the movement or communication of information across the swarm. Finally, tests with physical robots are necessary to determine actual effectiveness in real-world applications.

■ Conclusion

This study investigated whether a hierarchical communication structure could improve the accuracy or speed of collective decision-making in a swarm of simulated E-puck robots performing a best-of-two perception task. The results did not show a clear performance advantage over the decentralized voter-model baseline. This means that the null hypothesis that the implemented method has no effect is not rejected, and there is insufficient evidence to prove the alternative hypothesis - that the implemented method does have an effect. While accuracy may have improved slightly, consensus times increased due to manager robots delaying termination. These findings suggest that the particular hierarchical mechanism implemented here does not enhance performance under the tested conditions, especially at small swarm sizes.

The study does still show that hierarchical structures can be integrated into collective decision-making processes, which gives room for further development in the area. The support shown in the literature for implementations of hierarchical structures still shows promise for this concept. Hierarchical control systems still may be able to measurably improve decision-making speed, accuracy, and algorithm design if implemented in an optimal way. By writing this paper, this concept might be pushed out to more capable authors. Improvements in decision-making and collective perception are crucial to the viability of implementing swarm robotics systems into real-world applications such as space exploration, agriculture, and search-and-rescue by allowing robot swarms to more effectively choose the right strategies to perform their tasks without human intervention. Future work should include changes such as improved communication methods, larger swarm sizes, more optimized behavior algorithms, and comparison to other decision-making models.

■ Acknowledgments

Acknowledgments to Dr. Nisha Rawindaran, who helped guide the research process with her knowledge.

■ References

- Seeley, T. D.; Visscher, P. K.; Schlegel, T.; Hogan, P. M.; Franks, N. R.; Marshall, J. A. R. Stop signals provide cross-inhibition in collective Decision-Making by honeybee swarms. *Science* 2011, 335 (6064), 108–111. <https://doi.org/10.1126/science.1210361>.
- Dorigo, M.; Theraulaz, G.; Trianni, V. Swarm Robotics: Past, Present, and Future [Point of View]. *Proceedings of the IEEE* 2021, 109 (7), 1152–1165. <https://doi.org/10.1109/jproc.2021.3072740>.
- March-Pons, D.; Múgica, J.; Ferrero, E. E.; Miguel, M. C. Honeybee-like Collective Decision Making in a Kilobot Swarm. *Physical Review Research* 2024, 6 (3). <https://doi.org/10.1103/physrevresearch.6.033149>
- Li, J.; Yang, S. X. Intelligent collective escape of swarm robots based on a novel Fish-Inspired Self-Adaptive approach with neurodynamic models. *IEEE Transactions on Industrial Electronics* 2024, 71 (11), 14460–14469. <https://doi.org/10.1109/tie.2024.3363723>.
- Kim, D.; Lee, J.; Kim, H.-Y. Navigating the Swarm: Deep Neural Networks Command Emergent Behaviors. arXiv (Cornell University) 2024. <https://doi.org/10.48550/arxiv.2407.11330>.
- Francesca, G.; Brambilla, M.; Brutschy, A.; Trianni, V.; Birattari, M. AutoMoDe: A Novel Approach to the Automatic Design of Control Software for Robot Swarms. *Swarm Intelligence* 2014, 8 (2), 89–112. <https://doi.org/10.1007/s11721-014-0092-4>.
- Cyberbotics. *Webots: robot simulator*. Cyberbotics.com. <https://cyberbotics.com/>
- Valentini, G.; Hamann, H.; Dorigo, M. Self-organized collective decision making: the weighted voter model. *Adaptive Agents and Multi-Agent Systems* 2014, 1, 45–52. <https://doi.org/10.5555/2615731.2615742>.
- Valentini, G.; Ferrante, E.; Hamann, H.; Dorigo, M. Collective decision with 100 Kilobots: speed versus accuracy in binary discrimination problems. *Autonomous Agents and Multi-Agent Systems* 2015, 30 (3), 553–580. <https://doi.org/10.1007/s10458-015-9323-3>.
- Zakir, R.; Dorigo, M.; Reina, A. Robot swarms break decision deadlocks in collective perception through Cross-Inhibition. In *Lecture Notes in Computer Science*, 2022; pp 209–221. https://doi.org/10.1007/978-3-031-20176-9_17.
- Antonic, N.; Zakir, R.; Dorigo, M.; Reina, A. Collective Robustness of Heterogeneous Decision-Makers against Stubborn Individuals; 2024. <https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p68.pdf> (accessed 2025-08-31).
- Siemensma, T.; Chiu, D.; Ramshanker, S.; Nagpal, R.; Haghighat, B. Collective Bayesian Decision-Making in a swarm of miniaturized robots for surface inspection. In *Lecture Notes in Computer Science*, 2024; pp 57–70. https://doi.org/10.1007/978-3-031-70932-6_5.
- Chiu, D.; Nagpal, R.; Haghighat, B. Optimization and Evaluation of a Multi-Robot Surface Inspection Task through Particle Swarm Optimization. 2024, 8996–9002. <https://doi.org/10.1109/icra57147.2024.10611661>.
- Shan, Q.; Mostaghim, S. Collective Decision Making in Swarm Robotics with Distributed Bayesian Hypothesis Testing. In *Lecture Notes in Computer Science*, 2020; pp 55–67. https://doi.org/10.1007/978-3-030-60376-2_5.
- Varadharajan, V. S.; Soma, K.; Dyanatkar, S.; Lajoie, P.-Y.; Beltrame, G. Hierarchies define the scalability of robot swarms. arXiv (Cornell University) 2024. <https://doi.org/10.48550/arxiv.2405.02417>.
- Shankar, V. V.; Dyanatkar, S.; Beltrame, G. Hierarchical Control of Smart Particle Swarms. arXiv (Cornell University) 2022. <https://doi.org/10.48550/arxiv.2204.07195>.
- Zhu, W.; Oguz, S.; Heinrich, M. K.; Allwright, M.; Wahby, M.; Christensen, A. L.; Garone, E.; Dorigo, M. Self-organizing nervous systems for robot swarms. arXiv (Cornell University) 2024. <https://doi.org/10.48550/arxiv.2401.13103>.
- Zhang, Y.; Oğuz, S.; Wang, S.; Garone, E.; Wang, X.; Dorigo, M.; Heinrich, M. K. Self-Reconfigurable hierarchical frameworks for formation control of robot swarms. *IEEE Transactions on Cybernetics* 2023, 54 (1), 87–100. <https://doi.org/10.1109/tcyb.2023.3237731>.

10. Trabattani, M.; Valentini, G.; Dorigo, M. Hybrid control of swarms for resource selection. In Lecture Notes in Computer Science, 2018; pp 57–70. https://doi.org/10.1007/978-3-030-00533-7_5.
11. Reina, A.; Bose, T.; Trianni, V.; Marshall, J. A. R. Effects of spatiality on Value-Sensitive decisions made by robot swarms. In Springer Proceedings in Advanced Robotics; 2018; pp 461–473. https://doi.org/10.1007/978-3-319-73008-0_32.
12. Trianni, V.; De Simone, D.; Reina, A.; Baronchelli, A. Emergence of Consensus in a Multi-Robot Network: From abstract models to Empirical validation. IEEE Robotics and Automation Letters 2016, 1 (1), 348–353. <https://doi.org/10.1109/lra.2016.2519537>.

■ Author

Kevin Wu is a student of John Paul College in Rotorua, New Zealand, who has a long-standing love for the field of robotics. He hopes to major in mechanical engineering or mechatronics to work on robots that interface with humans.

■ Appendices

Appendix A – Simulation Files:

All Webots simulation files, including code and extracted data, can be found in a GitHub repo at the URL:

https://github.com/kevwano/swarm-hierarchy/tree/main/emitter_receiver/controllers/robot1

While Webots itself may be installed at:

<https://cyberbotics.com/>

The project can be accessed by opening the “worlds” folder and opening the file “emitterandreceivertesting_ver2.wbt” in Webots. It is recommended that, for optimal performance, the graphics settings be turned to the lowest.

Exploring the Growth and Limitations of *Ex Utero* Intrapartum Treatment

Kyle N. Alchaar

San Clemente High School, 6506 Ladera Brisa, San Clemente, California, 92673, USA; alchaarkyle@gmail.com

Mentor: Barbara Biney

ABSTRACT: *Ex utero* Intrapartum Treatment (EXIT) offers a method to secure fetal airways during delivery by temporarily preserving placental support, yet its use remains rare. This paper analyzes the administrative, financial, and organizational factors that continue to limit the expanded adoption of EXIT across the United States. The procedure is exclusively feasible at a small number of centers due to equipment costs, personnel costs, and management. Regional inequalities bolster this effect, and underserved areas are even further restricted. Research efforts face similar obstacles. The low number of cases and inconsistent process protocols hinder data collection. Additionally, the NIH, under the current administration, is confronting considerable budget cuts. Historically low federal support for maternal health research further halted progress. However, strengthening training pipelines and targeting grant allocation alongside expanding institutional support through public and private partnerships will open pathways to greater utilization. The remediation of funding and access issues is necessary for EXIT to move beyond specialized centers and become a viable and just option for a larger range of patients facing prenatal complications.

KEYWORDS: Translational Medical Sciences, Disease Treatment and Therapies, Maternal-Fetal Medicine, Fetal Surgery, *Ex utero* Intrapartum Treatment.

■ Introduction

Ex utero Intrapartum Treatment (EXIT) is a complex surgical technique that allows for the controlled delivery of fetuses with airway compromise. It does this through the maintenance of uteroplacental circulation during the establishment of a secure airway. In spite of its potential, EXIT remains underutilized in clinical settings and has struggled to gain the widespread recognition it needs for significant change.

Originally developed in the 1990s, EXIT was first used to “treat temporary tracheal occlusion in patients who had undergone fetal surgery for severe congenital diaphragmatic hernia.”¹ Since its initial application, the procedure has evolved to manage a wide variety of conditions. Some of these conditions include congenital high airway obstruction syndrome (CHAOS), large cervical or thoracic masses, and tracheoesophageal fistula.

The EXIT approach includes 4 possible procedures, including EXIT-to-Airway, EXIT-to-ECMO, EXIT-to-Resection, and EXIT-to-Separation. EXIT-to-Airway is the most common and is surgically indicated by CHAOS, severe micrognathia, or an obstructing mass.² EXIT-to-ECMO is a controversial procedure with no recent statistically significant difference in morbidity for severe congenital diaphragmatic hernia (CDH) patients.³ This cohort is primarily used in cases with CDH or Cardiothoracic anomalies.² In cases of mass obstructions in the head, neck, or chest, EXIT-to-Resection is often performed. Finally, EXIT-to-Separation is reserved for conjoined-twin procedures.² In general, EXIT has a low maternal complication rate, with only 3 of 40 responding institutions (7.5%) reporting major adverse maternal outcomes.⁴ 11 of the same 40 responding institutions (27.5%) report adverse events such as

hypoxic-ischemic encephalopathy (HIE), “death, need for fetal resuscitation, and ‘adverse events such as sequela of disease.’”⁴

Although it has potential, the use cases of EXIT remain limited, in part due to high resource requirements and the complicated nature of the procedure itself. The procedure demands a qualifying patient and the availability of a multidisciplinary surgical team. It also requires an institution with access to advanced surgical infrastructure. This combination of selective eligibility and high institutional costs restricts the general adoption of EXIT. In doing so, the production of procedural research is also limited. Moreover, clinical implementation and research into EXIT are mutually dependent: lacking frequent use, meaningful research is hindered; without sufficient research, hospital systems may lack the evidence base required to justify its use.

Another barrier to the advancement of EXIT-related research is the threat of federal budget cuts, particularly in proposed reductions to National Institutes of Health (NIH) funding under the current administration. The Department of State requested a ~40% cut to the NIH budget, entailing a reduction from \$27.5 billion to \$18 billion for 2026.⁵ Given the already limited research on this topic, if enacted in the future, such cuts could implicate serious threat to progress. Furthermore, despite large advances in maternal health over the past decade, considerable improvements are still necessary. Though the United States invests nearly \$5 trillion annually in healthcare, this investment does not consistently translate into positive maternal health outcomes. In fact, the U.S. continues to hold the highest maternal mortality rate among economically comparable nations.⁶

Overall, the potential for a wider implementation of EXIT in the United States is hindered by worsening funding limitations. Additionally, extensive qualifying criteria make the employment of the procedure currently hard to justify, given the high costs, most prevalently in the areas of research, staffing, and equipment. This paper intends to explore these limitations and present possible solutions that could improve both the accessibility of the procedure and the knowledge of its clinical value.

■ Results and Discussion

One noteworthy aspect of the underutilization of EXITs is found in equipment restrictions. Due to the highly specialized nature of the procedure, equipment is often costly and hard to get. While the specific required equipment varies on a case-by-case basis, generally, EXIT uses what is listed in Table 1.

Table 1: This table gathers estimated price ranges for critical equipment and materials required in Ex Utero Intrapartum Treatment. Using the midpoints of each range totals ~\$181,000. Quality and brand determine where in the listed ranges a specific item's cost may fall. In equipment like an ultrasound machine, fetal resuscitation bay, or adjacent operating room, specific styles and features also affect the price range. The cost of blood for transfusions is a generalization and not entirely indicative of what a specific institution may pay. *List of items cited from Jong, R. and Fordham, T.⁷

Item	Estimated Cost
Anesthesia circuits for both mother and fetus with invasive maternal blood pressure monitoring*	\$10,000-\$150,000
Foley catheter*	\$10-\$50
Two large-bore peripheral intravenous lines*	\$10-\$50
Type-and-crossed blood for possible maternal transfusion in the room*	\$180-\$300
Epidural catheterization*	\$300-\$1,000
Ultrasound machine with sterile sleeve*	\$5,000-\$120,000
Cesarean delivery surgical set*	\$300-\$1500
Electrocautery*	\$2,000-\$15,000
Uterine stapling device*	<\$1000
Amnioinfusion*	\$485-\$1000
Fetal resuscitation bay*	\$7,500
Adjacent operating room set up to receive infant*	>\$20,000

The expensive equipment required for EXIT procedures, as seen in Table 1, limits their workability. Only a small number of institutions, often academic medical centers containing dedicated fetal programs, can perform the surgery. A 2014 survey identified 59 fetal care centers in the United States. Of 29 survey responses, 23 have the capability to perform EXIT procedure, and only 15 have the capability to perform the similar procedure of EXIT to ECMO.⁸ 2023 Data from the North American Fetal Therapy Network (NAFTNet) indicated that, of a responding 41 institutions from a pool of centers from the United States, Mexico, Canada, Argentina, Belgium, and Switzerland (91% response rate for a survey sent to all NAFTNet institutions), 40 are capable of performing EXIT.⁴ There are currently 54 NAFTNet centers internationally, 47 of which are in the United States.⁹ These centers are concentrated on the West Coast, the Northeast, and select parts of the Midwest and South. Though some regions are concentrated, others are scarce; ultimately leaving significant areas of the country,

particularly the Mountain West and rural Southeast, without direct access to EXIT-capable facilities.⁹ The scarcity of these centers potentially leaves patients outside of these regions to face geographic difficulties in accessing potentially necessary critical surgical intervention.

Another costly limitation of EXITs is staffing. Having a multidisciplinary team of specialists is necessary because of the risks intrinsically associated with EXIT. Teams will regularly include an obstetrician, an obstetric anesthesiologist, a neonatal anesthesiologist, a pediatric surgeon, a pediatric otolaryngologist, a neonatal intensivist, a possible ECMO team, a possible pediatric cardiology or cardiothoracic surgery team, a circulating nurse, and a surgical technician.⁷

Additionally, examining the training patterns in maternal-fetal medicine reveals how some problems regarding the availability of doctors may persist into the future. Training in Maternal-Fetal Medicine still follows a structured, usually multi-year fellowship model, although some shifts are being made in program size and scheduling that could potentially affect workforce availability in the future. For instance, UCLA's maternal-fetal medicine fellowship. It will begin to alter its match cycle starting in August 2025, in which they admit two fellows in one year, then one fellow the next, thus reducing the rate of new specialists.¹⁰ This draws attention to potential training constraints in the future. ACGME's 2025 standards mandate the completion of at least 18 months of clinical training in addition to 12 months of research time with cross-disciplinary exposure. Many programs still adhere to this. Given the opportunity cost and length of obligation to the lengthy training durations, these requirements may also turn away trainees even though they demonstrate competency. Some programs, like UCSD, report one fellowship position annually or three fellows every three years. As a whole, the volume of trainees is stable although restricted, therefore potentially preventing further growth. A significant increase in fellowship slots or program flexibility is needed to ensure increased fellowship output and respond to the growing clinical demand in high-risk maternal-fetal surgery.

Despite this, maternal-fetal medicine is a growing field that could remedy the need for multidisciplinary personnel for EXIT. In 2025, the Society for Maternal-Fetal Medicine (SMFM) had over 6,500 members.¹¹ This is a marked increase from the approximately 1,355 board-certified MFM subspecialists reported in 2010.¹² The increased institutional focus on fetal medicine is reflected in this growth. Nevertheless, access to care varies by region. Academic reviews from 2019 say that over 90% of counties don't have a practicing MFM physician. The regional distribution is emphasized even more, as it shows that counties without a practicing physician experience higher preterm birth rates.¹³ This disproportion appears to persist into 2025, with continued concentration of MFM practitioners in larger cities and major fetal health centers. Although numbers are increasing, there is still a disconnect between population need and subspecialist presence. Fellowship training spots need to be expanded in zones where they are lacking. The growing number of specialists might not result in an equita-

ble distribution of services like EXIT. Thus, the expansion of training may prove effective in broadening treatment equity.

The final major limitation of EXIT is in research. The limited sample size due to EXIT's narrow scope of patients is one aspect of this. EXIT is a procedure reserved mainly for highly complex conditions (CHAOS, large neck masses, tracheo-esophageal fistula), and by the nature of this, studies are often small-scale or case reports, rather than large-scale clinical trials. A review noted the limited data on maternal morbidity, long-term fetal outcomes, and cost-utility.¹ Examples such as this, displaying a lack of complete data, may make it difficult for non-specialized hospitals to consider adopting EXIT.

Another aspect of research limitations is the effects of the current administration on the NIH. The NIH is the primary research institution for health in the United States. Following \$1.81 billion in grant terminations from February 28, 2025, to April 8, 2025¹⁴ and the DEI-related funding cut of \$783 million in August 2025,¹⁵ the current NIH has fallen around \$1 billion behind the spending pace of previous years.¹⁶ As of late March 2026, the National Cancer Institute has funded only \$72 million in New and Competitive Grants, in comparison to the \$250 million pace intended to be spent by an agreed-upon point in the fiscal year 2026.¹⁶ The effect of these cuts is amplified because maternal health already receives poor fund allocation from the government. Of the top 15 institutes funded according to the NIH's RePORT page, only ~5.5% of projects funded are performed under the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), where most studies on the subject of maternal health lie.¹⁷

A lack of standardization contributes to EXIT's variability, which further impairs its general utilization. Limited research and consensus consultation recommendations restrict the accumulation of long-term data based on outcomes, making it difficult for hospitals to weigh risks versus benefits or for hospital administrations to justify expanding funding. A survey given to NAFNet centers identified that 64.9% of centers (91% completed the survey) reported "at least one experience of airway intervention and management having the potential to be successful without EXIT," and from an under allocation perspective, adverse outcomes were reported by 10.8% of those centers when "the fetal surgery team was unaware of the patient in any regard and a potential mitigation of these events would have been possible with EXIT."⁴ Perpetuating this idea, maternal and neonatal research in the United States remains underfunded and inequitable,¹⁸ which are problems that disproportionately disadvantage niche procedures, such as EXIT, that do not have the clinical trial data of more established procedures.

Strategic investment in fetal surgical research, equipment, and personnel will mend the knowledge gaps surrounding EXIT; the same gaps that stall its progress and limit access for patients who would benefit the most.

Solutions:

One method to expand the uses of EXIT is through increasing access to federal and private research grants targeted

towards maternal-fetal care. As of 2025, NIH funding has been cut, with even more damage proposed; however, financial assistance and innovation in procedures like EXIT still exist. The NIH's Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs fund early-stage research for high-risk, high-reward projects. Inventions such as advanced fetal imaging, maternal monitoring systems, or surgical simulations for training, all related to EXIT innovation, fall within the eligibility of these research programs as well. Additionally, the NICHD offers grants under groups like the Human Placenta Project and Maternal Health Research Networks. These could be utilized for EXIT-focused studies with a wider perspective on maternal-fetal medicine. Pursuing these grants could allow specialized centers to track large-scale outcome data and conduct studies that would demonstrate the safety of EXIT. These demonstrations may provide the necessary data to allow hospitals to make well-informed decisions regarding EXIT.

Alongside grant access, alleviating infrastructure costs through public or private partnerships could help with the significant costs required to maintain the ability to perform EXIT. Existing groups like the Health Resources and Services Administration's (HRSA) Maternal and Child Health Services Block Grant Program offer funding to states to support services and systems for mothers and their children. States with low maternal health outcomes could potentially target these funds towards strengthening high-risk fetal centers capable of hosting EXIT procedures. In recent years, private organizations like the Bill & Melinda Gates Foundation and March of Dimes have provided funding for fetal technology and health programs. These joint efforts would be especially necessary in covering large-scale investments like fetal resuscitation bays and updated ultrasound technologies. These are often pieces of equipment that remain out of reach for many hospitals operating on smaller margins due to their substantial costs.

Finally, more access to specialized training and creating strong support structures within hospitals could contribute to the increase in the availability of EXIT procedures. Forming financial incentives for hospitals that develop the capabilities of EXIT could encourage growth, notably in areas that lack resources. It would also be beneficial to make it easier for EXIT-trained teams to function efficiently within the hospital environment. Changes like this prove necessary for EXIT's expansion, especially due to its high-risk nature. In addition, streamlining scheduling and procuring necessary equipment further helps to reduce barriers in implementation. For training, internet platforms such as OPENPediatrics and the Global Fetal Surgery Network provide growing infrastructure for online education, meaning specialized EXIT procedure knowledge could be more widely accessed than ever before.¹⁹

Researchers need to be aware of the specific grants available and the requirements to procure funding opportunities related to EXIT procedures. Considering the current administration, working through the current grant landscape endures as an essential step forward. Figure S1 provides examples of key grants, including federal, institutional, and private, that could be used

to expand EXIT-related research, training, and infrastructure expansion. These grants provide financial support and signal institutional interest in advancing maternal-fetal innovation for EXIT. By aligning EXIT-related proposals with the priorities of these funders, researchers and clinical teams could better position their work for further growth. With consistent support, EXIT can expand beyond its limited uses.

Limitations & Future Opportunities:

This study acknowledges possible limitations. The most notable of these is the scarce access to primary-source data. Figures like the number of fetal centers currently offering EXIT or analysis of EXIT outcomes per institution were referenced through secondary literature or older surveys rather than through direct and up-to-date institutional reports. Furthermore, detailed cost analyses relied on external equipment marketers and public health cost estimations rather than on internal hospital financial reports, meaning that the pricing estimates presented, while remaining conservative, are still generalizations and not specific.

Additionally, this research collects data from many non-standardized sources, including varying sample populations, clinical protocols, and reporting standards. As a procedure, EXIT is highly individualized, with outcomes heavily dependent on specific maternal status, fetal status, and institutional setup, and such diversity makes generalizations challenging. Because much of the existing research on EXIT is composed of case studies often following one specific instance of it, this paper is similarly limited in its power to draw definitive conclusions regarding 'best practices' or long-term cost-benefit analysis.

Further studies could help resolve these restrictions in several ways. First, interviews with hospital administrators and maternal-fetal medicine teams could provide helpful first-hand perspectives regarding EXIT. Furthermore, an analysis of cost-effectiveness, such as through hospital-accurate data collection, would be valuable. This would not only help measure the direct expense of EXIT, but also healthcare savings as a whole in relation to improving neonatal results and reducing hospital stays. Lastly, regularly updated versions of surveys administered to the North American Fetal Therapy Network could help determine EXIT readiness and how clinical attitudes have changed in response to changes in policy and funding cuts.

Conclusion

EXIT is a life-saving intervention for fetuses with a critical airway compromise; however, its adoption is limited by the factors that also hinder research into the procedure. High resource requirements and multidisciplinary dependencies both contribute to EXIT's underutilization and to the narrow evidence base available. Funding pressures on NIH from the current administration, coupled with limited maternal-fetal health resource assignment, may further restrict opportunities for substantial development.

However, progress is still achievable. Expanding the funding for EXIT research and adjacent studies through federal

and private grants, along with forming public and private partnerships to offset equipment and staff costs, could allow for greater access and procedural standardization. Furthermore, establishing data registries and accentuating transparent and reproducible study design (to the best of one's ability concerning patient variability) will build the research required to demonstrate safety, cost-effectiveness, and long-term benefit to hospital administrators.

EXIT has the potential to transition from a niche technology into a highly useful tool in improving outcomes for families facing complex prenatal diagnoses.

Acknowledgments

I would like to thank Dr. Biney, my mentor, for her support and overall help throughout my time on this project.

References

1. Hirose, S.; Harrison, M. R. The *Ex Utero* Intrapartum Treatment (EXIT) Procedure. *Seminars in Neonatology* 2003, 8 (3), 207–214. [https://doi.org/10.1016/s1084-2756\(03\)00029-0](https://doi.org/10.1016/s1084-2756(03)00029-0).
2. Bence, C. M.; Wagner, A. J. *Ex Utero* Intrapartum Treatment (EXIT) Procedures. *Seminars in Pediatric Surgery* 2019, 28 (4), 150820. <https://doi.org/10.1053/j.sempedsurg.2019.07.003>.
3. Shieh, H. F.; Wilson, J. M.; Sheils, C. A.; C. Jason Smithers; Kharasch, V. S.; Becker, R. E.; Studley, M.; Morash, D.; Buchmiller, T. L. Does the *Ex Utero* Intrapartum Treatment to Extracorporeal Membrane Oxygenation Procedure Change Morbidity Outcomes for High-Risk Congenital Diaphragmatic Hernia Survivors? *Journal of Pediatric Surgery* 2017, 52 (1), 22–25. <https://doi.org/10.1016/j.jpedsurg.2016.10.010>.
4. Joshi, D.; Stelson, M.; Antony, K.; Beninati, M.; Luks, F. I.; Puricelli, M.; Lobeck, I. N. Indications, Resource Allocation, and Outcomes Associated with *Ex-Utero* Intrapartum Treatment Procedures: A North American Fetal Therapy Network Survey. *Fetal Diagnosis and Therapy* 2023, 50 (5), 376–386. <https://doi.org/10.1159/000531615>.
5. Frank, R. G. The 2026 health and health care budget. Brookings. <https://www.brookings.edu/articles/the-2026-health-and-health-care-budget/>.
6. Gunja, M.; Gumas, E.; Masitha, R.; Zephyrin, L. Insights into the U.S. Maternal Mortality Crisis: An International Comparison. The Commonwealth Fund. <https://www.commonwealthfund.org/publications/issue-briefs/2024/jun/insights-us-maternal-mortality-crisis-international-comparison>.
7. Jong, R. D.; Fordham, T. *Ex utero Intrapartum Treatment (EXIT) Procedure*. Nih.gov. <https://www.ncbi.nlm.nih.gov/books/NBK604209/> (accessed 2025-07-31).
8. Kett, J. C.; Woodrum, D. E.; Diekema, D. S. A Survey of Fetal Care Centers in the United States. *Journal of Neonatal-Perinatal Medicine* 2014, 7 (2), 131–135. <https://doi.org/10.3233/npm-14814005>.
9. NAFTNet Centers | North American Fetal Therapy Network. [naftnet.org https://www.naftnet.org/i4a/pages/index.cfm?page-ID=3305](https://www.naftnet.org/i4a/pages/index.cfm?page-ID=3305) (accessed 2025-07-31).
10. Maternal-Fetal Medicine Fellowship Program. [uclahealth.org https://www.uclahealth.org/departments/obgyn/education/fellowship-programs/maternal-fetal-medicine-fellowship-program](https://www.uclahealth.org/departments/obgyn/education/fellowship-programs/maternal-fetal-medicine-fellowship-program) (accessed 2025-07-31).
11. Mission & Vision - Society for Maternal-Fetal Medicine. [smfm.org https://www.smfm.org/mission-and-vision](https://www.smfm.org/mission-and-vision).
12. Rayburn, W.; Klagholz, J.; Elwell, E.; Strunk, A. Maternal-Fetal Medicine Workforce in the United States.

- American Journal of Perinatology 2012, 29 (09), 741–746. <https://doi.org/10.1055/s-0032-1316445>.
13. Greiner, A. L.; Haeri, S.; Nidey, N. L. Preterm Births and Maternal-Fetal Medicine Physician Workforce Location in the United States. *American Journal of Perinatology* 2025. <https://doi.org/10.1055/a-2531-2783>.
 14. Santoro, C. NIH Grants Terminated Amid Trump Administration, Raising Concerns for US Research, Minority Health Disparities. *AJMC*. <https://www.ajmc.com/view/nih-grants-terminated-amid-trump-administration-raising-concerns-for-us-research-minority-health-disparities>.
 15. Whitehurst, L. Supreme Court allows Trump to cut nearly \$800M of research funding. *AP News*. <https://apnews.com/article/supreme-court-trump-nih-dei-320a6b3749bf56703b-50739362d1238c>.
 16. Mueller, B.; Hwang, I. Pace of N.I.H. Funding Slows Further in Trump's Second Year. *Nytimes.com*. <https://www.nytimes.com/2026/04/22/science/trump-nih-funding-research.html>.
 17. Home | RePORT. report.nih.gov. <https://report.nih.gov/>.
 18. *Birth Settings in America*; Scrimshaw, S. C., Backes, E. P., Eds.; National Academies Press: Washington, D.C., 2020. <https://doi.org/10.17226/25636>.
 19. *About Us*. *OPENPediatrics*. <https://www.openpediatrics.org/aboutus>

■ Author

Kyle is a student attending San Clemente High School in Orange County, California. He is interested in biological sciences, particularly from a biomedical perspective. He begins attendance at the University of California, Los Angeles in the fall of 2026 as a biology major and wishes to pursue a position in healthcare in the future.

A Stochastic Model of Sabre Fencing

Lynn Uemoto

Franklin School, 80 Town Square Place, Jersey City, NJ, 07310, USA; lynn.uemoto@gmail.com

ABSTRACT: This paper presents a statistical analysis and model framework for simulating and analyzing sabre fencing bouts using discrete stochastic processes. Sabre fencing is a ‘first-to-15’ sport without time constraints, which aligns well with event-based modeling. The model uses a Poisson race model, where each fencer scores touches based on their average scoring rate (λ). The parametric approach allows for the derivation of analytical win probabilities via the negative binomial distribution and enables comparison of fencers who have never competed against each other. To capture the real-world impact of penalties, a state-dependent Markov chain was used to model the FIE’s escalating card rules, moving beyond a simple memoryless model. Simulations of 10,000 bouts validate the framework against theoretical results and quantify the influence of scoring rates and penalties on bout outcomes. This study establishes a foundational, parametric model for sabre, demonstrating how classical probability models can effectively represent dynamic, rule-driven competition and provides a framework for understanding and predicting outcomes in sabre fencing.

KEYWORDS: Mathematics, Probability and Statistics, Stochastic Processes, Poisson Distribution, Sport Analytics.

■ Introduction

In recent years, the field of statistical modeling has become fundamental to advanced performance analysis across many sports. The rise of sports analytics, popularized by “sabermetrics” as seen in the movie *Moneyball*, signaled a major shift: the use of probabilistic models to uncover hidden values that traditional scouting methods could not.¹ This concept has since altered many other major sports. For instance, Poisson distribution-based models have become prominent in predicting goals scored in hockey and soccer. In contrast, Markov chain models have been used to evaluate possession value in basketball, and linear probability models for tennis.²⁻⁵ At the center of this revolution is the use of probability not just to describe what happened, but to forecast what will happen.

Despite this revolution, fencing remains largely unexplored. Existing quantitative research falls into three groups. The first focuses on physical and technical characteristics for technical and tactical analysis, notably biomechanical or motion analysis.^{6,7} While valuable, there is a gap, as strength in fencing is not only a product of isolated physical attributes. A fencer with high speed or reach may lose to an opponent with great and practical strategy or decision-making. A great example is Aron Szilagyi, the three-time individual Olympic sabre champion. While not being the fastest or tallest, his exceptional footwork and tactical intelligence have consistently allowed him to outperform athletes with more physical advantages. The second and third focus on visualization and deep learning tools for interpreting bout data and statistical models, as exemplified by Bradley-Terry to infer aggregate “player strength” from win-loss records.⁸⁻¹¹ Addressing the limitations in the first group, a recent and complementary research focuses on simulating bouts from video data by modeling sequences of discrete actions. Lin *et al.* (2025) developed a pipeline to extract fencer motion and positions from video, clustering them into discrete actions, and

then modeling bout progression using a probabilistic state machine.⁹ This data-driven model effectively captures the tactical interplay of action in a touch.

While these existing tools have advanced fencing analytics, two limitations remain. First, biomechanical and technical analyses describe physical attributes or movement patterns but are not designed to produce win probabilities or to isolate how a fencer’s ability translates to bout contexts. Second, approaches, such as Bradley-Terry ratings and video-based state machine simulators, can rank fencers or replicate actions, yet they do not provide a compact, parametric model that yields closed-form solutions and enables prediction for fencers who have never faced each other. Consequently, a gap exists for a foundational, interpretable probability model at the bout level. The approach of Lin *et al.* is complex, while foundational works in sports analytics have consistently shown the power of simple, parametric models, for example, the Poisson distribution, for general insights and closed-form solutions.^{2,3} This proposed Poisson race model addresses this gap, positioning as the “xG of fencing,” whereas the model of Lin *et al.* functions as a detailed “play-by-play” simulator that generates tactical sequences.⁹

The sport’s ‘first-to-15’, discrete-points structure naturally aligns with discrete-event models. Bouts are treated as a race between two independent Poisson processes, providing closed-form win probabilities via the negative binomial distribution. Penalties are incorporated through a discrete-time Markov chain that captures the FIE’s escalating penalty card rules. The research addresses two core questions: (1) How can discrete probability distributions simulate and analyze sabre bouts? (2) What insights can these simulations provide about match outcomes and the influence of penalties? By addressing these, the study aims to provide a simple, parametric framework that yields interpretable win probabilities and quantifies risk from

penalties, establishing a benchmark for future fencing analytics.

■ Methods

This study combines both theoretical probability models and computational simulations to model sabre fencing. A sabre bout is won by the first fencer to score 15 touches. As touches occur as separate, random events, the scoring process can be described using discrete probability distributions. The model uses a Poisson process and its exponential inter-arrival times, while penalties (yellow and red cards) are modeled as a discrete-time Markov chain. Analytical win probabilities are derived from the negative binomial distributions and simulations, and Python is used to validate and extend these analytical results.

Poisson Race Model and Closed-Form Win Probability:

A sabre bout is conceptualized as a race between two independent Poisson processes.¹² A scoring parameter characterizes each fencer, λ , which represents the average number of touches scored per bout. In this context, since a bout ends when reaching 15 points and has no time limit, the “time” variable is abstract. It represents the bout through scoring exchanges, not physical clock time. Therefore, the “time interval” is set to $t=1$ bout, simplifying the model. The number of touches, X , scored by a fencer thus follows a Poisson distribution.¹²

The Probability Mass Function (PMF) of a Poisson distribution gives the probability of how many times k events occur in a fixed interval.^{13,14}

$$P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \quad k = 0, 1, 2, \dots$$

This model relies on key assumptions that are reasonable approximations for sabre: (1) events occur independently of previous touches, (2) scoring rate λ is constant throughout the bout (no fatigue or momentum), and (3) in a single exchange, the probability of more than one touch is negligible. Strictly, only one scoring touch can occur per exchange, but a penalty point (awarded for a red card) may be added during the same exchange. The model, therefore, treats the scoring touch and penalty point as separate events rather than multiple touches.

Because the two Poisson processes are independent, the probability that the next touch in the bout is scored by Fencer A against Fencer B is constant and given by¹³

$$p_A = \frac{\lambda_A}{\lambda_A + \lambda_B}$$

where this probability does not depend on the state of previous touches; a memoryless property inherited from the Poisson processes. Although the number of exchanges between touches is random (and would follow an exponential distribution if we measured abstract “bout time”), only the order of touches matters to determine the winner.^{13,14} Thus the bout can be treated as a sequence of independent Bernoulli trials with success probability p_A .

A key property of the Poisson distribution is that its variance is equal to its mean (λ).¹²⁻¹⁴ This means that a fencer's scoring is unpredictable, where the number of touches they score in a bout will naturally vary from one bout to another,

with a spread defined by λ . This built-in variance is a strength of the model, as it captures the element of chance in any sport. Adding to the memoryless property of the exponential distribution, this ‘first-to-15’ race amplifies that underlying randomness, ensuring that even a significantly stronger fencer ($\lambda_A > \lambda_B$) has a non-zero probability of losing, which aligns with the real-world dynamics of sabre fencing, where upsets are possible, and no fencer is immune to it.

The number of touches scored by Fencer B before Fencer A reaches 15 therefore follows a negative binomial distribution. (see Appendix A). In general, for a sequence of independent trials with success probability p , the cumulative probability of at most k failures before the r -th success is:¹⁴

$$P(\text{at most } k \text{ failures}) = \sum_{k=0}^r \binom{k+r-1}{k} (1-p)^k p^r$$

Applying this to a first-to-15 sabre bout with $r=15$, $k=14$, and $p=p_A$, Fencer A wins if Fencer B scores at most 14 touches before A's 15th touch. Therefore:

$$P(A \text{ wins}) = \sum_{k=0}^{14} \binom{k+14}{k} (1-p_A)^k p_A^{15}$$

Here the binomial coefficient $\binom{k+14}{k}$ counts the number of sequences in which Fencer B scores exactly k touches before Fencer A's 15th touch.¹³ This closed-form theoretical probability serves as a benchmark to compare with simulation results and validate the simulation.

For example, with $\lambda_A=12$ and $\lambda_B=8$, we have $p_A=0.6$ and closed-form probability as follows:

$$P(A \text{ wins}) = \sum_{k=0}^{14} \binom{k+14}{k} 0.6^{15} 0.4^k \approx 0.864$$

Modeling Penalties with State Dependent Thresholds:

The initial baseline model for penalties treated each card as an independent event, modeled with a Bernoulli process. However, the FIE's penalty escalation rules are state-dependent.¹⁵ To accurately capture this, we extend the model to a discrete-time Markov chain, where the probability and consequence of a penalty depend on the fencer's current card state.

In sabre fencing, we consider 2 penalty types relevant to bout scoring: yellow cards and red cards. Black cards, which result in immediate disqualification from the entire competition, are excluded from this model due to their extreme rarity in bout-level outcomes. Under FIE rules, a fencer starts in a two-yellow threshold regime: two yellow cards accumulated before any red card result in a red card and a penalty point for the opponent. Once a fencer receives any red card (directly or through two yellow cards), they enter a one-yellow threshold regime: any subsequent yellow card immediately yields another red card (and another penalty point).

This system is naturally modeled as a discrete-time Markov chain, where the probability of transitioning between penalty states depends only on the current state.¹⁶ The model defines three distinct states based on the current threshold and the number of yellow cards a fencer has at touch k , where $S_{0,k}$ is the starting state for every bout. The states S_0 and S_1 belong to the initial two-yellow regime; T_0 is the post-red-one-yellow threshold regime.

- $S_{0,k}$: The probability that the fencer is in the initial “two-yellow threshold” regime and has zero yellow cards

- $S_{1,k}$: The probability that the fencer is in the initial “two-yellow threshold” regime and has one yellow card
- $T_{0,k}$: The probability that the fencer is in the post-red “one-yellow threshold” regime and has zero yellow cards

On any touch, a fencer may receive a yellow card with probability p_y , a direct red card with probability p_r , or no card ($p_N = 1 - p_y - p_r$). The state transitions are illustrated in Figure 1, which shows the three penalty states and probabilities of moving between them.

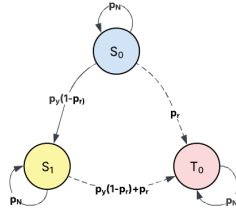


Figure 1: State transition diagram for the sabre fencing penalty model. This diagram illustrates the three penalty states and the probabilistic transitions between them through a Markov chain.

It is important to distinguish the triggering mechanism from the consequences of a card. While p_y and p_r are constant, memoryless probabilities, whether a yellow card awards a penalty point is state-dependent: in the two-yellow regime (S_0 , S_1), two yellows are needed for a red card; in the one-yellow regime (T_0), a single yellow immediately yields a red card. This hybrid model is set with Bernoulli card events with a state-dependent scoring impact, which is the key extension beyond a simple binomial penalty model.

The probability of a penalty point being awarded on touch k is therefore

$$a_k = S_{0,k}p_r + S_{1,k}(p_y + p_r) + T_{0,k}(p_y + p_r)$$

The state probabilities update as follows:

$$\begin{aligned} S_{0,k+1} &= S_{0,k}p_N \\ S_{1,k+1} &= S_{0,k}p_y + S_{1,k}p_N \\ T_{0,k+1} &= T_{0,k}p_N + a_k \end{aligned}$$

with $p_N = 1 - p_y - p_r$ and initial conditions $S_{0,0} = 1$, $S_{1,0} = T_{0,0} = 0$. This models the probability of a red card at each exchange, capturing the escalation rules compactly.

Integration of Models:

To unify the scoring and penalty processes, we frame the bout using joint probability distributions.¹⁶ The framework assumes that the scoring fencer is not the one receiving a card in the same exchange; situations where a touch is annulled because the scored committed an offence (e.g., guard punch, leaving piste) are excluded. This simplification keeps the joint distribution tractable and focuses on the more common case where the penalty goes against the non-scoring fencer.

Let X represent the scoring outcome and Y represent the penalty outcome for each exchange. Their joint probability mass function is $P(X=x, Y=y)$.¹⁶ This gives the probability that the outcome x for scoring and outcome y occur together on the same touch.

For each exchange, we consider six mutually exclusive joint events:

1. A scores, no red card for either $\rightarrow A + 1$
2. B scores, no red card for either $\rightarrow B + 1$

3. No score, A receives a red $\rightarrow B + 1$ (penalty point)
4. No score, B receives a red $\rightarrow A + 1$ (penalty point)
5. A scores and B receives a red card. $\rightarrow A + 2$ (touch + penalty point)
6. B scores and A receives a red card. $\rightarrow B + 2$ (touch + penalty point)

Let R_B denote the event that fencer B receives a red card. Then $P(A \text{ scores}, Y = R_B) = P(A \text{ scores}) \cdot P(R_B)$, assuming independence between the scoring outcome and card events as the bout progresses.

And similarly for the symmetric case. This reflects situations where a fencer scores a valid touch while their opponent gets a penalty card.

The overall bout outcome marginalizes over the joint distribution of penalty states:

$$P^*(A \text{ wins}) = \sum_{y_A} \sum_{y_B} P(A \text{ wins} | Y_A = y_A, Y_B = y_B) \cdot P(Y_A = y_A, Y_B = y_B)$$

where, $P(Y_A = y_A, Y_B = y_B)$ is the joint probability of penalty states from the Markov chain. This is estimated by running a Markov chain simulation and counting how often each state combination occurs. $P(A \text{ wins} | Y_A = y_A, Y_B = y_B)$ incorporates the adjusted probabilities for all six joint outcomes and is the conditional probability.

This comprehensive joint framework captures the full complexity of sabre fencing, where scoring and penalty can occur simultaneously, sometimes resulting in multiple point awards within a single exchange.

Simulation Design and Parameter Justification:

To operationalize the theoretical model and explore scenarios, a discrete-event simulation was developed in Python (v3.10). The simulations of 10,000 bouts were implemented using the NumPy library for random number generation and the Matplotlib and Seaborn libraries for visualizations. The Python code was developed with the assistance of large language models to generate code for the simulation and is publicly available in a GitHub repository: <https://github.com/LynnAchamo/sabre-fencing-model>.

The baseline matchup uses $\lambda_A = 12$ and $\lambda_B = 8$. The relative advantage of 50% creates a clear but not overwhelming skill gap. Analytically, Fencer A wins 86% of bouts, leaving a 14% upset rate. This reflects the reality that no fencer is unbeatable. However, the parameters are illustrative, and λ for each fencer could be estimated from historical bout data.

Yellow and direct-red card probabilities were set at $p_y = 0.10$ and $p_r = 0.02$ per touch. Public datasets of sabre penalty rates are not available: for this reason, values are synthetic. They are set somewhat higher than elite-level bouts so that the effect of the state-dependent escalation is observable across 10,000 bouts, while keeping cards rare.

Every figure in the Results section is generated from the same baseline parameter set ($\lambda_A = 12$, $\lambda_B = 8$, $p_y = 0.10$, $p_r = 0.02$) and 10,000 bout simulations unless explicitly noted otherwise.

Sensitivity Analysis:

A sensitivity analysis will be conducted to examine how changes in fencer strength affect win probabilities.^{14,18} We define fencer strength through a relative scoring rate or the ratio of the two Poisson rate parameters $\rho = \lambda_A/\lambda_B$. This ratio is different from the probability that Fencer A wins a single scoring exchange described as $p_A = \lambda_A/(\lambda_A + \lambda_B) = r/(1+r)$, from which the race-to-15 win probability $P(\rho)$ follows the negative-binomial cumulative distribution (Appendix A):

$$P(A \text{ wins}) = P(\rho) = \sum_{k=0}^{14} \binom{k+14}{k} (1-p_A)^k p_A^{15}, \quad p_A = \frac{\rho}{1+\rho}$$

This relationship is nonlinear, where small changes in r near parity (i.e., when $\rho \approx 1$, meaning the fencers are relatively even) produce larger changes in $P(\rho)$, while the same amount of change far from parity has a much smaller effect.

Sensitivity is measured by $dP/d\rho$. Applying the chain rule:

$$\frac{dP}{d\rho} = \frac{dP}{dp_A} \cdot \frac{dp_A}{d\rho}, \quad \frac{dp_A}{d\rho} = \frac{1}{(1+\rho)^2}$$

For two evenly matched fencers ($\rho = 1$, $p_A = 0.5$), we have $dp_A/d\rho = 1/4$.

The derivative dP/dp_A can be calculated from the PMF of the negative binomial distribution at $p_A = 0.5$, and is approximately 4.380, giving $dP/d\rho \approx 1.10$. Thus, a 10% relative improvement in ρ near parity translates into an expected increase of 11% in win probability, confirming the large impact of small skill improvements in a 1:1 skill matchup.

To reflect the inherent variability of one bout, we construct a 95% confidence interval around the theoretical win probability. Treating the bout as a sequence of independent point-level trials, the effective sample size is the expected number of total exchanges:^{13,14}

$$n_{eff} = \frac{15}{p_A} + \frac{15}{1-p_A} - 1$$

The standard error of the win probability is $SE = \sqrt{P(1-P)/n_{eff}}$, and the resulting 95% interval for a single bout is $P \pm 1.96 SE$ (reported in Table 1).

A note is that this interval quantifies the “bout-level” confidence interval, rather than the Wilson confidence interval used later in the simulation validation. The Wilson interval captures the faithful representation of the empirical win proportion across 10,000 independent simulations, whereas the interval here computed reflects the uncertainty of a single bout under the same fixed parameters.

Simulation Validation and Statistical Methods:

To assess whether the discrete-event simulation faithfully represents the closed-form Poisson race model, we will perform a hypothesis test and construct a confidence interval for the simulated win probability. The theoretical win probability for Fencer A is derived from the negative binomial distribution (see “Poisson Race Model and Closed-Form Win Probability”) and serves as the null-hypothesis value p_0 . The simulation will be run for $N=10,000$ independent bout simulations for the base scenario ($\lambda_A = 12$, $\lambda_B = 8$) and will record the number of bouts won by Fencer A. Let $\hat{p}$ denote the simulated win proportion.

For the one-sample proportion test, the null hypothesis $H_0: p = p_0$ is tested against the two-sided alternative $H_A: p \neq p_0$. Because the number of trials is large and p_0 is not extreme, the test statistic given by

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

approximately follows a standard normal distribution under H_0 .¹⁴ The corresponding two-tailed p -value is $P(|Z| > |z|)$ to determine whether the true probability is different from p_0 (either larger or smaller). We use a significance level of $\alpha=0.05$, where failure to reject H_0 indicates that the simulation and the theoretical model are statistically consistent.

To quantify the uncertainty around $\hat{p}$, we construct a 95% confidence interval using the Wilson score method. The Wilson interval was selected over the Wald (normal approximation) interval due to its improvement over the normal approximation interval for handling proportions near 0 or 1 better and does not rely on the assumption that $\hat{p}$ is normally distributed when probability is skewed.¹⁷ The interval is given as:

$$CI = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}}$$

Here, $z_{\alpha/2} = 1.96$ for 95% confidence; that leaves the tail area of $\alpha/2$ on each side. An interval that contains the theoretical value p_0 provides further evidence that the simulation faithfully represents the theoretical model.

After running the simulation, the observed $\hat{p}$, z -score with its p -value, and Wilson confidence interval will be reported in the Results section. Additionally, we will compute Wilson intervals at the 90% and 99% level to examine robustness.

Model Assumptions and Limitations:

The model is built upon several key assumptions, where the consequences are important in interpreting results and limitations. These assumptions fall into two categories: those for the scoring process and those for the penalty events.

For the scoring process, the model assumes a fencer's scoring rate (λ) remains constant through the bout, without accounting for fatigue, momentum, or tactical changes. This means the model may underestimate the probability of comebacks and a performance decline that occur in real bouts. Additionally, each touch is treated as an independent event following a Poisson process, thereby ignoring psychological factors, namely confidence shifts after winning consecutive touches or frustration following losses.

The penalty model introduces a state-dependent logic through a discrete-time and discrete-state-space Markov process (i.e., Markov Chain), where the consequence of the card depends on the current penalty state.¹⁶ A key assumption of this Markov model is time-homogeneity, where the probabilities of a yellow card (p_y) and a direct red card (p_r) are constant for every touch. This modeling convenience provides a baseline but may understate real-world dynamics where a fencer, feeling pressured late in a bout, becomes more aggressive and prone to penalties, or a referee's card frequency changes.

However, this approach still assumes the fundamental probabilities of receiving yellow (p_y) and direct red (p_r) cards are constant and not influenced by scoring dynamics, fencer behavior, or referee biases. In reality, a fencer losing by a large margin may become more aggressive and susceptible to penalties, or a referee's card frequency may change as a bout progresses.

A fundamental limitation comes from the interaction between these two modeling approaches. The scoring model assumes memoryless, independent events, while the penalty model explicitly incorporates memory through state tracking. This hybrid provides a baseline representation but cannot capture more subtle interactions, notably how receiving a yellow card might affect a fencer's subsequent scoring probability.

Finally, the model treats the two fencers' processes as independent, neglecting adaptive strategies where a fencer may alter their tactics based on the opponent's habits. A high λ might be suppressed by effective defensive strategies from the opponent, and interaction is not accounted for here.

These limitations are not weaknesses but rather clear directions for future work, as discussed later in the conclusion, where paths for incorporating more dynamic, interactive elements into the modeling framework will be outlined based on these assumptions.

■ Results and Discussion

Validation of Poisson Race Model:

To confirm that the simulation faithfully implements the Poisson race model, we first compare its output to the theoretical win probability. For the baseline matchup ($\lambda_A = 12, \lambda_B = 8$), the theoretical win probability derived from the negative binomial model is 0.864 (see Methods). The simulation of 10,000 bouts produced an empirical win proportion $\hat{p} = 0.861$ for the stronger fencer. The close agreement between the theoretical and simulated win probabilities, a difference of 0.3 percentage points, validates that our Poisson race framework captures the core dynamics of a sabre bout.

To assess whether this discrepancy is statistically significant, a one-sample proportion test was performed as described in the Methods. The resulting z score was -0.758, yielding a two-tailed p-value of 0.448. At the standard significance level of $\alpha=0.05$, we fail to reject the null hypothesis that the true win probability equals the theoretical value. Therefore, the simulation and the analytical model are statistically consistent.

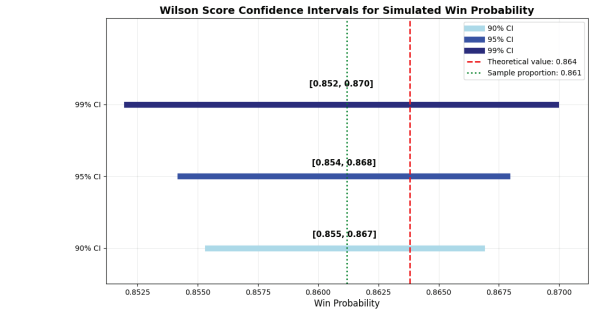


Figure 2: Wilson confidence intervals for simulated win probability ($\lambda_A = 12, \lambda_B = 8, 10,000$ bouts). The theoretical value 0.864 from the negative binomial formula falls inside all intervals, validating the simulation's consistency with the model.

Figure 2 displays Wilson confidence intervals for the simulated win proportion at 90%, 95%, and 99% confidence intervals. The 95% Wilson interval was [0.854,0.868], which contains the theoretical value of 0.864. The 90% and 99% confidence levels also include 0.864, showing consistency of simulation results across different statistical confidence levels.

These validation results demonstrate that a simple Poisson race model can serve as a reliable benchmark for sabre bout outcomes. If a fencer's scoring rate λ can be estimated from historical bout data, then the model offers a way to compare two fencers even when they have never faced each other. Because the win probability follows directly from the two λ values, the model provides a quantitative baseline against which real-world outcomes can be compared. While the current parameters are illustrative, the structure demonstrates how such a model could, with empirical data, help produce interpretable competitive bout expectations, where more complex psychological or tactical effects can later be measured.

Sensitivity Analysis:

Having established that the simulation faithfully reproduces the closed-form win probability, we next examine how that probability responds to changes in relative fencer strength. The table below reports the theoretical win probabilities and single-bout 95% confidence intervals for various ratios of ρ (derivation of sensitivity and confidence-interval setup is given in the Methods). Three patterns emerge from Table 1.

Table 1: Sensitivity analysis of win probabilities to scoring rate variations. A small increase in relative scoring rate yields substantial gains in win probability.

λ_A	λ_B	λ_A/λ_B	Advantage	Theoretical P(A wins)	Win Probability	95% CI
10	10	1.00	0%	0.500	50%	(0.422, 0.578)
11	10	1.10	+10%	0.602	60.2%	(0.527, 0.677)
12	10	1.20	+20%	0.690	69.0%	(0.619, 0.761)
12	8	1.50	+50%	0.864	86.4%	(0.812, 0.916)
14	10	1.40	+40%	0.819	81.9%	(0.760, 0.878)
15	9	1.67	+67%	0.916	91.6%	(0.876, 0.956)
8	12	0.67	-33%	0.136	13.6%	(0.078, 0.194)
9	11	0.82	-18%	0.293	29.3%	(0.210, 0.376)

First, the relationship between relative scoring rate and win probability is strongly nonlinear: a +10% advantage ($\rho=1.10$) increases win probability from 50% to 60.2%, while a +50% advantage ($\rho=1.50$) results in an 86.4% win probability. This nonlinear amplification effect, where small skill differences create substantial changes in win probability, occurs because the first-to-15 structure magnifies small per-exchange advantages.

Second, the most efficient range for competitive improvement is between parity and about a 20% advantage. The sensitivity of win probability to relative scoring rate at parity is $dP/d\rho \approx 1.10$ (Methods), meaning that near a 1:1 skill ratio, each 10% increase in ρ translates into an approximate 11% increase in win probability. Beyond a 50% advantage, the resulting increase in win probability diminishes sharply, creating a critical competitive threshold where further skill improvements add little to the probability of winning.

Third, the confidence intervals in Table 1 highlight the enormous bout-to-bout variability inherent in a race to 15. Even for evenly matched fencers ($P=0.500$), the 95% interval spans from 0.422 to 0.578, so a single bout is very noisy, indicating the underlying skill difference. This interval narrows as P approaches 0 or 1, but they remain wide enough that upsets are plausible even when there is a clear winner.

This analysis supports the paper's premise that a compact parametric model can capture a key competitive dynamic where small skill differences, when in the first-to-15 format, can create significant changes in expected outcomes. The model therefore provides a benchmark backing for discussing the importance of technical improvements, even if they are small, and explains why stronger fencers are not guaranteed victory in any bout.

Score Margin Distribution:

Sensitivity analysis reveals how win probability shifts with scoring rates, but it does not capture characteristics of the bout outcome. We therefore examine the distribution of winning and losing margins for the baseline matchup. We define a "decisive victory" as a win by four or more touches and a "close victory" as a win by three or fewer.

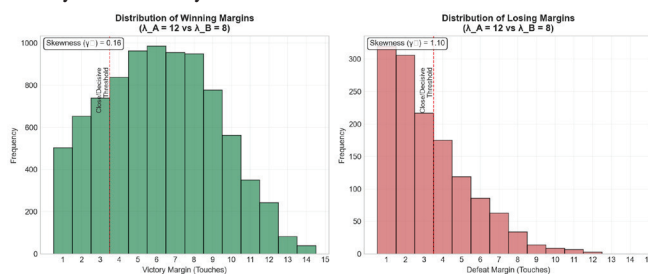


Figure 3: Distribution of victory and defeat margins for a stronger fencer ($\lambda_A = 12$ vs $\lambda_B = 8$) from 10,000 simulated bouts. Wins tend to be decisive (≥ 4 touches), while losses tend to be close (1-3 touches). This asymmetry shows that when a stronger fencer loses, it is most likely a narrow upset, while their victories are by a larger margin, underscoring how the first-to-15 format amplifies randomness.

Figure 3 displays the histograms of margins for wins and losses of the stronger fencer. When Fencer A won (85.83% of bouts), the victories were mostly decisive: 68% of wins were by a margin of four or more touches. In contrast, when the stronger fencer lost (14.17% of bouts), the defeats were close for an overwhelming number of times: 72% of losses came from a margin of one to three touches.

The sample skewness quantifies this contrast. For winning margins, Pearson's coefficient $\gamma=0.16$, indicating approximate symmetry. For losing margins, Pearson's coefficient $\gamma=1.1$, reflecting the right skewness visible in the distribution of Losing Margins (Figure 3), where most losses are narrow. The variance of the margin distribution is significant (35.340, see Table 2), a direct consequence of the Poisson assumption linking mean and variance and of the amplification introduced by the first-to-15 rule. It quantitatively demonstrates the high level of unpredictability in bout outcomes. Thus, even a decisive-looking 15 -10 win and a narrow 15 -14 loss are both plausible outcomes for the same skill matchup, purely due to random variation in the touches. This model can be used in the future

to examine whether an observed sequence of bout outcomes deviates from what randomness would produce, tracking progression of player growth or stagnation.

Penalty Analysis with State-Dependent Model:

Finally, we incorporate the penalty model to assess FIE's escalating card rules influences bout outcomes. All simulations in this section use the baseline scoring parameters and the penalty probabilities $p_y = 0.10$, $p_r = 0.02$. Although sabre bouts end when one fencer reaches 15 touches, the penalty-state model was simulated both under bounded (race-to-15) and extended (up to 30 total touches) conditions. Tables and figures are generated from 10,000 bounded bout simulations, unless noted. The unbounded results appear in Appendix B.

Table 2: Penalty Dispersion and Descriptive Statistics Bounded (race-to-15). Penalty events show high variance within the first penalty, typically occurring around touch 7.

Variable	Trials	Mean	Std. Dev.	Var
margin all	10000	4.496	5.945	35.340
margin penalty	9720	4.497	5.957	35.488
margin no penalty	280	4.458	5.504	30.293
yellow cards per bout	10000	2.909	1.696	2.963
red cards per bout	10000	0.438	0.586	0.368
first penalty touch	9720	7.337	3.112	11.972

Table 2 summarizes the number of penalties and the timing of the first penalty. Under the assumed baseline condition, an average bout contained 2.909 ± 1.696 yellow cards and 0.438 ± 0.586 red cards. The first penalty was observed, on average, at touch 7.337 ± 3.112 , indicating that most bouts saw a transition into a state with higher card-escalation risk (S_1 or T_0) within the first third of the bout. The high variance for both yellow cards ($\text{Var} = 2.963$) and red cards ($\text{Var} = 0.368$) indicates that penalties are not evenly distributed across bouts but instead are clustered randomly. This is a consequence of two factors: the underlying Bernoulli process that governs the issuing of cards, and the state-dependent escalation rules that transform the independent card events into a non-homogeneous process for scoring penalty points. While the probability of a referee giving a card on any single touch is constant (a Bernoulli trial), the probability that a card results in a point-winning red card is state-dependent. This combination of a memoryless trigger and memory-based consequence is what creates the high variance and random clustering observed in the data.

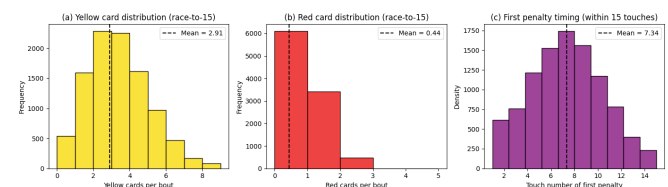


Figure 4: Marginal penalty distribution and first penalty timing under bounded (race-to-15) conditions ($\lambda_A = 12$, $\lambda_B = 8$, $p_y = 0.10$, $p_r = 0.02$, 10,000 bouts). Most first penalties occur in the middle stage of the bout, and red card distributions are heavily right-skewed. This non-uniformity is a direct consequence of combining the memoryless card issuing with state-dependent escalation.

Figure 4 shows the distribution of the first penalty touch and the yellow and red card count per bout. The first penalty timing distribution is unimodal with a peak in the middle phase of the bout. Red cards are heavily right-skewed: 62% of bouts have none, 25% have one, and 13% have two or more. This asymmetry comes naturally from the threshold rule.

Table 3: Penalty Win Probabilities Bounded (race-to-15). Penalties reduce the stronger fencer's win probability, especially in bouts where penalties occur, revealing high conditional impact despite low frequency.

Variable	Probability
win prob. (with penalties)	0.857
win prob. (no penalties)	0.864
win prob. (penalty bouts only)	0.835

Introducing penalties into the Poisson race model simulation lowered the stronger fencer's overall win probability from the penalty-free theoretical value of 0.864 to 0.857 (Table 3). While the reduction is modest in absolute terms (-0.8%), the impact becomes significant when conditioned on the presence of any penalty. In 97.2% of bouts where at least one yellow or red card was issued, Fencer A's win probability dropped to 0.835, a decrease of 3.4% relative to the penalty-free case. Thus, penalties function as high-impact, low-frequency events. While the probability of a penalty point due to a red card is rare on any single touch, when one does occur, it has a meaningful impact on the bout's trajectory, rather than being a negligible background factor.

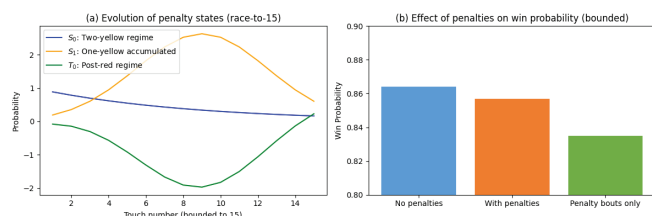


Figure 5: Markov penalty state probabilities over first 30 touches, computed analytically from the state transition equation with $p_y = 0.10$, $p_r = 0.02$. The initial state S_0 decays exponentially, while the post-red regime T_0 rises steadily, showing how penalty risk accumulates.

Figure 5 presents the evolution of Markov state probabilities across consecutive touches. The probability of remaining in the safe two-yellow regime (S_0) decreases exponentially from 1.0 to 0.04 by touch 30. The intermediate one-yellow state (S_1) peaks around touches 8-10 at approximately 0.38, and the post-red regime (T_0) rises steadily to 0.58. By the latter half of the bout, more than half of the fencers operate under the heightened one-yellow threshold, where any yellow card immediately results in a penalty point. This quantitatively captures the intuition that early warnings create long-lasting, escalating danger.

From a practical standpoint, the results could be used by coaches to underscore the importance of discipline from the first touch and through a bout. Additionally, the analysis supports the paper's central argument. By treating a bout as a stochastic race with state-dependent rules, we can move beyond simple outcome descriptions and begin to quantify how skill, risk, and rules interact to create the dynamics we see in competitive sabre fencing. From a modeling standpoint, these results demonstrate how integrating even a simple three-state chain can reproduce the cumulative escalation rules observed

in actual competition, providing a statistically interpretable bridge between rules and bout dynamics.

■ Conclusion

This paper presented an integrated probabilistic framework for modeling sabre fencing bouts using discrete stochastic processes. The Poisson race model, supported by negative binomial analytical results and large-scale Python simulations, demonstrated that fencers' scoring rates can accurately predict bout-level outcomes. Validation tests confirmed close agreement between theoretical and simulated win probabilities, with an absolute difference of less than 0.3 percentage points (0.864 theoretical vs 0.862 simulated). Incorporating the Markov-based penalty model extended this analysis to rule-dependent dynamics, revealing that although penalties are rare, their cumulative nature introduces asymmetry and a measurable reduction in win probability.

The primary contribution of this work is establishing a foundational, parametric model for Sabre, similar to how the early 'xG' models revolutionized soccer analytics. While this model excels at quantifying the baseline interplay of skill and discipline, its simplifying assumptions mean it does not directly capture psychological factors, namely, momentum or frustration. However, this is also its strength, as it provides a neutral expectation, where it becomes a benchmark against which these more complex effects can be identified and measured. For instance, a fencer who consistently outperforms their predicted win probability during a competition may have exceptional psychological resilience or tactical creativity under pressure.

The findings offer several practical insights for the fencing community. For the coaches and athletes, sensitivity analysis reveals a strong nonlinear relationship, where small, achievable improvements in scoring rates (e.g., 10-20%) can result in a large boost for win probability. This emphasizes the value of refining technique with a wide range of possibilities for development. Furthermore, the penalty model offers a concrete tool for teaching discipline management, visually showing how an early yellow card can significantly increase a fencer's vulnerability for the entire bout. For referees and officials, the state-dependent penalty model provides a quantitative perspective of the set rules of fencing. It underscores that a yellow card is not just an isolated warning but a fundamental shift in the bout's risk structure, potentially influencing the final bout outcome.

Looking forward, this framework opens many directions for future research. The most immediate step is adding empirical data: collecting historical bout data to estimate real-world scoring rates and penalty frequencies, then performing goodness-of-fit tests to compare simulated and actual victory margin distributions. The model can be extended to other fencing disciplines (foil and epee) by adjusting to time limits and incorporating discipline-specific rules. Further study could incorporate probabilities that change over time to reflect referee behavior or psychological effects (e.g., increased card frequency when a fencer is losing), and could also examine whether penalties disproportionately affect weaker fencers in

actual competition. Integrating this parametric approach with machine learning and motion analysis could also improve both accuracy and tactical insight.

The clearest takeaway from this work is that classical probability models, even with simple assumptions, can offer a useful framework for understanding strategy, risk, and outcome patterns in competitive sabre fencing. By treating a bout as a stochastic race with state-dependent rules, we can move beyond just outcome description to a deeper analysis of how skill, risk, and rules interact to shape bout dynamics, providing a foundation for the future of fencing analytics.

Acknowledgments

I wish to express my sincere gratitude and thanks to Adam Almakroudi of Imperial College London for his mentorship. Working on this research and learning was truly fun, and I am very happy to have the chance to explore and expand upon my passions.

References

- Baumer, B.; Zimbalist, A. *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*; University Of Pennsylvania Press, 2014.
- Campbell, M.; Hong, S. Predicting the Outcome of Sports Competitions Using Poisson Distribution. *Issues in Information Systems* **2024**, *25* (1), 188–198. https://doi.org/10.48009/1_iis_2024_116.
- Everson, P.; Goldsmith-Pinkham, P. S. Composite Poisson Models for Goal Scoring. *Journal of Quantitative Analysis in Sports* **2008**, *4* (2). <https://doi.org/10.2202/1559-0410.1107>.
- Skinner, B.; Guy, S. J. A Method for Using Player Tracking Data in Basketball to Learn Player Skills and Predict Team Performance. *PLOS ONE* **2015**, *10* (9), e0136393. <https://doi.org/10.1371/journal.pone.0136393>.
- Shirley, K. *A Markov Model for Basketball*; New England Symposium for Statistics in Sports, 2007. https://nessis.org/nessis07/Kenny_Shirley.pdf (accessed 2025-10-03).
- Bottoms, L.; Tarragó, R.; Muñoz, D.; Chaverri, D.; Irurtia, A.; Castizo-Olier, J.; Carrasco, M.; Rodríguez, F. A.; Iglesias, X. Physiological Demands and Motion Analysis of Elite Foil Fencing. *PLOS ONE* **2023**, *18* (2), e0281600. <https://doi.org/10.1371/journal.pone.0281600>.
- Chen, T. L.-W.; Wong, D. W.-C.; Wang, Y.; Ren, S.; Yan, F.; Zhang, M. Biomechanics of Fencing Sport: A Scoping Review. *PLOS ONE* **2017**, *12* (2), e0171578. <https://doi.org/10.1371/journal.pone.0171578>.
- Zhang, M.; Chen, L.; Yuan, X.; Huang, R.; Liu, S.; Yong, J. Visualization of Technical and Tactical Characteristics in Fencing. *Journal of Visualization* **2018**, *22* (1), 109–124. <https://doi.org/10.1007/s12650-018-0521-3>.
- Lin, Z.; Goel, P.; Yun, J.; Liu, K.; Araujo, J. P. VirtualFencer: Generating Fencing Bouts Based on Strategies Extracted from In-The-Wild Videos. **2025**. <https://doi.org/10.48550/arXiv.2507.00261>.
- Chang, C.-C.; Lee, T.-L. Fencing Tactics Analysis in Broadcast Video: A Point-By-Point Analytical System. *IEEE Transactions on Circuits and Systems for Video Technology* **2019**, *29* (7), 2162–2175. <https://doi.org/10.1109/tcsvt.2018.2855098>.
- Zeng, Y. *Modeling Player Strength in Fencing*; The University of Iowa https://homepage.divms.uiowa.edu/~dzimmer/sports-statistics/slides_yiran.pdf (accessed 2025-10-03).
- Haight, F. A. *Handbook of the Poisson Distribution*; John Wiley & Sons. New York, NY, US, 1967.
- Ross, S. M. *The Exponential Distribution and the Poisson Process. In Introduction to Probability Models (Tenth Edition)*, Academic Press, 2010; pp. 291–370.
- Casella, G.; Berger, R. L. *Statistical Inference*, 2nd ed.; Thomson Learning: Belmont, CA, 2002; pp. 92–97.
- Fédération Internationale D'Esclime. *The Penalties and the Competent Juridical Authorities*. In Technical Rules; FIE, 2025; pp. 158–170.
- Ross, S. In Introduction to Probability Models, Eleventh Edition.; Academic Press, 2014.
- Wallis, S. Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods. *Journal of Quantitative Linguistics* **2013**, *20* (3), 178–208. <https://doi.org/10.1080/09296174.2013.799918>.
- Saltelli, A.; Tarantola, S.; Campolongo, F.; Ratto, M. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*; John Wiley & Sons, 2004; Vol. 1.

Author

Lynn Uemoto is a Junior at Franklin School (New Jersey, United States), with research interests in applied mathematics and statistics. She plans to pursue a degree in mathematics or statistics and continue exploring quantitative and risk analysis.

Appendix A: Derivation of the Negative Binomial Distribution

This appendix provides the mathematical derivation of the negative binomial distribution from first principles.

Imagine a sequence of independent Bernoulli trials (e.g., coin flips), each with a probability of success p and failure $1-p$. Let the random variable X represent the number of failures observed before the r -th success. The event $\{X=k\}$ occurs if and only if exactly $r-1$ successes occur in the first $k+r-1$ trials, and the $(k+r)$ -th trial is a success.

The number of ways to arrange $r-1$ successes and k failures in $k+r-1$ trials is given by the binomial coefficient:

$$\binom{k+r-1}{k} = \frac{(k+r-1)!}{k!(r-1)!}, \text{ or equivalently } \binom{k+r-1}{r-1}$$

The probability of any specific sequence with $r-1$ successes and k failures is:

$$p^{r-1}(1-p)^k$$

Therefore, the probability of getting the $(r-1)$ -th success exactly on the $(k+r-1)$ -th trial is:

$$\binom{k+r-1}{k} p^{r-1}(1-p)^k$$

We now need one more success on the next trial, or the final trial. The probability of this is p . Since the trials are independent, the joint probability is the product of the two probabilities above:

$$P(X=k) = \binom{k+r-1}{k} p^{r-1}(1-p)^k \cdot p$$

Simplifying the exponents of p gives us the classic PMF of the Negative Binomial distribution:

$$P(X=k) = \binom{k+r-1}{k} p^r (1-p)^k, \text{ for } k = 0, 1, 2, \dots$$

■ Appendix B: Unbounded Penalty Analysis

This appendix presents the results of the Markov state-dependent penalty model under unbounded conditions, where the simulation was run for a cumulative 29 touches (the maximum number of exchanges in a 15-14 bout). This is different from the bounded (race-to-15) model in the main paper. The unbounded model removes the variability in total touches, allowing us to observe the full evolution of the penalty state probabilities and get a smoother estimate of long-term penalty dynamics.

Results and Comparison with Bounded Model:

Under the same baseline conditions ($p_y = 0.10$, $p_r = 0.02$), the 10,000 unbounded simulations over 29 touches produced the following averages: 2.995 ± 1.722 yellow cards, 0.494 ± 0.609 red cards per bout (Table B1). The mean first penalty occurred at touch 7.78 ± 3.5 . As expected, the average number of yellow cards and red cards is higher than in the bounded model (2.909 ± 1.696 and 0.438 ± 0.586 , respectively), because bouts are artificially extended to their maximum possible length. The conditional win probabilities (Table B2) revealed a more significant effect in unbounded conditions, with Fencer A's win probability in penalty bouts at 0.832 compared to 0.864 in a case without penalties.

Table B1: Penalty Dispersion and Descriptive Statistics Unbounded (total 29 points). Unbounded simulations show higher average penalties than bounded bouts, confirming that longer exchanges increase the likelihood of penalty points.

Variable	Mean	Std. Dev.
yellow cards per bout	2.995	1.722
Red cards per bout	0.494	0.609
bouts with ≥ 1 penalty	97.14	N/A
touch of the first penalty	7.780	3.500

Table B2: Penalty Win Probabilities Unbounded (total 29 points). The win probability reduction due to penalties is slightly more pronounced in unbounded simulations, highlighting the increasing influence of penalty escalation in extended bouts.

Variable	Probability
win prob. (with penalties)	0.858
win prob. (no penalties)	0.864
win prob. (penalty bouts only)	0.832

The distribution of red cards (Figure B1) remains heavily right-skewed, but the proportion of bouts with two or more red cards increases compared to the bounded model. This demonstrates that if bouts were to consistently run longer, the cumulative penalty rules would lead to a higher frequency of point-awarding red cards. The state dynamics (Figure B2) show the clear, long-term trajectory that is truncated in a real bout. Compared to Figure 5 with the bounded model.

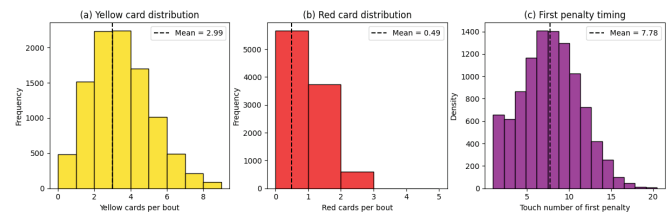


Figure B1: Marginal Penalty Distribution and Timing under unbounded conditions (total 29 points). Extending the bout length increases the frequency of red cards.

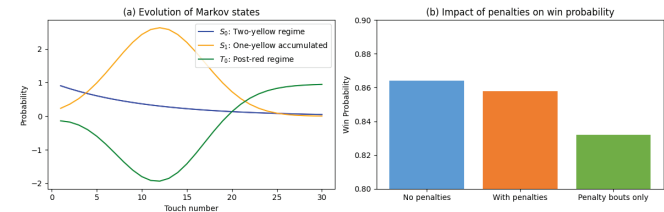


Figure B2: State Dynamics and Win Probability are unbounded (not depending on JPMF). Under unbounded conditions, penalty state evolution is smoother and more pronounced, illustrating the long-term trajectory of escalating card risk.

Comparing Visual and Working Memory of Teenage and Elderly Populations

Prathiksha Raguram

J. L. Mann High School, 160 Fairforest Way, Greenville, South Carolina, 29607, USA; prathiksha.raguram08@gmail.com

ABSTRACT: Teenage participants aged 14 to 18 and elderly participants aged 65 and above completed a survey to gauge their visual and working memory abilities. Each individual's visual memory was measured through their proportion of accurately identified fragmented images, while working memory was measured through accurate recognition of a previously shown sequence of numbers. The teenage population had a significantly higher proportion of accurately identified fragmented images than the elderly population, suggesting teenagers have more efficient visual memories. While identifying fragmented images, participants were also asked to rank their confidence in recognition on a scale of 1 to 10. The difference between each population's average confidence was not statistically significant. However, the correlation between confidence and accurately identified fragmented images was significant within the teenage population, yet nonsignificant within the elderly population. The difference in the proportions of accurately identified number sequences was not statistically significant, suggesting that varying working memory efficiency between the two populations was also not statistically significant. Future research should include more questions regarding the working memory aspect to increase the validity of the results.

KEYWORDS: Behavioral and Social Sciences, Cognitive Psychology, Neuropsychology, Short-Term Memory, Cognitive Aging.

■ Introduction

The average brain of a newborn baby contains over 100 billion neurons,¹ a feat of nature that enables humans to facilitate a plethora of complex processes, including but not limited to thinking, decision-making, language interpretation and formation, sensation, perception, and memory. Throughout childhood and adulthood, the brain rapidly processes and stores unfamiliar stimuli, making way for the burgeoning formation of neural pathways while neuroplasticity is at its prime. As the brain ages, however, it diminishes in a variety of aspects, including brain volume loss, neuron cell shrinking, and demyelination.² This physical deterioration is reflected through diminished mental efficiency: as reported by the Centers for Disease Control and Prevention (CDC), approximately 1 in 10 adults ages 45 and older reports worsening memory loss or cognitive decline.³

Significantly, a common side-effect of the aging brain is its impact on short-term memory, often measured through decreased processing and executive cognitive functioning speeds.⁴ The Cleveland Clinic describes short-term memory as a “temporary storage space that holds information you learned,”⁵ albeit only a limited amount of information for a relatively short duration (seconds or minutes). Controlled by the prefrontal cortex and the hippocampus, short-term memory can be categorized into distinct types, typically depending on the type of information stored. Some examples include visual memory, which stores images; echoic memory, which stores sounds; and short-term spatial memory, which stores spatial relations between objects, with each type of memory lasting varying durations. Certain types of memory decline as the

brain ages, such as episodic memory, while others remain relatively unchanged, like procedural memory.⁶ This study focuses specifically on visual memory and a type of memory similar to short-term memory called working memory. Working memory is “the short-term maintenance and manipulation of information necessary for performing complex cognitive tasks such as learning, reasoning, and comprehension.”⁷ Through the use of short-term memory, individuals can navigate everyday activities successfully, seamlessly learning, comprehending, and solving tasks that require immediate attention.⁸

Historically, several techniques have been developed to measure short-term memory.⁹ In 1913, surgeon Howard A. Knox developed an assessment used to examine immigrants at Ellis Island called the Cube Imitation Test, where participants had to repeat a pattern tapped on four cubes by an examiner.¹⁰ The Cube Imitation Test was superseded in 1972 by Canadian psychologist Philip Michael Corsi's block-tapping test. In Corsi's block-tapping test, nine wooden blocks are quasi-randomly spaced out on a flat surface, with numbered faces only facing the examiner and not visible to the participant. The participant would then copy an arbitrary sequence tapped out by the examiner, with difficulty increased by altering the length of the sequence.¹¹ While these techniques measure a participant's visual short-term memory, assessments like the Digit Span Backward test, where an individual repeats a series of digits in reverse order, gauge working memory.¹² In recent decades, researchers have been experimenting with new metrics for short-term memory, including exploring the connection between memory efficiency and the development of disorders like Alzheimer's.

The purpose of this study is to compare teenagers' and elderly individuals' perceived accuracy and confidence in their own visual and working memories.

■ Methods

Participants:

This study focuses on two populations with distinct age ranges. The youth population consisted of 98 participants aged 14 to 18 years old, while the elderly population consisted of 50 participants aged 65 years old and above. The inclusion criteria for this study were any individual between 14 and 18 years old or 65 and above; the exclusion criteria for this study were anyone outside of these age ranges. All data were collected by the student researcher: the teenage responses were collected through a voluntary response survey distributed at a high school in Upstate South Carolina, while the elderly responses were primarily collected through a local senior center. This primary analysis study aimed to evaluate and compare the short-term visual and working memories of these populations in order to determine whether younger individuals have significantly more accurate short-term memories than elderly individuals.

Study Design:

Visual memory. The survey was adapted from Anne Cleary *et al.*'s 2004 study, "Recognition without picture identification: Geons as components of the pictorial memory trace", in which participants were asked to: identify the item depicted by a set of fragmented lines (see Figure 1); identify whether the picture fragment belonged to an image they were previously shown; and rate on a scale of 0 to 10 the likeliness that the picture fragments originated from an image at study.¹³

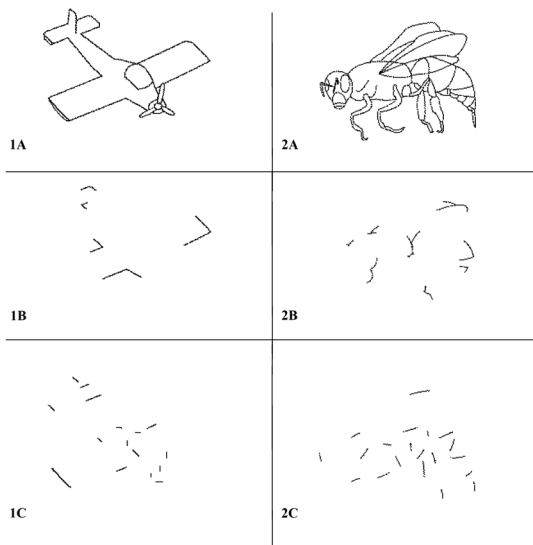


Figure 1: Sample stimuli. Sample stimuli from the original study. 1A and 2A depict the original images, with 1B and 2B depicting the geon fragments, and 1C and 2C depicting the non-geon fragments (Cleary *et al.*, 2004).

This observational study utilized voluntary responses submitted through a Google Form. This study adapted the 2004 study by incorporating "filler" elements, which were used to gauge the participants' natural recollection ability when not

initially prompted to remember a stimulus. The theory was that, when these seemingly unrelated questions were interspersed throughout the survey, the participant would be less likely to dedicate energy specifically to remembering any one. These filler elements were presented as general knowledge questions, such as "Who was the first president of the United States of America?" and "What is 3×7 ?"

Among the filler elements included were questions asking the participant to identify an unfragmented image from a drop-down of nine answer choices ('Horse,' 'Car,' 'Turkey,' 'Airplane,' 'Octopus,' 'Parrot,' 'Helicopter,' 'Seahorse,' & 'Hummingbird'). There were a total of five unfragmented images, shown in the following order: Airplane, Turkey, Parrot, Seahorse, Helicopter.

Following the initial survey sections containing filler questions, the participant was then directed to the Fragmented Image section (see Figure 2). In this section, each prompt depicted a set of lines and asked the participant to rate on a scale of 1 to 10 the familiarity of the fragmented image. Four out of six of the fragmented images provided were derived from images shown to the participant earlier in the survey, but the remaining stimuli (serving as red herrings) were derived from unfragmented images not previously shown to the participant. The fragmented images were shown in the following order: red herring, Seahorse, red herring, Turkey, Airplane, Parrot.



Figure 2: Unfragmented and fragmented image comparison. On the left is the unfragmented Parrot; on the right is the fragmented Parrot. Below the fragmented image, the 1-10 scale was provided for participants, which was used to indicate the participants' confidence in recognizing the image.

The participant was then asked to identify each fragmented image to the best of their ability without returning to a previous section. The provided drop-downs were identical to the ones provided earlier in the survey, with an additional tenth choice stating, "I have not seen this image," ideally to be used for the red herrings.

Working memory. Among the unfragmented images and filler questions, there were two questions regarding a series of numbers used to evaluate the short-term working memory of the participants, the secondary purpose of this study. The first question asked the participant to identify an item in a sequence of numbers, in order to subtly introduce the numbers to the participants: "In the given sequence {2, 3, 9, 5, 7, 1}, what is the *second* number?" Later in the form, the participant was asked to recall the sequence they saw earlier, including the correct order of the numbers.

Outcome Measures:

Demographic information was collected, including participants' highest level of education, preferred gender orientation, race, and whether they have been diagnosed with any memory-affecting conditions, such as concussions, traumatic brain injuries, or Alzheimer's disease. Certain demographics, such as education and memory-affecting conditions, were collected due to potential associations they may have with recognition and recall ability. For instance, those who have taken higher levels of education may exhibit better visual and working memory compared to those who have not. Furthermore, memory-affecting conditions are likely to have an impact on a participant's ability to recall and recognize visual and numeric components of the survey.

Visual memory was used in this study as "the capacity to remember what has previously been seen in the form of visual images".¹⁴ Visual memory was captured through the usage of the unfragmented and fragmented images. For the unfragmented images, participants selected the answer choice they felt best identified the image from a drop-down menu. During the fragmented image section, participants first indicated on a scale from 1 to 10 how familiar the fragmented image was, which suggested their confidence in recognizing the image. After that, they identified the image to the best of their abilities by selecting an answer choice from a similar drop-down menu, only with an additional option stating, "I have not seen this image."

Working memory was used in this study as "the short-term maintenance and manipulation of information necessary for performing complex cognitive tasks such as learning, reasoning, and comprehension."¹⁷ Working memory was captured by the participants' retention of a given sequence of numbers. The participants were initially introduced to the sequence by identifying the second item, and were later asked to identify which sequence from five answer choices was the same as the one they were shown earlier.

Data Analysis:

Descriptive statistics, such as mean, median, and mode, were collected for the following demographics: participants' highest level of education, preferred gender orientation, race, and whether they have been diagnosed with any memory-affecting conditions, such as concussions, traumatic brain injuries, or Alzheimer's.

Visual Memory. For this study's primary purpose of comparing visual memories of the two populations, Levene's Test for Equality of Variances was utilized to determine the skewness of the results. Then, a two-sample t-test was performed in order to determine whether the difference in accuracy rates of the two populations was statistically significant.

Working Memory. For this study's secondary purpose of comparing working memories of the two populations, a two-proportion z-test was used to determine the statistical significance of the differing population proportions of those who correctly identified the sequence of numbers.

Correlation between average confidence and accuracy. Additionally, Levene's Test and a two-sample t-test were performed

again to determine whether each population's average confidence indicated by the recognition scale for each fragmented image was significantly different. Correlation plots were also created to visualize the association between average confidence rates and average accuracy rates of each population.

Results and Discussion

Demographics:

98 responses were collected from the 14-18-year-old teenage population, with an average age of 15.71 ± 1.06 years. 50 responses were collected from the 65- 93-year-old elderly population, with an average age of 72.92 ± 6.02 years. Notably, over 80% of both populations reported having been diagnosed with no memory-affecting conditions, while 11.22% of teens reported having been diagnosed with a concussion at some point, and only 6% of elderly individuals reported having been diagnosed with stroke, dementia, or multiple sclerosis. Additional demographics, including preferred gender orientation and race, were collected, as seen in Table 1.

Table 1: Demographics table. Demographics collected included age, preferred gender orientation, race, and memory-affecting conditions. Most notably, more than 80% of both populations reported having been diagnosed with no memory-affecting conditions.

Demographic Category	Sub-categories	Percentage (Count)	
		14-18 years	65+ years
Age	Avg. Years $\pm$ Std. Dev. (Range)	15.71 $\pm$ 1.06 (14-18)	72.92 $\pm$ 6.02 (65-93)
Preferred Gender Orientation	Male	39.80 (39)	48.00 (24)
	Female	53.06 (52)	46.00 (23)
	Nonbinary	2.04 (2)	0
	Other (i.e., genderqueer, transgender)	3.06 (3)	0
	Missing data	2.04 (2)	6.00 (3)
Race	White	57.14 (56)	46.00 (23)
	Black or African American	22.45 (22)	8.00 (4)
	Asian	10.20 (10)	46.00 (23)
	Hispanic	7.14 (7)	0
	Other (i.e., Brazilian, MENA, Mixed Race)	3.06 (3)	0
Memory-Affecting Conditions	None	83.67 (82)	88.00 (44)
	Concussion	11.22 (11)	0
	Multiple Sclerosis	0	2.00 (1)
	Stroke / Dementia	0	2.00 (1) - Stroke & Dementia 2.00 (1) - Dementia only
	ADHD / ADD	2.04 (2)	2.00 (1)
	Other (i.e., Major Depressive Disorder, lack of sleep)	2.04 (2)	0
	Missing Data	1.02 (1)	4.00 (2)
	TOTAL	98	50

Comparison of Accurately Identified Fragmented Images

To evaluate the differences between the visual memories of teenagers ages 14-18 and elderly individuals ages 65 and older, equality of variances was first tested using Levene's Test for homogeneity. The F-test indicated no significant difference in variance between the two groups, $F(1, 146) = 1.42$, $p = 0.18$, suggesting that the assumption of equal variances was met.

A standard independent two-sample t-test was then conducted to compare the percentages of accurately identified fragmented images between participants in the teenage popu-

lation ($n = 98, 0.39 \pm 0.20$) and those in the elderly population ($n = 50, 0.29 \pm 0.17$), as seen in Figure 3. Results revealed a **statistically significant** difference in the proportion of accurately identified fragmented images between the two groups, $p = 0.00$, Cohen's $d = 0.54$, indicating a moderate effect size in favor of the teenage population. These results suggest that teenage participants aged 14-18 have significantly more accurate visual memories compared to those of elderly individuals aged 65 and older.

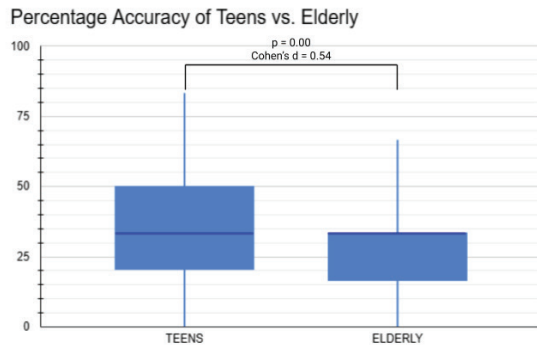


Figure 3: Percentage Fragmented Image Accuracy of Teens vs. Elderly. This boxplot compares the percentage of accurately identified fragmented images between the teenage population and the elderly population. This difference is statistically significant, with $p = 0.00$. Notably, the medians of both populations are equivalent, at 33.33%.

Comparison of Working Memories:

A two-proportion z-test was conducted to compare the working memories of the teen population and the elderly population. In the teenage group, 33 out of 98 participants (34%) accurately recognized the sequence of numbers, compared to 18 out of 50 participants (36%) in the elderly group. The difference in proportions was **not statistically significant**, $z = -0.28$, $p = 0.60$, indicating that varying working memory efficiency between these two populations is also not statistically significant.

Confidence in Fragmented Image Recognition:

To compare the average confidence of the two populations, equality of variances was first tested using Levene's Test for homogeneity. The F-test indicated unequal variance between the two groups, $F(1, 146) = 2.43$, $p = 0.00$.

Due to the unequal variance, an unequal-variance Welch's two-sample t-test was conducted to compare the average confidence on a scale of 1 to 10 of the teenage group ($n = 98, 4.74 \pm 1.92$) and of the elderly group ($n = 50, 5.01 \pm 2.98$). Results revealed a difference that was **not statistically significant** between the average confidence of the two groups, $p = 0.57$, Cohen's $d = -0.12$, indicating a small effect size.

Correlation of Confidence and Accurately Identified Fragmented Images:

A Pearson correlation was conducted to examine the relationship between confidence, indicated on a scale of 1 to 10, and the accuracy of identifying fragmented images in the teenage population. Results showed a **significant** moderate, positive correlation between the two variables, $r(97) = 0.42$, $p = 0.00$. This suggests that participants who indicated greater

confidence in recognizing the fragmented image are associated with greater accuracy. See Figure 4.

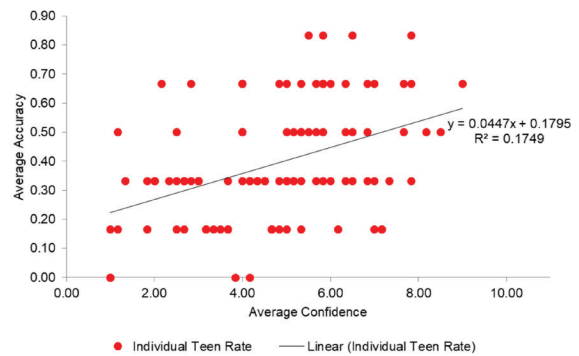


Figure 4: Correlation of confidence and average accuracy rate in the teenage population. This correlation is statistically significant. Results showed a moderate, positive correlation between the two variables.

A Pearson correlation was conducted to examine the relationship between confidence indicated on a scale of 1 to 10 and the accuracy of identifying fragmented images in the elderly population. Results showed a **nonsignificant**, weak, negative correlation between the two variables, $r(49) = -0.09$, $p = 0.55$. This suggests that participants who indicated greater confidence in recognizing the fragmented image are associated with lower accuracy rates. See Figure 5.

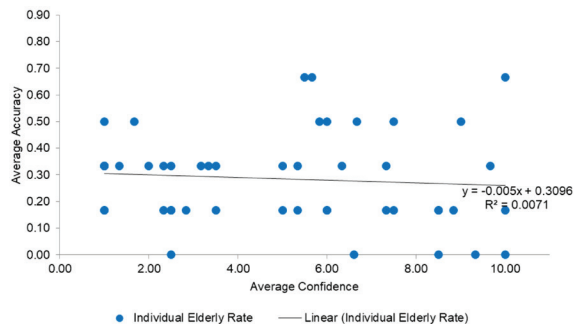


Figure 5: Correlation of confidence and average accuracy rate in the elderly population. This correlation is not statistically significant. Results showed a weak, negative correlation between the two variables.

Limitations:

This study is limited by its sampling technique, utilization of a voluntary response survey, inclusion of participants with memory-affecting conditions, and the testing method of working memory. Participants were primarily based in local senior centers and a local high school, possibly affecting the generalizability of the results. Due to the use of a Google Form survey, it is possible that participants may have returned to previous sections of the Form. Self-reported information could potentially lead to skewed results. Additionally, the inclusion of participants with memory-affecting conditions, such as concussions, ADHD, or stroke, may have skewed the findings. Finally, this study's objective to compare the working memories of two populations may be limited by the fact that there was only one pair of questions pertaining to this objective.

Overall, these limitations may affect the generalizability and validity of the findings. As a result, further research and exploration are recommended to develop more definitive con-

clusions. Even so, this research remains significant despite the limitations because it provides valuable insight regarding changes in visual and working memory throughout the aging process, inviting new understandings of how short-term memory is affected by age.

■ Conclusion

In summary, this study compared the visual and working memories of teenagers and elderly individuals. The findings demonstrate that there was a **statistically significant** difference in the proportion of accurately identified fragmented images between the two populations, suggesting that teenagers may have better visual memories than elderly individuals. This finding is supported by Ebaid & Crewther's 2019 study: "A decline in cognitive functions, including ... speed of visual information processing [and] working memory ... associated with the normal aging process is commonly reported in the scientific literature."¹⁵ It is reasonable to infer that the typical cognitive decline associated with aging is responsible for this difference in short-term visual memories.

The difference in proportions of accurately identified number sequences was not statistically significant, suggesting that differences in working memory efficiency of these populations may not be significant either. Interestingly, Freitas *et al.*'s 2007 study states, "Numerous studies have suggested that performance in [complex cognitive] tasks [in context of working memory] is related to age, where older adults have a lesser performance than the young."¹⁶ When considering this, a possible explanation for the nonsignificant finding could be the limitation that working memory is only tested using one pair of questions in the survey.

When comparing the average confidence of the two populations, the results revealed that the differences were **not statistically significant**. Curiously, however, a moderately strong positive correlation between the variables of confidence and accurately identified fragmented images was found specifically within the teenage population, while the elderly population indicated a **nonsignificant** weak negative correlation between the two variables. This could indicate that teenagers exhibit better recall when they have more confidence in their memory abilities, possibly due in part to the fact that younger people typically do not experience memory issues.

In future research comparing the working memories of these two populations, the researchers should incorporate several test questions to increase the validity of their results. Additionally, future studies should more closely examine the relationship between confidence and recall ability, particularly in populations that have been diagnosed with memory-affecting conditions like Alzheimer's or strokes.

New findings regarding how different types of short-term memory change throughout the aging process could offer crucial insight into the development of serious memory-affecting conditions like dementia and Alzheimer's. Additionally, they could enhance current understanding of how the role of memory in disorders such as ADHD, where the brain's chemistry is regulated in a different way than typical, or Major Depressive Disorder, where there is oftentimes a chemical imbalance in

the brain, deviates from expectation. These insights could shed light on various aspects of society, from reforming standardized testing accommodations so that they more efficiently aid students to informing current treatments of memory conditions.

■ Acknowledgments

I attest that the ideas, graphics, and writing in this paper are entirely my own. I am very thankful to Indigo Research for their support of this study, and I am specifically indebted to Dr. Amanda Rounds, PhD, whose guidance and support were invaluable as I developed and navigated this research. A special thank you to my AP Statistics teacher, Mr. Zachary Sondrini, who guided me with the statistical analyses utilized in this paper. Finally, I am thankful to my parents, who helped me greatly in the data collection process.

■ References

1. Ackerman, S. *The Development and Shaping of the Brain*; National Academies Press (US), 1992.
2. Blinkouskaya, Y.; Caçoilo, A.; Gollamudi, T.; Jalalian, S.; Weickenmeier, J. Brain Aging Mechanisms with Mechanical Manifestations. *Mechanisms of Aging and Development* 2021, 200, 111575. <https://doi.org/10.1016/j.mad.2021.111575>.
3. CDC. *Cognitive Health and Caregiving*. Chronic Disease Indicators. <https://www.cdc.gov/cdi/indicator-definitions/cognitive-health-caregiving.html>.
4. Murman, D. The Impact of Age on Cognition. *Seminars in Hearing* 2015, 36 (03), 111–121. <https://doi.org/10.1055/s-0035-1555115>.
5. Cleveland Clinic. *Short-Term Memory: What It Is, How It Works & Duration*. Short-Term Memory. <https://my.clevelandclinic.org/health/articles/short-term-memory>.
6. Nyberg, L.; Lövdén, M.; Riklund, K.; Lindenberger, U.; Bäckman, L. Memory Aging and Brain Maintenance. *Trends in Cognitive Sciences* 2012, 16 (5), 292–305. <https://doi.org/10.1016/j.tics.2012.04.005>.
7. APA Dictionary of Psychology. *Working Memory*. APA Dictionary of Psychology. <https://dictionary.apa.org/working-memory>.
8. Seibert, P. Short-term memory | EBSCO. EBSCO Information Services, Inc. | www.ebsco.com. <https://www.ebsco.com/research-starters/health-and-medicine/short-term-memory>.
9. Richardson, J. T. E. Measures of Short-Term Memory: A Historical Review. *Cortex* 2007, 43 (5), 635–650. [https://doi.org/10.1016/s0010-9452\(08\)70493-3](https://doi.org/10.1016/s0010-9452(08)70493-3).
10. Richardson, J. T. E. Knox's Cube Imitation Test: A Historical Review and an Experimental Analysis. *Brain and Cognition* 2005, 59 (2), 183–213. <https://doi.org/10.1016/j.bandc.2005.06.001>.
11. Oxford Reference. Corsi Blocks. Oxford Reference. <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095640799>.
12. APA Dictionary of Psychology. Digit Span. [dictionary.apa.org](https://dictionary.apa.org/digit-span). <https://dictionary.apa.org/digit-span>.
13. Cleary, A. M.; Langley, M. M.; Seiler, K. R. Recognition without Picture Identification: Geons as Components of the Pictorial Memory Trace. *Psychonomic Bulletin & Review* 2004, 11 (5), 903–908. <https://doi.org/10.3758/bf03196719>.
14. APA Dictionary of Psychology. *Visual Memory*. APA Dictionary of Psychology. <https://dictionary.apa.org/visual-memory>.
15. Ebaid, D.; Crewther, S. G. Visual Information Processing in Young and Older Adults. *Frontiers in Aging Neuroscience* 2019, 11. <https://doi.org/10.3389/fnagi.2019.00116>.

16. Freitas, M. I. d'Ávila; Ribeiro, A. F.; Radanovic, M.; Mansur, L. L. Working Memory: Differences between Young Adults and the Aged in Listening Tasks *. *Dementia & Neuropsychologia* 2007, 1 (2), 147–153. <https://doi.org/10.1590/s1980-57642008dn10200006>.

■ Author

Prathiksha Raguram is a high school senior at J. L. Mann High School, Greenville, South Carolina. She hopes to apply her computer engineering degree to the fields of neuroscience and psychology.

Microplastics and Female Reproduction: Comprehensive Review of Hormonal and Fertility Alterations

Precious Ojo

Indigo Research; preciousdara.ojo@gmail.com

Mentor: Dr. Barbara Biney

ABSTRACT: Microplastics (MPs) and nanoplastics (NPs) are ubiquitous pollutants derived from plastic degradation. These plastics emerged as silent disruptors of human health. While early studies centered on environmental contamination, recent research reveals alarming links between MP/NP exposure and female reproductive health. This paper synthesizes evidence from studies published between 2019 and 2025, sourced from PubMed and Google Scholar, to evaluate how MPs and NPs influence female fertility, hormonal regulation, and reproductive diseases. Findings indicate that MPs are now being detected in human placental tissue and ovarian follicles, follicular fluid, confirming the ability to cross biological barriers. Exposure has been associated with hormonal dysregulation (altered FSH, LH, and estradiol levels), ovarian fibrosis, reduced oocyte quality, precocious puberty, and uterine fibroids. Moreover, co-exposure to MPs and plastic additives such as DEHP exacerbates endocrine disruption via the TGF- β 1/Smad3 pathway, mirroring PCOS-like changes. These outcomes suggest that MPs are not passive pollutants but active agents of reproductive dysfunction. The study underscores an urgent need for more toxicological research, detection, and monitoring of MPs and NPs.

KEYWORDS: Biomedical and Health Sciences, Cell, Organ and Systems Physiology, Earth and Environmental Science, Microplastics, Reproductive Health, Female Fertility.

■ Introduction

Global plastic production reached 4.9 billion tons, which entered the environment in 2015. By 2025, this amount will have doubled, leading to 11 billion tons of plastic polluting the earth.¹ The increasing presence of plastic and the degradation of its products in the environment has become an important global emergency. Although the use of plastics has greatly increased convenience, it has also led to the environmental release of a significant amount of plastic waste, which has accumulated in ecosystems. Right now, only 9% of the 460 million tons of plastic items produced in 2019 were recycled, and by 2060, this number is predicted to rise to 1.2 billion tons.²

From the process of plastic degradation, particles of plastic on a micrometric scale are created. This is commonly known as microplastics. Microplastics (MPs) are typically defined as synthetic plastic particles ranging from 5 mm to 1 μ m in size, and nanoplastics with a diameter smaller than 1 μ m.³ “They are hydrophobic, synthetic, solid, or polymer matrices with a regular or irregular shape⁴ (fibers, fragments, spheres, beads, films, scales, pellets, and foam) and colors depending on the original form of the large plastics from which they derive.”^{3,4} Their compositions mainly include polyethylene (PE), polyethylene terephthalate (PET), polypropylene (PP), polystyrene (PS), and polyvinyl chloride (PVC).⁵ These MPs are classified into two major types; the first is the primary, which arose from intentional products, smaller in size and mostly rounded in shape. Primary MPs are specifically produced by humans both for their abrasive properties and to improve the stability of certain products; consequently, MPs are intentionally added to cosmetics, toothpaste, household products, detergents, and

paint formulas.³ The other type is the secondary, which originates from the fragmentation of larger particles and is usually irregular in shape. Secondary MPs are also derived from the fragmentation of plastic wastes created from thermal destruction, UV radiation, photodegradation, and biodegradation.³ These MP particles are found everywhere in all natural environments, including water, sewage, air, soil, and the human food chain.¹

Data have shown that humans can ingest up to 5g of MPs per week via multiple routes of exposure.⁶ MPs can enter the human body via three primary routes: oral (consumption), inhalation (breathing), and dermal exposure (skin contact). Of these, oral ingestion is the main route of human exposure to MPs.⁵ Given the ubiquity of MPs in the environment, people inevitably inhale or ingest them daily. MPs have been found in human feces, which is strong proof that they were consumed through the food chain and exposed to the stomach.⁷ Therefore, evaluating the risks of exposure to MPs is an urgent public health priority.⁸ Recent studies have detected microplastics in various human organs and tissues, including the placenta of pregnant women, liver, lung tissues, penis, blood, urine, amniotic fluid, and the human brain.¹ Several studies show that MPs smaller than 10 μ m can enter cells and cause serious health problems, including oxidative stress and damage to cells, inflammation, immune issues, neurotoxicity, reprotoxicity, and carcinogenicity.³

Furthermore, MPs, due to their hydrophobic surface, act as a Trojan horse for other types of notoriously toxic environmental contaminants (such as bisphenols, phthalates, and heavy metals) and can bind to them through processes of bioaccu-

mulation and biomagnification, causing additional harm to living organisms.³ These substances have been associated with various health issues in humans, such as epigenetic changes, reproductive toxicity, and cancer.⁷

These developments raised urgent questions about the long-term impacts of MPs and NPs on female reproductive health, especially in areas such as hormonal regulation, fertility, and transgenerational health. It is within this critical context that this paper synthesizes current research on the impact of microplastics (MPs) and nanoplastics (NPs) on female reproductive health, examining their potential roles in hormonal disruption, abnormal puberty onset, infertility, pregnancy complications, uterine fibroids, and reproductive cancers. This paper provides a comprehensive overview of how MPs and NPs may influence female reproductive biology at multiple stages by merging findings across molecular and hormonal pathways. Moreover, despite this growing body of evidence, the unique risks MPs pose to female reproductive health remain under-researched and overlooked in global health discourse. Addressing this gap requires urgent investment in gender-specific toxicology research and proactive public health strategies.

■ Methods

An electronic literature search was performed across Google Scholar and PubMed for material published between 2019 and 2025. The search terms used were: microplastics, nanoplastics, microplastics and reproductive health, effects of microplastics on reproductive health, microplastics and the ovaries, microplastics in human health, microplastics and female reproductive health, microplastics and female reproductive hormones, plastic pollution and human health, microplastics estradiol, microplastics estriol, microplastics and puberty, microplastics fibroids, microplastics and PCOS, microplastics female fertility, microplastics estriol, microplastics and FSH, LH, polystyrene and female reproductive health.

Each publication was reviewed individually, focusing on the aim of the study and specifically on microplastics and female reproductive health. There were no language restrictions.

■ Results

The Global Decline of Fertility amid Plastic Pollution:

Over the past decades, global fertility rates have declined sharply. Since the 1960s, global total fertility rates (TFR) have begun to fall, and according to the United Nations, two-thirds of the global population currently live in countries with fertility rates below replacement levels. Increasing attention is being drawn to the role of environmental toxins, especially microplastics (MPs), in this global trend, even though social, economic, and medical factors are equally important contributors.⁹ As plastic production continues to rise exponentially, human exposure to MPs and nanoplastics (NPs) has become unavoidable due to them infiltrating air, water, food, and ultimately the human body. Microplastics and nanoplastics have become a great concern for public health experts due to their evident prominence in impacting the reproductive system. Predictably, both male and female reproductive systems

are susceptible to environmental and lifestyle-related damage. However, research has historically focused more on male fertility, largely due to the easier accessibility of sperm for analysis compared to oocytes.⁹

Research on reduced fertilization and failure of offspring to thrive due to micro and nano-plastics was first done primarily on different marine animal models like *Tigriopus japonicus* and *Daphnia magna*. This showed a visible decrease in fecundity, survival, growth, development, feeding capability, and reproduction capacity of offspring. Testing also revealed that the toxicity of MPs-NPs in egg follicles made them remain unfertilized and fail to develop into new offspring. Moreover, it was observed that the MPs induced trans-generational effects on offspring, affecting their phenotypes.⁴

Controlled laboratory studies on mice further demonstrate the reproductive toxicity of microplastics at the mammalian level. In one study, male and female mice were exposed to polystyrene microplastics (PS-MPs) via oral gavage over 44 days. Post-mating analysis revealed that PS-MPs accumulated significantly in both the testes and ovaries, with a higher degree in ovarian tissue ($p < 0.05$), confirming their ability to cross biological barriers. The fertility outcomes in this study showed that in females, this accumulation was associated with reduced ovary-to-body weight ratios and a lower number of embryos.⁵ Both sexes exhibited reduced pregnancy rates and fewer embryos post-mating, indicating compromised fertility. When neither was exposed to MP, the pregnancy rate was 89%; when only the female was exposed, the pregnancy rate was 38%; when only the male was exposed, the rate was 78%; and when both were exposed, the pregnancy rate was 75%. Meaning that exposure to PS-MPs reduced the pregnancy rate by up to ~57% when only the female was exposed. This was the most significant reduction across all groups; a summary of these findings is provided in Appendix 1, Table 1.⁵

Several studies focus specifically on correlations between MP exposure and female fertility disruption. One systematic review (Sharma *et al.*) analyzed 12 studies (totaling 234 subjects) and confirmed the presence of MPs ranging from 2.1 to 100 μm in human placental and fetal tissues. The findings showed correlations of MP levels with indicators of reduced fetal growth, including low birth weight and altered gestational age. This suggested broader reproductive implications, though direct fertility data remain limited.² Another breakthrough study showed that co-exposure to MPs and DEHP triggers PCOS-like ovarian damage via activating the TGF- β 1/Smad3 pathway. This pathway is normally inhibited, preventing fibrosis, oxidative stress, and hormone imbalance. The study also found serious molecular damage via proteomics, including barrier dysfunction and metabolic disruption. This makes it a key contribution to understanding how plastic pollution may harm female fertility.⁸

These are clear indicators of reduced fertility. These findings suggest that MPs may affect ovarian function and have a negative impact on ovarian capacity, oocyte maturation, and oocyte quality through endocrine disruption, oxidative stress, inflammation, and ultimately affect reproductive success.⁴

The Hormonal Cycle and Microplastics Disruption:

Figure 1 illustrates the normal female reproductive cycle regulated by the hypothalamic–pituitary–ovarian (HPO) axis, highlighting the coordinated roles of follicle-stimulating hormone (FSH), luteinizing hormone (LH), estrogen, and progesterone across the follicular, ovulatory, and luteal phases. Within this axis, estradiol (E2) drives follicular growth and ovulation, estrone (E1) contributes to baseline estrogen balance, estriol (E3) supports pregnancy and placental function, while progesterone stabilizes the luteal phase and prepares the endometrium for implantation. Disruption of estradiol, estrinol, or progesterone synthesis impairs gonadotropin signaling, folliculogenesis, luteal function, endometrial receptivity, and pregnancy maintenance, linking hormonal imbalance to infertility, PCOS, and adverse reproductive outcomes.

One of the first ways MP and NP impact female reproductive health is through the disruption of these major hormones: FSH, LH, estrogen, and progesterone. Each of these has a specific role in regulating the menstrual cycle and reproductive success in females. Research has shown that MP exposure led to increased FSH and LH levels in mouse models when exposed to polystyrene microplastics.⁵ This suggests MPs can cause reduced ovarian reserve, decreased fertility, and a lower chance of a successful pregnancy, potentially leading to conditions like Polycystic Ovary Syndrome (PCOS) and ovarian failure.

Furthermore, a study, detailed in Appendix 1, Table 1, on human adrenal carcinoma cells (H295R) found that exposure to 50 nM of PS-MNPs at 100 µg/mL for 48 hours reduced estriol (E3) levels by ~37.5%, with no significant changes to estrogen and androgen receptor activity. Since adequate estriol is necessary for placental and uterine function, this drop suggests that PS-MNPs may disrupt steroidogenesis and trigger oxidative stress, with potential consequences for reproductive and fetal health during pregnancy, particularly when compared to normal physiological estrogen dynamics illustrated in Figure 1.¹⁰ Another study on female Sprague-Dawley rats utilized doses of 0.5 mg/day of PS-MPs, found that co-exposure to polystyrene microplastics (MPs) and DEHP (a plastic additive) lowered estradiol (E2) while raising testosterone levels, a hormonal profile consistent with PCOS-like ovarian damage via the TGF-β1/Smad3 pathway.⁸ Relative to the normal cycle shown in Figure 1, where rising estradiol drives follicular maturation and ovulation, those reduced levels are insufficient to support ovulation, thereby promoting follicular atresia and infertility.

In addition, there is also growing evidence that microplastics may accelerate pubertal onset in females.¹¹ Precocious puberty (PP), the unusually early onset of sexual development, is a prevalent pediatric endocrine disorder affecting children's growth and sexual development with documented links to environmental pollutants, including microplastics. Observational research found that early MP exposure activated pubertal signaling in the brain: microglial and astrocytic activation increased GnRH expression in the hypothalamus, pushing puberty forward.¹² As illustrated in Figure 2, GnRH is the primary neuroendocrine trigger for HPA axis activation, driv-

ing downstream estrogen and progesterone secretion, which connects MP-induced neural changes directly to accelerated development.

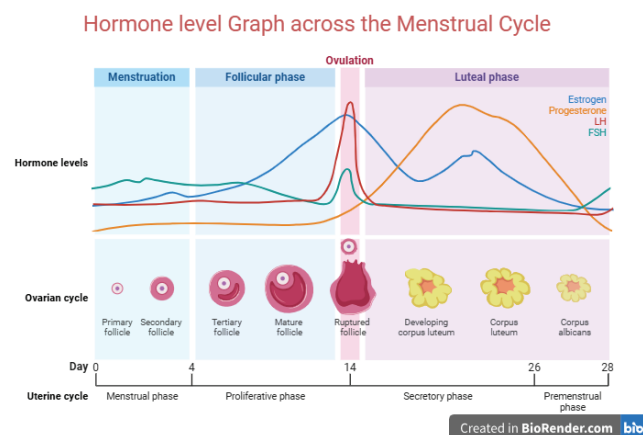


Figure 1: Hormone Level Graph Across the Menstrual Cycle. Created in BioRender.com. The LH surge at day 14 triggers ovulation, marking the shift from the follicular to the luteal phase. Rising progesterone after ovulation reflects corpus luteum formation and prepares the uterine lining for implantation.

In a human case–control study of 14 girls, those with central precocious puberty (CPP) had significantly higher MP concentrations in their blood serum compared to healthy controls (68.19 ± 26.49 mg/kg vs. 36.94 ± 13.22 mg/kg). The dominant polymers were polyethylene (PE), CPP: 28.22%; Control: 31.62%, and polyvinyl chloride (PVC), CPP: 62.67%; Control: 58.14%. Hormonal analysis revealed a negative correlation between MPs and progesterone levels, alongside a positive correlation with dehydroepiandrosterone sulfate (DHEAS), indicating that MPs may disrupt steroid balance during critical developmental windows.¹² Under normal physiology, progesterone rises in the luteal phase to prepare the endometrium for implantation (Figure 1); the observed reduction with MP exposure suggests impaired reproductive function and increased infertility risk. Complementary animal experiments confirmed these findings. Young female mice exposed to PE (3 mg/day) and PVC (6 mg/day) showed advanced vaginal opening, a biological marker of early puberty. Brain tissue analysis confirmed elevated hypothalamic GnRH expression and glial cell activation, as seen in Figure 2. These findings provide some of the first experimental evidence that PE and PVC specifically can drive premature puberty by disrupting ovarian hormones and overstimulating the HPG axis. Early puberty carries long-term risks for pediatric endocrine health, including infertility, metabolic disease, and reproductive cancers.

As women move into adulthood, microplastics can further impact the reproductive system by inducing and intensifying pathological dysregulated hormonal states like Polycystic Ovarian Syndrome (PCOS), one of the most common endocrine disorders in reproductive-age women, defined by cystic follicles, hyperandrogenism, and hormonal dysregulation. One study examined the effects of co-exposure to polystyrene microplastics (MPs) and di-2-ethylhexyl phthalate (DEHP), a widely used plastic additive, finding PCOS-like phenotypes

in animal models. Key findings include the visible disruption of the ovarian morphology, particularly the formation of cystic follicles and elevated follicular atresia. Exposed female Sprague-Dawley rats in the MP + low-DEHP and MP + high-DEHP groups developed approximately 64 and 58 atretic follicles, respectively, compared to only ~25–30 in saline and MP-only controls. Ovarian fibrosis was also observed, with excess connective tissue damaging normal ovarian structure. Hormonally, the exposed groups showed roughly a 20% drop in serum estradiol and a concurrent 21–29% rise in testosterone, producing a significant downregulation of the E2/T ratio by up to 29%. This hormonal profile aligns with hyperandrogenism, a defining trait of PCOS.⁸

Furthermore, the study found that microplastics and DEHP together activated the TGF- β 1/Smad3 pathway. This pathway is an intracellular serine/threonine kinase signaling cascade that drives fibrosis and inflammation. It is activated by the binding of TGF- β 1 to cell surface receptors and is downregulated by inhibitory proteins such as Smad7 or targeted chemical compounds. When dysregulated, this pathway promotes excessive extracellular matrix deposition and ovarian tissue remodeling, contributing to follicular loss and impaired ovarian function.¹³ This signaling pathway increased fibrosis (scarring), quantified by an elevated collagen volume fraction of over 40% in the MP + H-DEHP group, and led to higher oxidative stress in the ovary. When researchers blocked Smad3 with the compound called SIS3, the detrimental effects were significantly reversed. Hormone levels improved, the E2/T ratio by ~42% (from ~5.0 to ~7.1) following SIS3 treatment, returning it to the Saline control level of 7.0, and tissue damage was reduced, the collagen volume fraction by ~25% (from 40% to ~30%).⁸ This shows that TGF- β 1/Smad3 plays a central role in how pollutants harm reproductive health.

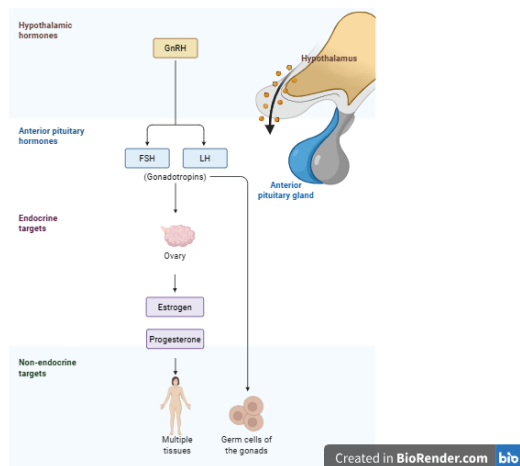


Figure 2: The Hypothalamic-Pituitary-Gonadal (HPG) Axis and its role in hormone release. Created in BioRender.com. GnRH from the hypothalamus stimulates the pituitary to release FSH and LH, which in turn drive ovarian production of estrogen and progesterone. These hormones act on multiple tissues, coordinating the body's reproductive response.

Pathological and Developmental Risks of Microplastic Exposure: Uterine Fibroids and Cervical Cancer:

Emerging evidence suggests that microplastics have a dual pathological effect on female reproductive tissues. They can

stimulate abnormal tissue proliferation (hyperplasia, tumor formation) while also damaging healthy reproductive structures through inflammation, oxidative stress, and endocrine disruption. A 2025 case-control study of 40 fibroid patients and 40 healthy controls provided direct evidence of MPs in uterine fibroid tissue.¹⁴ Uterine fibroids (UFs) are benign smooth muscle tumors of the uterus that affect an estimated 70–80% of women of reproductive age globally, often leading to infertility, heavy bleeding, and chronic pain.¹⁵

The results of the study showed that MP concentrations were highest in fibroid tissue, followed by patient myometrium, and lowest in healthy controls. Polyethylene (PE) dominated, and its exposure was linked to a threefold increased fibroid risk (OR = 3.327, $p = 0.012$).¹⁴ This suggests not all MP types carry equal risk, with PE posing a particular concern. While genetics and hormones are established contributors to fibroid development, these findings add environmental pollutants to the list of plausible risk factors.

Regarding cervical cancer, a preliminary cross-sectional study found that MPs may promote pathological growth, inflammation, and potentially carcinogenic processes in reproductive tissues. A study done in China analyzed blood, cancerous cervical tissue, and adjacent paracancerous tissue from 26 cervical cancer patients using Raman spectroscopy. Thirteen microplastic (MP) types were identified, with polyethylene (PE), polypropylene (PP), and PE-co-PP as the dominant polymers. MP abundance followed the pattern: blood > cancerous tissue > paracancerous tissue, with significantly higher diversity in tumors than in adjacent tissue ($p < 0.05$). Although these findings are significant, further research comparing cancerous vs non-cancerous populations is needed to further investigate the association between microplastics and cancerous tissue.

The pathogenesis of cervical cancer and fibroids from microplastics may be due to the upregulation of amino sugar and nucleotide sugar metabolism pathways. These pathways may support abnormal cell and fibroid growth. In addition, microplastics may cause a downregulation of cofactor biosynthesis and platelet activation pathways, leading to decreased cell repair or detox, allowing fibroids to thrive and poor healing or inflammation control, worsening fibroid symptoms.¹⁴

Extracorporeal Therapeutic Apheresis in Microplastic Removal:

Until very recently, there were no established medical interventions capable of actively removing microplastics (MPs) or nanoplastics (NPs) from the human body. Research on the biological impact of these particles has grown rapidly, but therapeutic strategies have lagged, largely because of the microscopic size of MPs/NPs and their ability to accumulate in deep tissues. Other than detection, the extraction has been a major impediment.

In 2025, Bornstein and colleagues conducted the first human study exploring the possibility of using extracorporeal therapeutic apheresis (TA) as a method for eliminating microplastic- and nanoplastic-like particles circulating in the blood plasma. As shown in Figure 3, this process involves blood extraction via venous access. The plasma is separated from blood

cells using a porous membrane, then filtered through adsorption columns or membranes designed to capture pathogenic or toxic components before the blood components are reinfused into the patient. The study recruited 21 patients diagnosed with post-infectious myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), all of whom underwent at least two sessions of double-filtration apheresis (INUSpheres).¹⁶

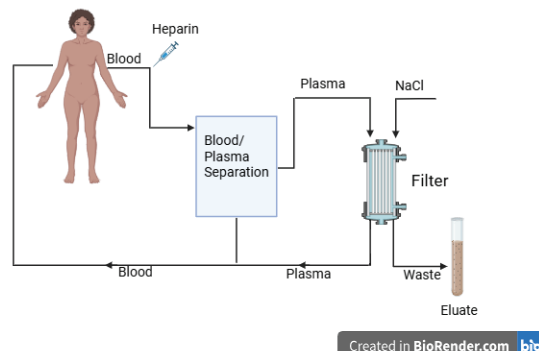


Figure 3: Schematic representation of the extracorporeal therapeutic apheresis process used for microplastic and nanoplastic removal. Blood is withdrawn from a large vein and circulated through an apheresis system, where a semipermeable membrane separates plasma from cellular components. The isolated plasma is then directed through specialized filtration columns that selectively remove targeted molecules like microplastics. Once filtration is complete, the remaining plasma and blood cells are returned to the patient via a second venous line. The filtered waste, or eluate, may contain retained microplastic and nanoplastic particles. Created in BioRender.com.

Analysis of the concentrated eluate, which is the waste fraction collected after filtration, revealed that 14 polymer-like substances, including polyamide 6 and polyurethane, were identified using attenuated total reflectance Fourier-transform infrared (ATR-FT-IR) spectroscopy. Figure 3 contextualizes this workflow, showing how the eluate fraction is isolated and analyzed for polymer content. The polymer signatures matched up to 67.5% for polyamide 6 and 35.3% for polyurethane. This suggests that apheresis may be capable of physically trapping and removing certain microplastic residues from human circulation. However, the study did not quantify total polymer concentrations removed, nor did it measure blood levels before and after treatment.

The study authors cautioned that these findings are preliminary and novel. Larger patient cohorts and more extensive clinical trials are still needed. Future work should use quantitative analytical tools like pyrolysis–gas chromatography–mass spectrometry (Py-GC/MS) to measure microplastic levels in blood plasma both before and after therapy, and within the eluate fractions collected during apheresis. Tracking these levels across multiple treatment cycles would help determine how consistently and effectively apheresis removes MNPs from circulation.

Despite these limitations, the findings are considered groundbreaking because no prior human intervention had ever demonstrated active removal of microplastic-like materials from the bloodstream. The absence of such approaches historically stems from some major challenges, which include the invisibility and chemical resilience of microplastics in biological systems, the lack of analytical tools sensitive enough to

detect microplastics at nanoscales, and uncertainty over their clinical relevance and effect until recent evidence confirmed microplastic presence in blood and reproductive tissues.

Polyamide and polyurethane, which are the same polymer types detected in the apheresis eluates, have also been identified in female reproductive tissues, including the placenta and the uterus.¹⁷ Although the 2025 apheresis study did not quantify how much of these polymers were removed from circulation, the overlap in polymer classes is scientifically meaningful. It suggests that apheresis is capable of capturing some of the microplastics accumulating in reproductive organs. This raises the possibility that, with further refinement, extracorporeal therapeutic apheresis could one day be adapted to target the specific polymer types most frequently detected in female reproductive systems.

However, achieving this goal requires future studies that measure blood microplastic levels before and after apheresis, track polymer concentrations across multiple treatment cycles, answer the question of whether removal improves with repeated sessions, and directly compare the polymer profiles found in circulation with those accumulating in reproductive tissues. These future steps are essential because they determine whether removing specific polymer types from the bloodstream can reduce the very same microplastics that accumulate in female reproductive tissues, thereby directly addressing the sex-specific risks highlighted in this study.

■ Discussion

The collective findings reviewed in this paper demonstrate that microplastics and nanoplastics are not inert environmental contaminants but biologically active agents capable of disrupting female reproductive health through endocrine imbalance, oxidative stress, inflammatory signaling, and scarring of reproductive tissues. Evidence from animal models, *in vitro* systems, and emerging human studies consistently links microplastic exposure to altered hormonal regulation. In addition, exposure to microplastics is correlated with impaired ovarian function, abnormal pubertal timing, and increased risk of uterine fibroids and cervical cancer. Despite growing evidence of these effects, significant challenges remain in accurately characterizing human exposure. Currently, there are no standardized clinical blood tests for microplastics; however, the techniques mentioned earlier, such as Raman spectroscopy and Fourier-transform infrared (FTIR) analysis, demonstrate the initial methodology of detection.⁷ Simultaneously, the recent discovery and use of therapeutic apheresis represents a promising strategy for addressing internal microplastic accumulation.

These methodologies are helpful steps in the right direction in addressing this problem. The female reproductive system is highly sensitive to microplastic exposure, and even minor fluctuations in FSH, LH, estradiol, or progesterone can cascade into significant physiological disruption. This sensitivity, along with increasing plastic contamination in the environment, makes for a compounding effect and a public health concern. Ultimately, prevention remains the most practical and sustainable approach to this problem. On an individual level, using

high-quality water filters, avoiding single-use plastics, and choosing fresh foods over heavily packaged alternatives are a few of the many solutions to the problem. Additional individual practices to reduce microplastics exposure include limiting shellfish consumption due to high levels of bioaccumulation in the shellfish.¹⁸

Community-level interventions are also critical, and implementing recycling programs in communities that do not currently have them is an initial step. In areas where recycling systems are already in place, more advanced recycling methods such as chemical pyrolysis can also be utilized. Chemical pyrolysis uses heat in the absence of oxygen to break plastics down into their basic molecular components.¹⁹ This approach is often more effective than traditional mechanical recycling, which typically involves shredding and melting plastic into pellets and can reduce material quality over time. Lastly, using policy to decrease the manufacturing of plastics is another solution. For example, the EU Single-Use Plastics Directive in 2021 banned the manufacture and sale of single-use plastic items across all 27 EU member states and thus directly eliminated the entire category of plastics from sale.²⁰ Ultimately, combating the pollutants of microplastics necessitates a shift from personal lifestyle adjustments to strict source-control legislation. Future global health frameworks must prioritize toxicological standards to protect the unique sensitivities of the female reproductive system and ensure that the documented decline in global fertility is not exacerbated by unregulated plastic pollution.

The collective findings reviewed in this paper demonstrate that microplastics and nanoplastics are not inert environmental contaminants but biologically active agents capable of disrupting female reproductive health through endocrine imbalance, oxidative stress, inflammatory signaling, and scarring of reproductive tissues. Evidence from animal models, *in vitro* systems, and emerging human studies consistently links microplastic exposure to altered hormonal regulation, impaired ovarian function, abnormal pubertal timing, and increased risk of gynecological pathologies, including uterine fibroids and cervical cancer.

Limitations & Future Work:

The strength of current conclusions is constrained by a number of limitations, despite the mounting evidence that microplastics are linked to the reproductive health of females. First, there is still a lack of human data, as most research uses invasive, non-scalable sampling or small sample sizes. It is not possible to accurately measure microplastics or differentiate particle toxicity from chemical additives using detection techniques like Raman and FTIR spectroscopy. Standardized worldwide measurement protocols and longitudinal cohorts will be necessary to close these gaps. Furthermore, comprehending the practical effects of long-term, low-level environmental exposure over a person's lifetime requires going beyond the high-dose laboratory models used to determine biological thresholds.

Table 1: Summary of Experimental Studies on Microplastic Exposure and Female Reproductive Outcomes.

Study Type	Polymer Type	Experimental Dose	Target Tissue	Key Reproductive Outcome	Proposed Mechanism
Animal (Mouse) ⁵	Polystyrene (PS)	Oral gavage (44 days)	Ovary	Reduced the ovary-to-body weight ratio; decreased pregnancy rate by up to 57%	Oxidative stress and endocrine disruption
Human (Placental) ²	Mixed Microplastics (2.1–100 µm)	Environmental exposure	Placenta and fetal tissue	Altered gestational age and reduced fetal growth	Barrier penetration and inflammation
Animal (Mice) ⁵	Polystyrene (PS)	Oral gavage (44 days)	ovary	Decreased Pregnancy rate (up to 57%)	Oxidative stress; Endocrine disruption
<i>In vitro</i> (H295R cells) ¹⁰	Polystyrene (50 nm)	100 µg/mL (48 hours)	Adrenal steroidogenic cells	Reduced estril levels by approximately 37.5%	Disrupted steroidogenesis and oxidative stress
Animal (female Sprague-Dawley rats) ⁸	Polystyrene + DEHP	Low and high DEHP co-exposure	Ovary	PCOS-like phenotype; reduced estradiol; increased testosterone	TGF-β1/Smad3 activation and fibrosis
Human Uterus: (40 fibroid patients; 40 healthy controls) ¹⁴	Polyethylene (PE)	Environmental exposure	Uterine fibroid tissue	3x higher risk of fibroids	Metabolic dysregulation ²¹ nd tissue prolif ² at ²³
Human (Cervical) ²¹	PE, PP, and PE-co-PP	Environmental exposure	Cervical tissue	Higher microplastic diversity in tumor tissue	Inflammation and carcinogenic signaling
Human (Puberty Study) ¹¹	PE and PVC	Environmental exposure	Blood serum	Reduced progesterone; increased DHEAS	HPG axis disruption
Animal (Juvenile Mouse) ¹¹	PE and PVC	3 to 6 mg per day	Hypothalamus and ovary	Accelerated pubertal onset	GnRH overstimulation

AI Acknowledgment

During the preparation of this work, the author used Google Gemini in order to brainstorm research topics and key terms, etc. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgments

I extend my deepest gratitude to my research mentor, Dr. Barbara Biney, for her exceptional mentorship throughout this year-long project. Her rigorous instruction in research methodology, consistent feedback through regular meetings, and thoughtful review of my work were invaluable to the development and execution of this study. I also thank the Indigo Research Program for awarding me the Next Nobel Research Scholarship and providing all the financial support, mentorship structure, and academic resources that made this independent research possible.

References

1. Jaafarzadeh Haghghi Fard, N.; Mohammadi, M. J.; Jahedi, F. Effects of Nano and Microplastics on the Reproduction System: *In*

- Vitro and in Vivo Studies Review. Food Chem. Toxicol.* **2023**, 178, 113938. <https://doi.org/10.1016/j.fct.2023.113938>.
2. Sharma, R. K.; Kumari, U.; Kumar, S. Impact of Microplastics on Pregnancy and Fetal Development: A Systematic Review. *Cureus* **2024**. <https://doi.org/10.7759/cureus.60712>.
 3. Montano, L.; Raimondo, S.; Piscopo, M.; Ricciardi, M.; Guglielmino, A.; Chamayou, S.; Gentile, R.; Gentile, M.; Rapisarda, P.; Oliveri Conti, G.; Ferrante, M.; Motta, O. First Evidence of Microplastics in Human Ovarian Follicular Fluid: An Emerging Threat to Female Fertility. *Ecotoxicol. Environ. Saf.* **2025**, 291, 117868. <https://doi.org/10.1016/j.ecoenv.2025.117868>.
 4. Afreen, V.; Hashmi, K.; Nasir, R.; Saleem, A.; Khan, M. I.; Akhtar, M. F. Adverse Health Effects and Mechanisms of Microplastics on Female Reproductive System: A Descriptive Review. *Environ. Sci. Pollut. Res.* **2023**, 30 (31), 76283–76296. <https://doi.org/10.1007/s11356-023-27930-1>.
 5. Wei, Z.; Wang, Y.; Wang, S.; Xie, J.; Han, Q.; Chen, M. Comparing the Effects of Polystyrene Microplastics Exposure on Reproduction and Fertility in Male and Female Mice. *Toxicology* **2022**, 465, 153059. <https://doi.org/10.1016/j.tox.2021.153059>.
 6. Estimation of the Mass of Microplastics Ingested – A Pivotal First Step towards Human Health Risk Assessment. *J. Hazard. Mater.* **2021**, 404, 124004. <https://doi.org/10.1016/j.jhazmat.2020.124004>.
 7. Damaj, S.; Trad, F.; Goevert, D.; Wilkesmann, J. Bridging the Gaps between Microplastics and Human Health. *Microplastics* **2024**, 3 (1), 46–66. <https://doi.org/10.3390/microplastics3010004>.
 8. Xu, K.; Wang, Y.; Gao, X.; Wei, Z.; Han, Q.; Wang, S.; Du, W.; Wan, J.; Wan, C.; Chen, M. Polystyrene Microplastics and Di-2-Ethylhexyl Phthalate Co-Exposure: Implications for Female Reproductive Health. *Environ. Sci. Ecotechnol.* **2024**, 22, 100471. <https://doi.org/10.1016/j.ese.2024.100471>.
 9. Aitken, R. J. What Is Driving the Global Decline of Human Fertility? Need for a Multidisciplinary Approach to the Underlying Mechanisms. *Front. Reprod. Health* **2024**, 6. <https://doi.org/10.3389/frph.2024.1364352>.
 10. van Boxel, J.; Khargi, R. R. J.; Nijmeijer, S. M.; Heinzlmann, M. T.; Pereira, D. D. C.; Lamoree, M. H.; van Duursen, M. B. M. Effects of Polystyrene Micro- and Nanoplastics on Androgen- and Estrogen Receptor Activity and Steroidogenesis *in Vitro*. *Toxicol. Vitro Int. J. Publ. Assoc. BIBRA* **2024**, 101, 105938. <https://doi.org/10.1016/j.tiv.2024.105938>.
 11. Zang, S.; Wang, F.; Ouyang, Y.; Cheng, W.; Li, P.; Yin, X. Exposure to Polyethylene and Polyvinylchloride Microplastics Caused Advanced Puberty Onset in Females through Promoting Hypothalamic GnRH Expression. *Ecotoxicol. Environ. Saf.* **2025**, 291, 117906. <https://doi.org/10.1016/j.ecoenv.2025.117906>.
 12. Zang, S.; Wang, F.; Ouyang, Y.; Cheng, W.; Li, P.; Yin, X. Exposure to Polyethylene and Polyvinylchloride Microplastics Caused Advanced Puberty Onset in Females through Promoting Hypothalamic GnRH Expression. *Ecotoxicol. Environ. Saf.* **2025**, 291, 117906. <https://doi.org/10.1016/j.ecoenv.2025.117906>.
 13. Shi, L. B.; Zhou, F.; Zhu, H. Y.; Huang, D.; Jin, X. Y.; Li, C.; Dai, Y.; Pan, Y. B.; Zhang, S. Y. Transforming Growth Factor Beta1 from Endometriomas Promotes Fibrosis in Surrounding Ovarian Tissues via Smad2/3 Signaling. *Biol. Reprod.* **2017**, 97 (6), 873–882. <https://doi.org/10.1093/biolre/iox140>.
 14. Xu, H.; Dong, C.; Zhang, B.; Jin, Y.; Cheng, J.; Yu, Z.; Ozaki, Y.; Yu, J.; Ma, D.; Xie, Y. The Role of Microplastics in Elevated Risk and Tissue Metabolic Alterations in Uterine Fibroid of Female Patients. *Environ. Pollut. Barking, Essex* **2025**, 374, 126251. <https://doi.org/10.1016/j.envpol.2025.126251>.
 15. Katz, T. A.; Yang, Q.; Treviño, L. S.; Walker, C. L.; Al-Hendy, A. Endocrine Disrupting Chemicals and Uterine Fibroids. *Fertil. Steril.* **2016**, 106 (4), 967–977. <https://doi.org/10.1016/j.fertnstert.2016.08.023>.
 16. Bornstein, S. R.; Gruber, T.; Katsere, D.; El Attaoui, A.; Wohlsperger, L.; Yaman, M.; Kanczkowski, W.; Piwernetz, G.; Straube, R.; Voit-Bak, K.; Licinio, J.; Steenblock, C. Therapeutic Apheresis: A Promising Method to Remove Microplastics? *Brain Med. Neurons Behav. Better Health* **2025**, 1 (3), 52–53. <https://doi.org/10.61373/bm0251.0056>.
 17. Ali-Hassanzadeh, M.; Arefinia, N.; Ghoreschi, Z.-A.-S.; Askarpour, H.; Mashayekhi-Sardoo, H. The Effects of Exposure to Microplastics on Female Reproductive Health and Pregnancy Outcomes: A Systematic Review and Meta-Analysis. *Reprod. Toxicol.* **2025**, 135, 108932. <https://doi.org/10.1016/j.reprotox.2025.108932>.
 18. Cherian, A. G.; Liu, Z.; McKie, M. J.; Almuhtaram, H.; Andrews, R. C. Microplastic Removal from Drinking Water Using Point-of-Use Devices. *Polymers* **2023**, 15 (6), 1331. <https://doi.org/10.3390/polym15061331>.
 19. Hasan, M. M.; Haque, R.; Jahirul, M. I.; Rasul, M. G. Pyrolysis of Plastic Waste for Sustainable Energy Recovery: Technological Advancements and Environmental Impacts. *Energy Convers. Manag.* **2025**, 326, 119511. <https://doi.org/10.1016/j.enconman.2025.119511>.
 20. Kasznik, D.; Łapniewska, Z. The End of Plastic? The EU's Directive on Single-Use Plastics and Its Implementation in Poland. *Environ. Sci. Policy* **2023**, 145, 151–163. <https://doi.org/10.1016/j.envsci.2023.04.005>.
 21. Xu, H.; Dong, C.; Yu, Z.; Ozaki, Y.; Hu, Z.; Zhang, B.; Yao, W.; Yu, J.; Xie, Y. Detection and Analysis of Microplastics in Tissues and Blood of Human Cervical Cancer Patients. *Environ. Res.* **2024**, 259, 119498. <https://doi.org/10.1016/j.envres.2024.119498>.

■ Author

Precious Ojo is a gap-year student from Nigeria. She is deeply passionate about women's health and biomedical science. She is fascinated by understanding reproductive health, uncovering the mechanisms behind different diseases, and exploring how to improve women's well-being. She plans to major in biology and minor in global health to pursue these interests.

Exploring Venture Capital Investment Impact on Innovation

Safina Kajani

Castilleja School, 1310 Bryant St, Palo Alto, CA, 94301, USA; safinakajani@gmail.com

ABSTRACT: It is widely held that venture capital investments play an important role in stimulating innovation; however, most of the existing research centers around venture capital funding of innovation. Different quantitative empirical studies on the topic are utilized in this paper to achieve a qualitative, in-depth understanding of the effect venture capital investment has on innovation and vice versa. In this research, the intersection of innovation and venture capital will be explored under three thematic frameworks: patent count, productivity & patent count, and persistence of innovation. This paper supports the innovation-first hypothesis as it suggests that venture capital investment may not spur innovation, but instead, likely benefits from innovation by selecting firms with high innovative capabilities, as they have high potential for growth. In fact, venture capital investment can often prevent further innovation to monetizing existing innovation.

KEYWORDS: Economics, Venture Capital, Innovation, Patents, Productivity.

■ Introduction

As an industry, venture capital was created to close the gap between institutional funding and early-stage companies, often referred to as startups. Venture capitalists make investments in startups in the form of equity.¹ Thus, they partake in the profits and losses from the company's innovation.¹ Venture capital plays an important role in driving growth and is likely to continue to fuel growth and prosperity.² As a result, it has played a crucial role in changing the trajectory of the innovation economy around the world. It has funded the launch of major markets and leading companies like Google, Oracle, and Apple.^{3,4} Other leading venture-backed companies include Alphabet, Facebook, Microsoft, and Alibaba.⁵ However, picking a winning company continues to remain one of the greatest challenges for a venture capitalist.⁴ Innovation and technological progress can be difficult, taking years to make a major impact on the market before finding an unexpected result.² However, multiple studies have found that venture-backed firms grow faster, have better financial results, and report more innovation than firms that are not venture-backed.⁶ To effectively encourage innovation, venture capitalists subsidize new firms and provide them with capital, along with managerial resources.⁷ Documented investment cycles on venture capital have often shown a relationship between radical innovations and financial market activity.⁸

In addition to venture capitalists, the other element to consider is governmental impact on innovation. If this paper finds that venture capital investment does, in fact, help spur innovation, there are multiple policy implications. Economists and policymakers have agreed that greater innovation is needed, especially due to lagging productivity growth in multiple Western nations.⁹ Policymakers should consider the impact venture capital has on an economy due to its ability to encourage innovation.⁶ The government is critical to fueling the venture capital industry, which can, in turn, promote economic growth, job creation, and better standards of living.¹⁰ To promote in-

novation, governments can make policies that help facilitate venture capital investment.⁶ In addition, the government can take steps to ensure that the legal ecosystem is favorable for venture capital to drive more investments, and thus increase innovation.⁶

While it is clear that innovation is a crucial element in venture capital, the measures of innovation can vary. A study by Kortum and Lerner,¹¹ a study conducted by Engel and Keilbach,¹² and a study by Caselli, Gatti, and Perrini¹³ attempt to use patent count to measure innovation because innovation as an ideology is difficult to quantify. Patent count measures the growth of new technology even if it has not been adopted yet.⁷ Therefore, it is believed to be an effective measure of innovation, as newly developed technology does not necessarily have to be put into use and be adopted to be considered innovation. In addition, patents granted to companies backed by venture capitalists are usually cited more often, implying that venture capitalists do play a role in innovation. Venture capitalists helping stimulate patent count is associated with the "VC-first hypothesis"⁷ created by Hirukawa and Ueda⁷ to explain the idea that venture capitalists spur innovation.

However, Hirukawa and Ueda⁷ bring to the surface the point that this approach is one-sided, as it fails to consider the opposite: that innovation opportunities lead to more firms and start-ups tapping into that new opportunity. This will then lead to a greater desire among venture capitalists to invest in these new startups, accelerating the growth in the venture capital market. Hirukawa and Ueda described this concept as the "innovation first hypothesis."⁷ This concept essentially states that venture capitalists choose to invest in more innovative firms. Both ideologies must be considered when understanding the relationship between venture capital investment and innovation.

Some researchers mention how innovation can exist without the creation of any patents.^{2,7,14,15} Instead, existing technologies can exist uniquely, creating a new product or service without

having to create a new technology.² Arvanitis and Stucki¹⁴ found that the two main reasons why patents count as a measure of innovation can be limited. Firstly, not all innovations can be patented, and secondly, the process of applying for and receiving a patent can take time and is associated with high costs.¹⁴ The main disadvantage of only measuring patents is that many other forms of innovation are not captured.¹⁵ To address this issue, Hirukawa and Ueda⁷ measure innovation in two different ways: total factor productivity and patent count. Unlike patent counts, total factor productivity tests innovation and growth resulting from adopting new technology.⁷ Therefore, innovation driven by utilizing existing technology in a new way will still be accounted for through this approach. It is crucial to account for both measures of innovation to achieve a well-rounded understanding of venture capital and innovation.

Others argue that labor productivity growth could be a better measure of innovation than total factor productivity growth, as it does not require non-labor inputs to be measured. These non-labor inputs include capital, material, and energy. The measurements of non-labor inputs are difficult, so the estimated total factor productivity is greatly dependent on the method used to measure it.¹⁶ On the other hand, labor productivity growth is not the most accurate measure of innovation either. Labor productivity only measures one aspect of productivity, unlike total factor productivity. Labor productivity could increase simply due to other productive inputs being used instead of labor inputs, even when there is no actual improvement in efficiency. Therefore, the proportion of the productive inputs must remain the same in order for labor productivity to be an accurate measure of innovation.¹⁶ Thus, even though each measure of innovation has its drawbacks, it is important to consider each of them when evaluating the impact on innovation, as they each fit into a slightly different, but crucial, definition of innovation.

However, venture capital does not always have positive impacts. Many potential problems exist between companies and venture capitalists, including conflicts with the allocation of effort, information asymmetries, and the valuation of the commitment the financier has to the company's development.¹³ Therefore, it is critical to consider that venture capital could have a neutral or even negative impact on innovation. To solve this, the difference in innovation in venture-backed and non-venture-backed firms has to be reviewed further by understanding the effect on innovation before and after funding. Innovation should undoubtedly be found before venture capitalist involvement if the venture capitalists are searching for innovation in potential firms. Since venture capital is designed to aid firms in using innovation to give them an advantage in the market, innovation should also be evaluated after investments from venture capitalists. Thus, it is necessary to test if innovation is sustained after venture capital involvement.¹³

It is also crucial to consider that venture capital investment may have a different effect on innovation in the short-term activities of firms versus the long-term activities of firms. This is especially important because venture capital investment may stimulate innovation in the long-term, even if it does not have an effect in early stages.¹⁴ Thus, the time horizon of the invest-

ment needs to be taken into account. The innovative results one year after venture capital investment may vary from four years post-investment. Therefore, persistence of innovation activities must be taken into consideration.

Other researchers have focused specifically on how venture capital investment promotes open innovation or corporate innovation.^{17,18} Additionally, some researchers go as far as stating that innovation is no longer consistently driven by research and development within corporations, but, instead, venture capitalists are needed to spur breakthrough innovation.¹⁹ In 2021, Yi, Wang, Lyu, and Xia conducted a study to determine whether venture capital investment promotes open innovation in China. Open innovation is a process in which sharing and combining innovation are used to reduce development costs, prevent long-term development cycles, and improve efficiency in innovation.¹⁷ They found that venture capital investment does have a positive effect on open innovation, but the extent varies in different regions.¹⁷ In 2025, Wang, Ma, Yu, and Guo explored how venture capital investment influences corporate innovation. They concluded that venture capital investment does positively affect corporate innovation, but the effect is more significant when the firms have intense industry competition or lack financial resources. Their findings also suggest that earlier venture capital investment has a more significant effect on the innovation of the firm.¹⁸ Lee, Oh, Shin, and Yoon²⁰ also determined that venture capital investment has a positive influence on corporate innovation even in areas with government intervention like Korea.²⁰

While there are multiple studies researching the way in which venture capital affects certain aspects or sections of innovation, there is a lack of resources discussing how venture capital affects innovation using multiple definitions of innovation. Considering all of the above, the present paper is a synthesized review of the empirical literature in order to explore the intersection of venture capital and innovation through patent count as innovation, productivity as innovation, and short-term and long-term effects on innovation qualitatively, and discuss whether venture capital or innovation comes first. This paper utilizes existing quantitative empirical studies to extrapolate a qualitative perspective of venture capitalists' effect on innovation through three thematic analytical components:

1. Patent count

Venture-backed firms patent more than non-venture-backed firms, but this exists before the venture capital investment.

2. Productivity and patent count

The "innovation-first hypothesis"⁷ is supported in regard to both total factor productivity and patent count.

3. Persistence of innovation

Venture capitalists do not spur innovation in initial innovative activities or the persistence of innovative activities.

The findings from the three themes explored inform a synthesized review of the effect venture capital has on innovation.

■ Methods

This paper will perform a thematic analysis based on secondary data from existing literature. For the data accessed, a

keyword search has been performed to identify the necessary quantitative literature regarding 'venture capital influence on innovation'. This search finds three relevant articles, but a greater selection of literature is needed. Therefore, 'innovation and venture capital' is used as a more general key search term, resulting in five more articles to be found. Each of the articles is read in depth to choose a more selective list of six articles that have the appropriate quantitative research needed for the study.

The chosen six papers are then categorized into three themes after a deeper analysis of their research inquiry. Each of the papers within each theme used different empirical models to test innovation. The organization of the papers into the three themes (patent count, productivity & patent count, and persistence of innovation) is illustrated in Table 1, Table 2, and Table 3. Each paper is categorized into one of the three themes to be analyzed individually, to eventually gather a holistic understanding of the theme, and to inform a review of the venture capital effect on innovation.

Table 1: Organization of articles into patent count thematic framework.

Themes	Title	Authors & Year of Publication	Research Inquiry	Methods	Limitations
Patent Count	1. Does venture capital spur innovation?	1. Samuel Kortum & Josh Lerner (1998)	1. Direct impact of patent count on innovation in 20 industries from 1965 to 1992	1. Linearized patent production function with VC investment and R&D as inputs	1. Limited analysis with a simple stylized model because they were the first in the field, and they only used patent count as an innovation measure. ¹¹
Patent Count	2. Firm-level implications of early-stage venture capital investment—an empirical investigation	2. Dirk Engel, Max Keilbach (2006)	2. Difference in innovation and firm growth in 21,541 young German firms with 0.66% venture-backed firms before and after venture capital involvement	2. Statistical matching procedures, probit estimation	2. They used only patent count as a measure of innovation, which is a narrow measure. ¹²
Patent Count	3. Are venture capitalists a catalyst for innovation	3. Stefano Caselli, Stefano Gatti, & Francesco Perrini (2008)	3. Difference in innovation and firm growth in 37 venture-backed and 37 non-venture-backed firms that got an IPO in Italy in 1995-2004 by measuring innovation before and after venture capital involvement to test if innovation existed before venture capital intervention or if venture capital was spurring the innovation, as well as measuring firm growth	3. Statistical matching procedure, Selection process with consideration to the time horizon model on the development of patents	3. They used only patent count as a measure of innovation, and findings can only be generalized to venture-backed firms that use IPO as an exit strategy. ¹³

The patent count thematic framework contains three articles: Kortum and Lerner's¹¹ study in 1998; Engel and Keilbach's¹² study in 2006; and Caselli, Gatti, and Perrini's¹³ study in 2008. Each article has its own research focus and limitations, but all of them use patent count as the measure of innovation.

Table 2: Organization of articles into productivity and patent counts thematic framework.

Themes	Title	Author & Year of Publication	Research Inquiry	Methods	Limitations
Productivity & Patent Count	1. Venture capital and industrial "innovation"	1. Masako Ueda & Masayuki Hirukawa (2008)	1. Re-test Kortum and Lerner's study on patent count in 19 US manufacturing industries with an extended time period of 1968-2001, and additionally test for total factor productivity growth and labor productivity growth	1. Linearized patent production function with VC investment and R&D as inputs *Use instrumental variables to lessen bias	1. They patent count, total factor productivity, and labor productivity as measures of innovation, each of which each have their own drawbacks. Additionally, when extending Kortum and Lerner (1998) to 2001, they had to hand-collect observations for data with undeterminable SIC codes and had to extend the database themselves, as the NBER discontinued in 1997. (They used the same method the NBER used to construct their database). ¹⁴
Productivity & Patent Count	2. Venture capital and innovation: which is first	2. Masako Ueda & Masayuki Hirukawa (2011)	2. Venture capital investment impact on total factor productivity growth and patent count from 1968 to 2001 in the manufacturing industry	2. Panel AR (autoregressive) models, including testing for Granger-type causality	2. Used total factor productivity and patent count as measures of innovation, which each have their own limitations by definition. ¹⁵

The productivity and patent count thematic framework contains two articles: Ueda and Hirukawa's¹⁶ study in 2008 and Ueda and Hirukawa's⁷ study in 2011. Each article has its own research focus and limitations, but both use productivity (total factor productivity or labor productivity) and patent count as measures of innovation.

Table 3: Organization of articles into the persistence of innovation thematic framework.

Themes	Title	Authors & Year of Publication	Research Inquiry	Methods	Limitations
Persistence of Innovation	1. The impact of venture capitalism on the persistence of innovation activities of startups	1. Spyros Arvanitis & Tobias Stucki (2013)	1. Early stage venture capital effect on short-term and long-term innovation of start-ups by analyzing startups founded in Switzerland in 1996/7 over a span of 10 years *Initial innovation activities are defined as 3 years after startup creation	1. Selection model and propensity score matching model	1. There are natural limitations to the selection model and propensity score matching model used in the study. ¹⁴

The persistence of innovation thematic framework contains one article: Arvanitis and Stucki's¹⁴ study in 2013 study. The article used the selection model and propensity score matching model as methods to measure the persistence of innovation.

In this paper, there is an important distinction between the selection effects and the treatment effects. In the six papers, venture-backed firms tended to have higher levels of innovation and non-venture-backed firms tended to be less innovative, in general. So, there is consistent evidence across the six review papers that venture capitalists tend to select more innovative firms. Thus, the selection effects are strong. However, the treatment effects were mixed, as some of the papers found that venture capital did spur certain measures of innovation, while others found no such evidence. The majority of the papers that did not find evidence were using a matched pair system or time-lag system that allowed them to measure innovation before and after venture capital investment.

The methodology used in this paper has various limitations. Firstly, this paper only uses existing data, each with its own limitations, to inform the understanding of venture capital's impact on innovation (Table 1, Table 2, and Table 3). Secondly, each theme uses only a limited number of articles in order to create a central concept on a subset of innovation. Thus, the restricted search categories may be a limitation that had an effect on the findings of the study. Additionally, the six papers used to understand each thematic framework are relatively old and may not reflect the most recent developments in the field. Therefore, it does not aim at empirical generalization, but a synthesized review of the existing literature in the field. However, venture capital's impact on innovation may differ across industries, but this paper doesn't explore industry differences, which may be a limitation of the findings. To mitigate the existing limitations, this paper uses a wider variety of data across themes to ensure a more informed conclusion on the overall effect of venture capital on innovation.

■ Results and Discussion

3.1. Patent Count:

To understand the direct impact that venture capital has on patent count as a measure of innovation, existing studies provide important context. In 1998, Kortum and Lerner

conducted a study on the intersection of venture capital and innovation, utilizing patent count as their measure of innovation. Their study was the first conducted on the topic of venture capital and innovation. It laid the framework for all future studies. They explored twenty industries in the United States manufacturing sector from 1965 to 1992. The industry of a firm was determined by the technological classification of the patent. Their study utilized reduced-form regressions to analyze the impact venture capital funding had on the number of innovations granted patents. Additionally, they controlled for research and development (R&D) funding. Their research found that venture capital funding is correlated with an increase in patenting. However, the simple reduced-form regression model Kortum and Lerner¹¹ utilized in their study might have overstated the effect of venture capital funding on innovation. Instead, venture capital and patents could be affected by the arrival of entrepreneurial opportunities, an additional unobservable variable. To address this, they considered the discontinuity in the venture capital industry in 1979 when the Employee Retirement Income Security Act was clarified, consequently allowing pensions to invest in venture capital and leading to an increase in venture capital funds.¹¹

The exogenous change allows for a separation of the arrival of entrepreneurial opportunities and the role of venture capital. Kortum and Lerner¹¹ explored this through instrumental variable regressions. In addition, they estimate venture capital funding's impact on the patent-R&D ratio, instead of solely patenting, to address the majority of the causality problems. Even after considering the causality concerns, venture capital funding was still found to have a significant positive impact on innovation. Their results imply that compared to a dollar of traditional R&D, a dollar of venture capital would be ten times more successful in stimulating patents. Yet, it is necessary to consider whether patents, the dependent variable utilized in the analysis, are an effective measure of innovation. It is possible that venture capital funding spurs patents without indicating any relationship between venture capital and innovation. Firms backed by venture capital may have a higher number of patents granted than non-venture-backed firms solely because they are concerned that their investor may expropriate their idea or hope to impress possible investors by obtaining a patent even if it is of marginal worth.¹¹

The increase in patenting could simply be caused by the venture financing process's natural inclination to patent innovations. Kortum and Lerner¹¹ studied 530 venture-backed and non-venture-backed firms in Middlesex County, Massachusetts, to understand if the increase in patent filing associated with venture capital was solely caused by a greater propensity to patent. They found that companies backed by venture capital do patent more, but not to the extent that the economic importance of the patent is diluted. Venture-backed patents just tend to be fought more aggressively and to be cited more in other patents. This suggests that the increase in patent count is not simply due to the greater inclination of venture-backed firms to patent.¹¹

When considering the validity of Kortum and Lerner's¹¹ research, it is important to consider if their study accurately

ly measured a strong positive correlation between venture capital and patents. There was an issue in the way they only estimated the impact venture capital has on patenting without considering research and development, as that would lead to a drastically positively skewed result. However, after further analyzing their own research and results, they realized this issue and instead estimated the impact venture capital had on the ratio of patents to R&D, still finding a positive correlation. There was also a concern that other unmeasured variables that had not been taken into account were affecting the positive relationship between venture capital and patents. Kortum and Lerner¹¹ did eventually address this, though, and considered that the unmeasured variable of entrepreneurial opportunities could have been impacting the correlation between venture capital and patents. But this theory was disproved by utilizing the discontinuity in 1979. Therefore, the positive impact of venture capital on patent count that Kortum and Lerner¹¹ found seems to have significant validity.

Additionally, it must be understood that Kortum and Lerner¹¹ assumed a proportional relationship between patents and innovation. However, in reality, this is not always the case, as multiple different innovative characteristics or variants can be part of the same patent. At the same time, though, if the two innovations are fundamentally different, separate patent applications will be needed, but this is not always the case. It is also crucial to consider that patenting, at a minimum, is a somewhat effective measure of innovation because it accounts for technological innovation that has not been created and may not have been adopted yet.⁷ Therefore, while patenting an innovation may not have a linear relationship, a direct link between patenting and innovation does exist. Thus, even though a strongly positive relationship was found between patenting and venture capital, the implied relationship between venture capital and innovation may be exaggerated in their results. Instead, their study suggests that a milder positive correlation between venture capital and innovation actually exists.

Another study conducted by Engel and Keilbach¹² in 2006 also analyzed the relationship between venture capital investment and innovation using patent count as a measure of innovation. The study also examines the impact venture capital investment has on firm growth. Engel and Keilbach¹² found the number of employees of a firm and sales to be almost linearly correlated, possibly signifying a strong correlation between the variables, so they chose to use employment level as the measure of firm growth. They used a dataset of 21,541 young German firms that were founded between 1995 and 1998. 0.66% of these young German firms were venture-backed. This dataset allowed them to identify the exact movement of the venture capital investment and allowed them to identify "twin" firms, with one venture-backed firm and one non-venture-backed firm. The use of matched pairs helped correct the industry and timing biases that may have existed in Kortum and Lerner's data.¹²

Their study finds that venture-backed firms have more patent applications than non-venture-backed firms, but these applications are obtained before the involvement of venture capitalists. However, after the venture capital investment, the

patent count has an incredibly weak significant difference between venture-backed firms and non-venture-backed firms. The likelihood of a venture-backed firm applying for a patent is greater, but the difference is insignificant, so no correlation between venture capital investment and driving innovation can be found. On the other hand, venture capital investment does correlate with greater employment growth rates. The average growth rate for venture-backed firms is twice the rate for non-venture-backed firms. These findings suggest that venture capitalists look to select firms with the greatest growth potential, and innovative potential is an important factor in finding these firms. So, the higher levels of innovation found in venture-backed firms are likely due to the venture capital selection process rather than being a result of the funding. Thus, it is likely that venture capitalists mainly aid their firms in commercialization, instead of aiding with further innovation, as that would maximize return.¹² While it is very possible that patent applications do not account for all innovation, there is a correlation between patents and innovation. Thus, Engel and Keilbach's¹² findings of a greater number of patent applications in venture-backed firms before the venture capital investment came can be concluded.

Engel and Keilbach's¹² findings, however, are contradictory to Kortum and Lerner's¹¹ study, which found that venture capital investment does have a correlation with innovation measured as patent count. The difference may be because Kortum and Lerner¹¹ did not examine innovation before and after venture capital investment. Instead, they solely looked at whether venture-backed firms had a higher patent count and found this to be true. If this is the case, then their findings are in line with the study conducted by Engel and Keilbach.¹² Both Engel and Keilbach¹² and Kortum and Lerner¹¹ found that venture-backed firms do patent more. However, when Engel and Keilbach¹² looked at the difference in innovation between venture-backed and non-venture-backed firms before and after investment, they found the difference to be significant only before venture capital investment. Thus, it is likely that Kortum and Lerner's¹¹ findings are correct in that venture-backed firms patent more, but due to the method of their study, they did not come to the conclusion that the greater number of patents disappears after venture capital investment. Therefore, this paper finds that venture-backed firms do patent more, but these are obtained before venture capital investment, implying that venture capital selects innovative firms rather than spurring innovation.

In 2008, Caselli, Gatti, and Perrini¹³ conducted a similar study to analyze the intersection between venture capital and innovation as well as firm growth. In regard to innovation, they examine whether innovation exists before and after venture funding, focusing on whether innovation is important for intervention by venture capital and whether venture capitalists can sustain and improve the firm's innovation. Innovation is measured as the number of registered patents, and firm growth is defined as the net sales growth rates. They utilized a sample of 37 Italian venture-backed firms and 37 of 112 non-venture-backed twin firms that all went public from 1995 to 2005 on the Italian Stock Exchange. The dataset allows for a list of all

innovations registered to be built, including innovations that were not registered in Italy. Caselli, Gatti, and Perrini¹³ compared the patent count in the five years before the firm received venture capital investment and the period between the initial venture capital investment and 2004. The non-venture-backed twin firms are examined during the same periods, even though the discontinuity of the venture capital investment does not exist.¹³

Caselli, Gatti, and Perrini¹³ found that venture capital investment does not promote innovation. Before the initial venture capital investment, venture-backed firms register 0.3327 patents more per year than non-venture-backed firms, but after the investment, venture-backed firms register 0.53825 fewer patents than their non-venture-backed counterparts. However, venture capital investment was found to be correlated with sales growth, as venture-backed firms had a sales increase of 5.5% over non-venture-backed firms. This implies that innovation is an important factor that is considered during the venture capitalists' selection process, but once funding begins, venture capitalists do not advance or increase innovation. In fact, it may actually block the firm's continued innovation. Instead, after funding begins, venture capitalists help aid the improvement of economic aspects of the firm, such as firm growth. These results could be interpreted as such that venture capitalists value innovation as a factor in growth potential, but after the investment, they focus their efforts on maximizing firm growth in the hope of a future deal.¹³ It could also suggest that venture capitalists see high innovation activities as a sign of potential monetary success and see that as an opportunity to exploit instead of looking to further advance the innovation of the firm.

When considering the findings of Caselli, Gatti, and Perrini,¹³ it is necessary to acknowledge that their dataset only included firms that had gone public on the Italian Stock Exchange. Thus, their results cannot be generalized to all venture-backed firms, just the venture-backed firms whose exit strategy was an IPO. Therefore, the great shift towards sales and away from innovation may be solely correlated with venture-backed firms that went to an IPO.¹³ However, Engel and Keilbach,¹² who did not solely look at firms that went public, found similar results that found that venture-backed firms patent more, but only before venture capital investment, while firm growth increases after venture capital involvement. Thus, this can be concluded for all venture-backed firms.

It is possible that the slight variation in results from Engel and Keilbach's¹² study and Kortum and Lerner's¹¹ study is due to the different structure and goals of venture capitalists in different countries.¹³ It could also be because Caselli, Gatti, and Perrini¹³ only studied firms that had gone public on the Italian Stock Exchange. However, overall, the results are consistent with those found by Engel and Keilbach¹² and Kortum and Lerner,¹¹ suggesting that venture-backed firms patent more than their non-venture-backed counterparts but only before venture capital involvement.

However, these findings only consider one measure of innovation, patent count, that has its own inherent limitations. Patents legally protect the innovation of a firm from being

copied, so it is incredibly likely that these firms would patent if they believed it was a worthwhile way of protecting their innovation.¹² However, it is possible that firms do not patent even if an innovation exists. This could be due to the fact that not all innovations can be patented. It is also possible that a patentable innovation is not patented, as the patent process is too long in the firm's eyes, and they would, instead, prefer to use secrecy or speed to protect their innovation. Lastly, a firm may not want to disclose knowledge in the innovation that would inevitably be disclosed in the patent application.¹² It is also necessary to consider that firms may apply for multiple patents if aspects of the innovation are fundamentally different, increasing the number of patent applications without increasing the amount of innovation. Yet, it is possible that if the different applications had gotten approved, that would signify that the innovations are fundamentally different and can be considered two separate innovations. It is also important to consider if the increase in patenting was simply due to a natural tendency of the venture capitalists to file for patents without a change in the innovation, but Kortum and Lerner¹¹ found no validity to this concern. Patenting must be at least a somewhat effective measure of innovation because it accounts for technological innovation that has not been created and may not have been adopted yet.⁷ Therefore, we must extend the measure of innovation to productivity in addition to patent count.

3.2. Productivity and Patent Count:

To understand the impact venture capital has on innovation, as measured by both patent count and productivity, existing studies provide important information. In 2008, Ueda and Hirukawa¹⁶ conducted a study by modifying the length and variables of Kortum and Lerner's study.¹¹ They extended the length of the sample to 2001 to include the NASDAQ bubble period, leading to a dataset between 1968 and 2001. They looked at nineteen U.S. manufacturing industries and utilized a linear patent production function with R&D and venture capital investment as inputs and instrumental variables to lessen bias, the same model Kortum and Lerner¹¹ used. Ueda and Hirukawa¹⁶ measure innovation as patent count, their total factor productivity growth, NBER total factor productivity growth, labor productivity growth, and production labor productivity growth. Their study finds that venture capital investment does have a significant positive effect on patent count even in the extended time period, reconfirming Kortum and Lerner's¹¹ results. A positive relationship between venture capital investment and labor productivity growth was also found, but no significant positive relationship was found between venture capital investment and total factor productivity growth.¹⁶

Extending the sample until 2001 had unexpected results. Hirukawa and Ueda¹⁶ found that venture capital investments, in comparison to R&D, are more effective in bringing about patents. This effect only increased during 1993 to 2001, the extended period of the sample, including the NASDAQ bubble period. When they disregard the period from 1993 to 2001 in the sample, their coefficients on the ratio of venture capital to R&D are significantly smaller. Therefore, the positive correlation between venture capital investment and patent count

simply increased during the 90s.¹⁶ It is also possible that this data implies that the NASDAQ bubble period increased innovation when measured as patent count. So, there may be a correlation between bubble periods and an increase in innovation, though this must be studied further for a more concrete answer to be found.

However, even when the sample period was restricted to 1968 to 1992, the estimated coefficients found by Hirukawa and Ueda¹⁶ were still considerably higher than the estimated coefficients found by Kortum and Lerner.¹¹ This is most likely due to survivorship bias within their data set, which would occur as a greater percentage of the older data points would be surrounding venture capital funds that have survived. The more successful venture capital funds will probably also make higher-quality investments.¹⁶ It is also necessary to consider whether firms backed by venture capital patent low-quality innovations as a strategy. Hirukawa and Ueda¹⁶ did not find that the quality of patents declines with an increase in venture capital. However, the breadth of knowledge that the patents contain was found to have a positive correlation with venture capital investment. These results may be a benefit of more access to technology and knowledge bases given to venture-backed firms, which are surfacing in the way in which those firms patent.¹⁶

In regard to productivity, Hirukawa and Ueda¹⁶ found a positive correlation between venture capital investment and labor productivity growth but did not find a significant correlation between venture capital and total factor productivity growth. This could be due to the measurability of total factor productivity and labor productivity, as well as their relationship with innovation. The main issue with total factor productivity growth as a measure of innovation is that it is difficult to measure quality improvements. To identify quality improvements, a great amount of knowledge is needed in assessing product quality and identifying the change in product quality. Therefore, this aspect of total factor productivity growth is often not incorporated, unlike cost-reducing innovation. This means that total factor productivity growth may be underrepresented.¹⁶ Thus, even though Hirukawa and Ueda¹⁶ did not find a positive correlation between venture capital investment and total factor productivity growth, it is possible that a correlation does exist and that it was just underrepresented due to the difficulty in measuring quality improvement.

On the other hand, some researchers believe that labor productivity growth is a better measure of innovation in comparison to total factor productivity growth, as nonlabor inputs, i.e., capital, material, and energy, do not need to be measured.^{7,16} Most of these nonlabor inputs are difficult to measure, so the estimate of the total factor productivity growth relies on whether the nonlabor inputs are measured accurately. However, there are also criticisms of labor productivity growth as a measure of innovation, as it only partially measures productivity. An unproportional change in any labor-productive input would change the labor productivity growth. So, the proportion of each would need to stay constant in order to successfully measure innovation. Thus, the positive impact venture capital investment has on labor productivity is most likely just driven

by substitutions of labor.¹⁶ Therefore, while a positive relationship between venture capital investment and labor productivity growth, this paper does not conclude that this relationship would be indicative of a relationship between venture capital and innovation. Labor productivity growth only partially measures innovation and can also positively affect the substitution of input factors, which are unrelated to innovation.

Another study was conducted by Ueda and Hirukawa⁷ in 2011 to understand the causality of venture capital investment and innovation. They considered both the “VC-first hypothesis” in which venture capital investment spurs innovation and the “innovation-first hypothesis” in which innovative opportunities are exploited by firms and venture capitalists invest in these innovative firms. They analyze five different U.S. manufacturing industries that account for 88% of all venture capital investment from 1963 to 2001. Ueda and Hirukawa⁷ measure innovation as both total factor productivity growth and patent count. Innovation is examined with a 1-year lag, 2-year lag, and 4-year lag in both first-round and follow-on-round investments.⁷

In the premise of total factor productivity growth, their study found weak evidence of the “VC-first hypothesis.”⁷ The 1-year lag of the first-round investment found a negative relationship between venture capital spurring innovation, but the 2-year lag found a positive relationship between venture capital spurring innovation. No evidence of venture capital spurring innovation was found in the follow-on round investment. However, Hirukawa and Ueda⁷ found that innovation does lead to venture capital investment, in the context of total factor productivity growth. While the 1-year lag of the first round and the 1-year lag of the follow-on investment had a negative correlation with total factor productivity growth, the 2-year lag of the first-round investment and the follow-on investment both had a significant positive relationship with total factor productivity growth. Thus, this finding provides evidence of the “innovation-first hypothesis”⁷ when considering total factor productivity.

Conversely, Hirukawa and Ueda⁷ found no relationship between venture capital investment and innovation regarding both the “VC-first hypothesis”⁷ and the “innovation-first hypothesis.”⁷ The patent count was found to be lower in both the 1-year lag of the first round and the follow-on round of venture capital investment in each industry. After the 1-year lag, no evidence was found to support the “VC-first hypothesis”⁷ in regard to patent count. Evidence was found that supported the “innovation-first hypothesis,”⁷ however, it was not significant. Therefore, no support was found for innovation spurring venture capital investment.⁷

Interestingly, a negative relationship is found with both total factor productivity growth and patent counts in the 1-year lagged venture capital investments, which suggests that venture capital slows down innovation in the first year after investment. This negative relationship may be due to venture capitalists supplying firms with immense resources, leading to inefficiencies in strategic decision-making immediately after the investment. This result appears to be consistent with the studies conducted by Engel and Keilbach¹² and Caselli, Gatti, and Perrini,¹³ who found that patenting decreases after venture

capital investment due to a desire to monetize existing innovation rather than create innovation.⁷

When considering the validity of Hirukawa and Ueda’s¹⁶ and Hirukawa and Ueda’s⁷ conclusions, we must first understand the drawbacks of patent count and total factor productivity as a measure of innovation, as they found a positive relationship between total factor productivity growth and venture capital investment. At the bare minimum, patent count has to be a somewhat effective measure of innovation, as it accounts for technological innovation that has not been created in the past, even if that technology has not yet been adopted.⁷ So, even though patent count does not always necessarily account for existing technology adapted in innovative ways, it is still a measure of innovation. Therefore, a positive relationship between venture capital and innovation, as patent count can be concluded from Hirukawa and Ueda’s¹⁶ study. Total factor productivity measures innovation resulting from adopting new or existing technology, which is not always measured by patent count.⁷ However, an accurate measure of total factor productivity is heavily reliant on the measurement of nonlabor inputs that affect total factor productivity, which are difficult to measure.¹⁶ While total factor productivity is not a perfect measure of innovation, it is still a sufficient measure of innovation, and therefore, this paper concludes that evidence of innovation spurring venture capital investment regarding total factor productivity in Hirukawa and Ueda’s⁷ study can be found.

However, these findings are initially not consistent with the findings of Hirukawa and Ueda,¹⁶ which found a positive correlation with venture capital and patent counts, but no correlation between venture capital and total factor productivity. It is possible that this could be due to the challenges in measuring total factor productivity. It is quite difficult to identify and assess quality improvements, so it is often not incorporated in the measure of total factor productivity growth. Thus, the results found by Hirukawa and Ueda¹⁶ in regard to total factor productivity could be underrepresented.¹⁶ Therefore, the results with total factor productivity growth are consistent in both Hirukawa and Ueda¹⁶ and Hirukawa and Ueda,⁷ supporting the “innovation-first hypothesis.”⁷

Yet, there is still a discrepancy with the findings of Hirukawa and Ueda⁷ with the other studies on the relationship between venture capital and patent count. Hirukawa and Ueda,¹⁶ Engel and Keilbach,¹² and Caselli, Gatti, and Perrini¹³ all found a positive relationship between venture capital investment and patent counts. Engel and Keilbach¹² and Caselli, Gatti, and Perrini¹³ additionally found that innovations spur venture capital investment in regard to patents. This discrepancy is likely due to the method by which Hirukawa and Ueda⁷ conducted their study. The patent filing process is extensive and takes a great period of time before being finalized. Therefore, the method of measuring innovation at the 1-year lag, 2-year lag, and 4-year lag of venture capital investment may not accurately capture the relationship between venture capital investment and patent count. The 4-year lag timeline is probably too short, proportionally to the length of the patent filing process. This is likely consistent with Hirukawa and Ueda,⁷ as they found positive, but insignificant evidence of the “innovation-first hy-

pothesis⁷⁷ in regard to patent count. This insignificance could be due to the underrepresentation of patents due to the short time period considered in the study. Therefore, this paper finds that innovation does lead to venture capital investment in regard to patent count.

This discrepancy could also be due to the varying locations of the samples of the studies. Hirukawa and Ueda⁷ conducted their study with data from the United States. However, Engel and Keilbach¹² utilized a German dataset, while Caselli, Gatti, and Perrini¹³ examined firms in Italy. The different goals of venture capitalists in different countries could be causing this discrepancy.¹³ How venture capitalists in the country choose venture capital firms to invest in could be heavily affected by the individual goals of that country.

3.3. Persistence of Innovation:

To understand the impact venture capital has on initial innovation and persistence of innovation, existing studies provide important context. In 2013, Arvanitis and Stucki¹⁴ conducted a study on whether venture capital stimulates early-stage innovation or persistence of innovative activities in start-ups. They utilized a representative cohort of all startups founded in 1996 or 1997 in Switzerland. To analyze the initial and continuous innovation activities, they measured innovation performance after three, six, and nine years after the firm's foundation for those that survived. In their study, initial innovative activities were defined as occurring three years after firm creation. Arvanitis and Stucki¹⁴ used five binary innovation measures: propensity of new products, propensity of modified products, propensity of either new or modified products, propensity of R&D activities, and propensity of patent applications.¹⁴ They found that venture-backed firms, on average, have a higher inclination to initial innovative activities in comparison to other firms in four of the five innovation measures. However, when Arvanitis and Stucki¹⁴ controlled for the innovation capabilities that the firms initially had, the statistical significance disappeared. Besides a weak effect on patenting, venture-backed firms do not have significantly higher innovation activities than non-venture-backed firms, even after matching. Thus, their study finds no positive correlation with venture capital investment and initial innovation activities.¹⁴

Arvanitis and Stucki¹⁴ also found no positive relationship between venture capital investment and the persistence of innovation activities. Before matching, venture-backed firms have a higher propensity for introducing new or modified products than non-venture-backed firms. However, after the initial stage and a longer period, there is no statistical significance for the persistence of innovation in any of the five measures of innovation. Thus, Arvanitis and Stucki¹⁴ found that venture capital investment does not affect the firm's innovativeness.¹⁴ It is possible that the insignificance found between persistence of innovation and venture capital investment could be due to the restrictive definition of persistence they used in the study. To test this, they utilize a less limited definition of persistence, defining persistence as innovative activities in two out of the three periods measured. Even though there were slight changes, the results were still statistically insignificant in regard to

all measures of innovation. Arvanitis and Stucki¹⁴ also considered whether the time period between the investment and the measurement of innovation was too short for real results to be found. However, no significant difference was found.¹⁴

In addition, Arvanitis and Stucki¹⁴ did not include the volume of funding in the definition of venture capital investment, even though innovation could have a positive relationship with the amount of funding. This change did have a statistically significant increase in patenting propensity after matching. The increase in patenting is logical, as patenting is a costly type of innovation, and a greater volume of funding is likely to increase patenting.¹⁴ Even though venture capital was not found to have a significant correlation with initial innovation or persistence of innovation, excluding a slight increase in patenting, the finding can signify validity in the "innovation first hypothesis."⁷⁷ The greater innovation propensity in venture-backed firms before matching could suggest that venture capitalists choose to fund startups with above-average innovation or at least the potential for above-average innovation. However, venture capital investment itself does not seem to be a stimulator of innovation.¹⁴ Thus, venture capitalists likely seek out innovative firms and benefit from that innovativeness to increase profit and scale instead of being the ones to spur the innovation.

These findings go hand in hand with the other studies included in this paper. Engel and Keilbach¹² and Caselli, Gatti, and Perrini¹³ found that innovation is greater in venture-backed firms in comparison to their non-venture-backed counterparts before they received venture capital investment. However, this disappeared after they were matched and received venture capital investment. At first glance, this seemed inconsistent with Hirukawa and Ueda¹⁶ and Kortum and Lerner,¹¹ who both found a positive correlation between venture capital investment and innovation as patent count. However, after deeper examination, these two studies that found that venture capital investment positively affects patent count did not consider the difference in innovation before and after venture capital investment. Instead, they simply found that venture-backed firms have a greater number of patents, which is consistent with the other results. Venture-backed firms are more innovative, but not due to the venture capital investment itself. On the contrary, it seems to be due to the selection process of the venture capitalists. They likely choose to fund more innovative ventures as they are likely to have more potential for profit and scaling. This is consistent with the findings of Engel and Keilbach¹² (2006) and Caselli, Gatti, and Perrini,¹³ who found that venture capital investment does spur firm growth even though it does not spur innovation. Hirukawa and Ueda⁷ found that innovation spurs venture capital investment in regard to total factor productivity, but not patent count, which is inconsistent with the other studies that found a positive relationship with patent count. This discrepancy is likely due to the method by which Hirukawa and Ueda⁷ conducted their study. The patent filing process is extensive and takes a long period of time before being finalized. Therefore, the method of measuring innovation at the 1-year lag, 2-year lag, and 4-year lag of venture capital investment may not accurately capture the relationship between venture capital investment and patent count, as the

timeline is too short. This is further supported by the fact that Hirukawa and Ueda⁷ found a positive, but insignificant relationship with patent count spurring venture capital investment, highlighting how this relationship could have been underrepresented. Thus, this paper finds support for the “innovation first hypothesis,”⁷ that venture capitalists chose more innovative firms to fund in regard to innovation, as both patent count and total factor productivity.

■ Conclusion

Since the selection effects are strong, but the treatment effects are mixed in the existing literature, this paper supports the “innovation-first hypothesis.”⁷ It suggests that capital may exploit innovation by selecting firms with high innovative capabilities that are likely to have more potential for profit and scaling instead of spurring innovation. Thus, this paper finds support for the “innovation-first hypothesis”⁷ in regard to both patent count and total factor productivity.

However, further research is needed as there are various limitations to this study. First, this paper has various data limitations. This study only utilizes data from existing studies that each have their own limitations to reach an understanding of venture capital's impact on innovation. In addition, each thematic framework only has a limited number of articles that are relatively old to inform the central concept of that theme. Secondly, this paper contains various scope limitations. A few of the articles used in this paper utilize a dataset that cannot be generalized to all venture capital investments, but rather a specific subset of venture capital. Some examples include venture capital investment in firms that went public through an IPO as an exit strategy, venture capital in specific countries, and venture capital during a bubble period. Thus, there may be specific nuances in each article that are not fully taken into account when developing a synthesized review of venture capital investment and innovation. To mitigate these methodological limitations, this paper uses a wide variety of data across themes to reach a more informed conclusion on the overall relationship between venture capital and innovation. This variety is also a strength as it allows for multiple lenses to be considered before reaching an informed conclusion.

The findings of this paper suggest multiple avenues for future research. There may be a difference in the relationship between venture capital and innovation across different industries or sectors, which was not a focus of this study. Future research could examine how various industries create nuance in regard to venture capital and innovation. Additionally, the papers utilized in each thematic framework are relatively old, as there is a lack of current research on the topic. While this study is a review of the empirical work up to this point, more primary empirical studies should be conducted to understand if the relationship between innovation and venture capital has changed. In this paper, there was a causality issue, which may be partly attributed to the lack of qualitative data in the field. Thus, more qualitative studies without any oversights of the existing papers' limitations would be an avenue for future research. There is also a minuscule possibility that venture capital

investment does have a mild effect on patent count in relation to funding intensity, but further research would need to be conducted to verify this. In addition, geographic variations in the dataset led to slight variations in the findings of the studies. Further research could be conducted on the ways in which geographic variation in venture capital affects the relationship between venture capital and innovation. In addition, Ueda and Hirukawa¹⁶ found that the NASDAQ bubble period had an interesting effect on innovation. Thus, a future study could be conducted focusing on how bubble periods affect the relationship between venture capital investment and innovation. Finally, the study conducted by Casselli, Gatti, and Perrini¹³ looked at a dataset of purely venture-backed and non-venture-backed companies that had gone public on the Italian Stock Exchange. Even though no extreme differences were found that affected the findings of this paper, future research could examine how venture capital's effect on innovation varies when the venture capitalist has an end goal of a specific exit strategy.

■ Acknowledgments

I would like to express my gratitude towards Dr. Iva Bimpli, my mentor, and Patrick Hao, my TA, for guiding me through this process. You both have been invaluable during this journey.

■ References

1. Popov, A., Roosenboom, P., Ichino, A., & Schivardi, F. (2012, July). *Venture capital and patented innovation: evidence from Europe*. Economic Policy.
2. Kressel, H., & Lento, T. V. (2010). *Investing in Dynamic Markets: Venture Capital in the Digital Age*. Cambridge University Press.
3. Kolympiris, C., & Kalaitzandonakes, N. (2013, July 4). *The geographic extent of venture capital externalities on innovation* (3). *An International Journal of Entrepreneurial Finance*. <https://doi.org/10.1080/13691066.2013.804749>
4. Ester, P. (2017). *Accelerators in Silicon Valley: Building Successful Startups*. Amsterdam University Press. <https://doi.org/10.2307/j.ctt1zrvhk7.7>
5. Lerner, J., & Nanda, R. (2020). Venture Capital's Role in Financing Innovation: What We Know and How Much We Still Need to Learn. *The Journal of Economic Perspectives*, 34(3).
6. Khan, N., Qu, H., Qu, J., Wei, C., & Wang, S. (2021, March 31). *Does Venture Capital Investment Spur Innovation? A Cross-Countries Analysis*. Sage Journals. <https://doi.org/10.1177/21582440211003087>
7. Hirukawa, M., & Ueda, M. (2011, October 10). *Venture Capital and Innovation: Which is first?* (Research Report No. 4). Pacific Economic Review. <https://doi.org/10.1111/j.1468-0106.2011.00557.x>
8. Nanda, R., & Rhodes-Kropf, M. (2017). Financing Risk and Innovation. *Management Science*, 63(4), 901-918. <https://doi.org/10.1287/mnsc.2015.2350>
9. Nanda, R., & Lerner, J. (2023). Venture capital and innovation. In *Handbook of the Economics of Corporate Finance: Private Equity and Entrepreneurial Finance* (pp. 77-105). <https://doi.org/10.1016/bs.hecf.2023.02.002>
10. Lee, P. (2022, December 31). *Enhancing the Innovative Capacity of Venture Capital*. University of California, Davis.
11. Kortum, S., & Lerner, J. (1998, December). *Does Venture Capital Spur Innovation?* (Research Report No. 6846). National Bureau of Economic Research. <https://doi.org/10.3386/w6846>
12. Engel, D., & Keilbach, M. (2007, March). *Firm-level implications of early-stage venture capital investment — An empirical investiga-*

- tion. *Journal of Empirical Finance*. <https://doi.org/10.1016/j.jempfin.2006.03.004>
13. Caselli, S., Gatti, S., & Perrini, F. (2008, December 25). *Are Venture Capitalists a Catalyst for Innovation?* *European Financial Management*. <https://doi.org/10.1111/j.1468-036X.2008.00445.x>
 14. Arvanitis, S., & Stucki, T. (2013, August). *The impact of venture capital on the persistence of innovation activities of start-ups* (4). *Small Business Economics*. <https://doi.org/10.1007/s11187-013-9499-3>
 15. Kerr, W. R., & Robert-Nicoud, F. (2020). Tech Clusters. *The Journal of Economic Perspectives*, 34(3), 50-76.
 16. Ueda, M., & Hirukawa, M. (2008, September). *Venture Capital and Industrial "Innovation."* SSRN. <http://doi.org/10.2139/ssrn.1242693>
 17. Yi, R., Wang, H., Lyu, B., & Xia, Q. (2021, May 17). *Does venture capital help to promote open innovation practice? Evidence from China*. Emerald Publishing. <https://doi.org/10.1108/EJIM-03-2021-0161>
 18. Wang, S., Ma, C., Yu, M., & Guo, X. (2025, March). *Impact of venture capital on innovation in key & core technology firms*. <https://doi.org/10.1016/j.pacfin.2025.102756>
 19. Srinivasan, S., Barchas, I., Gorenberg, M., & Simoudis, E. (2014). Venture Capital: Fueling the Innovation Economy. *Computer*, 47(8), 40-47. <https://doi.org/10.1109/MC.2014.230>
 20. Lee, K., Oh, F. D., Shin, D., & Yoon, H. (2022, May 25). *Does venture capital investment enhance corporate innovation? Evidence from Korea*. *Journal of Business Finance & Accounting*. <https://doi.org/10.1111/jbfa.12618>

■ Author

Safina Kajani is a student attending the Castilleja School in Palo Alto, California. Aside from being an accomplished equestrian and harpist, she is a problem solver with a passion to use her knowledge for the betterment of the broader community.

Animacy Decoding During Early Visual Processing Using Large-Scale EEG Data

Stewart S. Lee¹, Ioannis Delis

1. Stratford High School, 14555 Fern Dr, Houston, Texas, 77079, USA; trawets9788@gmail.com
Mentor: Ioannis Delis

ABSTRACT: The human visual system can rapidly recognize meaningful information from images, sometimes within a fraction of a second. Previous research has shown that animacy, the distinction between living and non-living objects, can be decoded from neural signals shortly after the stimulus presentation. However, the exact timing and spatial distribution of animacy in electroencephalography (EEG) remain under investigation. In this paper, I examined whether animacy can be decoded from EEG signals within the first 200 milliseconds after image presentation. Using the Naturalistic Object Dataset EEG (NOD-EEG), a large publicly available dataset containing responses to thousands of object images, I applied time-resolved decoding, permutation-based significance testing, Haufe-transformed forward-model scalp mapping, event-related potential (ERP) testing, and EEG-Behavior correlations. The results showed that animacy could be decoded above chance (>50%) beginning around 100 ms, with peak activity at approximately 200 ms. Haufe-transformed forward-model analyses also revealed the strongest activity over occipital and parieto-occipital areas (posterior scalp regions). Overall, these results indicate that animacy-related information emerges rapidly in EEG signals and appears in distributed posterior visual regions during the early stages of visual processing.

KEYWORDS: Behavioral and Social Sciences, Neuroscience, Electroencephalography, Animacy, Visual.

■ Introduction

The human visual system can quickly convert raw sensory input into meaningful representations of objects. Within just a few hundred milliseconds of seeing an image, the brain can recognize what the object is and use that information to guide behavior. Figuring out how and when these visual representations form over time is a major focus in visual neuroscience.¹ One key factor in object perception is animacy, which refers to whether an object is animate or inanimate. Animate objects are often more behaviorally important because they can signal potential danger or opportunities for social interaction.^{2,3} Because of this, the brain may prioritize animate objects more quickly or prioritize them over inanimate ones. Neuroimaging research using fMRI has shown that animate and inanimate objects are processed differently in the high-level visual and temporal cortex.^{2,3} However, while these studies provide strong evidence for where animacy is represented, they offer limited insight into when animacy information first emerges during visual processing.

Electroencephalography (EEG) is a valuable tool for studying early visual processing because it captures brain activity with very high temporal resolution. Previous EEG studies have shown that meaningful object information can be rapidly decoded from neural signals, even when images are presented briefly or under more naturalistic conditions.^{1,4} Certain event-related potential (ERP) components, such as the P1 (around 80-120 ms) and N170 (around 140-200 ms), are associated with early visual and categorical processing and have been linked to differences in how the brain responds to objects and faces.⁵

More recent work has demonstrated that animacy-related distinctions can be decoded from EEG signals within the first few hundred milliseconds after the stimulus onset.⁶ Other recent work has shown that transforming decoding weights into activation patterns using Haufe-transformed forward models provides a more interpretable representation of how neural activity contributes to decoding performance.⁷ However, it remains unclear how early animacy information can be reliably extracted from EEG data when using these image datasets.

Although previous studies have shown that animacy information can be decoded from EEG signals in early visual processing, not many studies have looked at whether these findings hold up across large naturalistic image datasets with thousands of object categories. This paper, therefore, examines whether the brain responds differently to animate versus inanimate stimuli within the first 200 milliseconds, and whether early EEG signals can reliably decode this information across a highly diverse set of stimuli. This question is important because it clarifies whether high-level semantic distinctions, such as animacy, emerge during early stages of visual processing or only after more extensive cortical processing.⁸ As a result, I used a combination of resolved decoding, permutation testing, ERP analyses, Haufe transformed forward models, and EEG-behavior correlations to examine both the timing and spatial distribution of related activity. From there, I hypothesized that animate objects would elicit stronger or earlier neural responses than inanimate objects during early stages of visual processing. In order to address this hypothesis, I analyzed EEG data from the Naturalistic Object Dataset (NOD-EEG). This is a large, publicly available dataset containing brain responses to thousands of object images.⁹ The images used in the NOD-

EEG are derived from the THINGS-EEG dataset, another large-scale stimulus set designed to sample naturalistic object concepts across a broad range.¹⁰ Although activity beyond 0 ms and 200 ms is examined for contextual purposes, the core analyses and conclusions of this paper focus specifically on the 0–200 ms early visual processing window.

■ Materials and Methods

EEG dataset and preprocessing:

EEG data were obtained from the Naturalistic Object Dataset EEG (NOD-EEG), which contains magnetoencephalography (MEG) and EEG recordings from human participants viewing thousands of naturalistic object images in rapid serial visual presentation streams. Specifically, the EEG recordings were collected using a 64-channel NeuroScan system using the international 10–20 electrode configuration, sampled at 500 Hz in an electromagnetically shielded room. This study used preprocessed and epoched EEG data from 15 participants (6 female, 9 male, mean age = 20.47 years, standard deviation = 1.28 years) with behavioral data. These 15 participants represented all subjects with usable EEG recordings and complete metadata available after preprocessing. The participant IDs were the following listed: sub-01, sub-02, sub-04, sub-07, sub-08, sub-10, sub-11, sub-12, sub-13, sub-14, sub-24, sub-26, sub-27, sub-29, and sub-30. Information about participants' handedness, height, and weight was not available in the dataset. Four participants (sub-03, sub-05, sub-06, and sub-09) were excluded because their EEG files were corrupted and could not be properly loaded during preprocessing. All the data used were originally identified and collected by the original authors, and I collected no new data for this project.

Each EEG .fif file contained epochs from -100 ms to 800 ms around image onset, along with metadata for each trial. This indicated whether the object was animate or inanimate, the participant's response, and their reaction time. The stimulus set consisted of a large, diverse collection of real-world objects, spanning thousands of unique images and categories, and each image was presented once per session. Participants subsequently performed a binary animacy judgment on each trial.

EEG data analysis:

Participants viewed rapid streams of animate (e.g., animals and people) and inanimate (e.g., tools and furniture) object images and had to judge whether each image was animate or inanimate on each trial. Behavioral responses and reaction times were recorded in the metadata. Additionally, behavioral accuracy was calculated as the proportion of correct responses, and reaction time was measured as the median response time across correct trials.

All analyses were done in Python using open-source libraries, including MNE, NumPy, SciPy, scikit-learn, and Matplotlib in Google Colab. ChatGPT was used to help with programming, mainly for debugging and figuring out code issues. All analytical decisions, statistical methods, interpretations, and conclusions were my own and independently verified by me. For each participant, I loaded the pre-epoched EEG data from

-100 ms to 800 ms relative to stimulus onset. The original files did not include baseline correction or active projections. Using the metadata field `stim_is_animate`, I then created a binary label for each trial (`animate = 1`, `inanimate = 0`). Trials missing behavioral data were excluded from behavior-related analyses but retained in the decoding analyses if they had a valid animacy label.

To focus on early visual processing, EEG data were cropped to the 0–200 ms window after the stimulus appeared and downsampled to 50 Hz to reduce computational demands and to help with signal stability, while preserving coarse temporal dynamics within the early visual processing window. Although this reduced temporal precision, the sampling rate was still sufficient to capture broad ERP timing differences within the early 0–200 ms window. A 3-fold cross-validation approach was used to try to balance computational efficiency with getting stable estimates, given the large number of trials each participant had. For each subject, this resulted in a data array with dimensions corresponding to trials, channels, and time. The number of trials varied slightly across participants (approximately 4,000 per subject) due to preprocessing and data-quality constraints. All available EEG channels in the decoding and forward-model analyses were used.

Time-Resolved Multivariate Decoding of Animacy:

To determine when animacy information could be decoded from EEG signals, I conducted a time-resolved multivariate decoding analysis using a ridge classifier and cross-validation. For each time point, I reshaped the EEG data into a trials-by-channels matrix and passed it through a pipeline that included a `StandardScaler` followed by a `RidgeClassifier`. Animacy decoding accuracy was estimated using a 3-fold cross-validation, which was performed separately for every time point in Python. This analysis produced a time-resolved decoding curve showing how classification accuracy changed over time for each participant. The highest decoding accuracy within the 0–200 ms window was then taken as that participant's peak decoding score. It was later used in the EEG-behavior analysis discussed in the results section.

Permutation-based group-level significance testing:

To assess whether decoding performance exceeded chance, I conducted subject-level permutation tests for each participant and then combined the resulting p-values across participants. For each participant, EEG data from the 0–200 ms window were reshaped into a trials by channels by time matrix and collapsed into a single feature vector. Decoding accuracy was calculated using the same `StandardScaler` and `RidgeClassifier` pipeline with 3-fold cross-validation, just like in the time-resolved analysis.

A null distribution was then generated by shuffling the animacy labels 1,000 times and recomputing the cross-validated decoding accuracy for each permutation. I assigned a p-value to each participant by calculating the proportion of permuted accuracies that were at least as high as the observed decoding accuracy. Across the 15 participants, real decoding accuracies generally ranged from 0.57 to 0.63, and all p-values

were below 0.05. Accordingly, for each individual, these values indicated statistical significance and above chance for animacy decoding. Group-level p-value distributions are reported in the Results section.

Forward models and group-average scalp maps

To understand where on the scalp animacy information was expressed, Haufe-transformed forward models (activation patterns) were estimated for each subject and then averaged at the group level. For each subject, a sensor-space pattern was derived to quantify how strongly each EEG channel contributed to animacy decoding within the early 0-200ms time window. Then, these subject-level pattern vectors were stacked into a subjects-by-channels matrix.

Because activation patterns derived from linear models can vary in sign across participants, each individual's pattern was sign-aligned to a reference pattern by checking their correlation and flipping the sign when the correlation was negative. After aligning the data, I averaged the patterns across all subjects to create a group-level scalp map, which I visualized using a topographic plot. The group map showed the strongest weights over occipital and parieto-occipital sensors, which matched early visual cortex involvement in animacy processing. To illustrate this, I computed Haufe-transformed, time-resolved maps for all the participants. I plotted them at evenly spaced time points (e.g., 40, 80, 120, 160, and 200 ms) to show how the spatial distribution of animacy-related activity changed over time. These examples are presented and discussed in the results section, which describes the temporal dynamics of these scalp patterns.

Event-related potentials (ERPs) and the occipital region of interest (ROI):

To complement the decoding analyses, I also examined event-related potentials (ERPs) to assess how EEG responses differed between animate and inanimate objects. ERPs were computed by averaging EEG responses across trials for each category, with the focus on an occipital region of interest (ROI) because it is closely associated with early visual processing.

Since the group forward model showed strong activity in posterior regions, I defined an ROI at the sensor level to get a clearer view of early visual responses. This group-average and difference-wave ROI included occipital and parieto-occipital electrodes (O1, O2, Oz, PO7, PO8) from the NOD-EEG dataset. The epochs were baseline-corrected using the -100 to 0 ms pre-stimulus interval and analyzed within the 0-200 ms window after stimulus onset. These ERP results, together with the forward-model scalp maps, support the idea that early animacy information mainly appears over occipital and parietal regions.

EEG-behavior relationships:

Finally, I examined whether early EEG decoding of animacy was related to behavioral performance. For each participant, I computed Behavioral accuracy (The mean proportion of correct animacy judgments, where the response correctly matched the true animacy label), Median reaction time (The median

reaction time (RT) across the correct trials), and Peak EEG decoding accuracy (the maximum decoding accuracy within 0-200 ms from the time-resolved decoding curve).

I then calculated Pearson correlations between peak decoding and behavioral accuracy, as well as between peak decoding and median RT across the subjects. In the results section, correlation coefficients and p-values, along with scatter plots showing that early animacy decoding relates to behavioral performance, are reported.

Results and Discussion

Early animacy decoding from EEG:

Across participants, animacy could be reliably decoded from early EEG signals. Using time-resolved decoding in the 0-200 ms window, the group average decoding accuracy rose above chance shortly after stimulus onset. It remained consistently above 0.5 for the rest of the early time window (Figure 1). The mean peak decoding accuracy across the 15 participants was 0.59 (Group peak = 0.59, ± 1 SEM = 0.02). Individual peak accuracies ranged from 0.57 to 0.63 (Figure 1). These results indicate that even a simple linear classifier can distinguish between animate and inanimate images based on early neural activity.

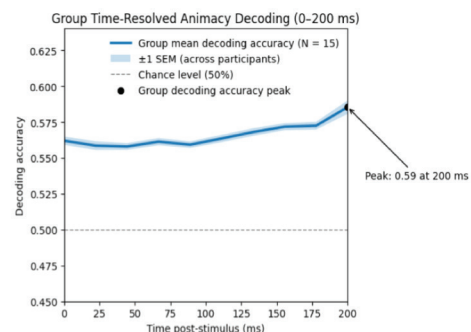


Figure 1: Group-average time time-resolved decoding accuracy for animate versus inanimate stimuli within the first 200 ms after stimulus onset. The solid line represents the mean decoding accuracy across participants, the shaded region indicates ± 1 SEM, the dashed line marks chance performance (50%), and the dot indicates the group's peak decoding accuracy. Decoding accuracy rose above chance beginning around 100 ms and reached its maximum near 200ms, demonstrating that animacy related information emerges rapidly during early visual processing.

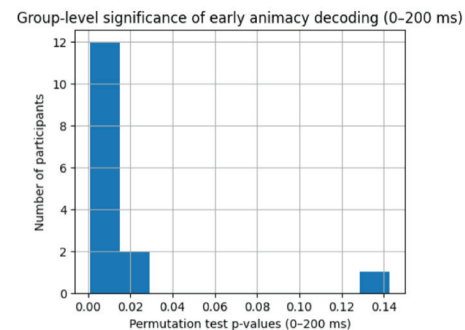


Figure 2: Histogram of participant participant-level permutation test p-values for animacy decoding accuracy within the 0-200 ms post post-stimulus window. Each bar shows how many participants reached a given significance level based on 1,000 label permutations. The strong clustering of values near zero indicates consistent above-chance decoding across participants. This indicates that animacy decoding was reliably above chance across the dataset.

Permutation tests confirmed that these decoding results were statistically significant for each participant. For each subject, decoding accuracy using the true animacy labels in the 0-200 ms post-stimulus period was compared to a null distribution of 1,000 permutation accuracies computed with shuffled labels (Figure 2). The null distributions were centered around chance (0.50), while all participants showed p-values below 0.05, with most under 0.01 (Figure 2). Together, these findings suggest that early animacy information was reliably detected at the individual and group levels.

Spatial distributions for forward models and occipital focus:

Haufe-transformed forward-model analyses showed that animacy-related EEG activity was strongest over posterior scalp regions. For each participant, I calculated a Haufe-transformed forward model for the 0-200 ms window to identify the channel-level patterns related to animacy decoding. These group-average forward models were then sign-aligned and averaged across participants. After this sign alignment was performed, the average spatial pattern remained highly consistent across the subjects. Overall, the strongest activity appeared over occipital and parieto-occipital electrodes, with weaker activity observed at central and frontal sites (Figures 3 and 4).

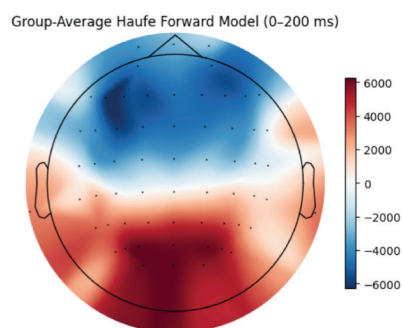


Figure 3: Group-average Haufe forward model for animacy decoding within the 0 0-200 ms postpost-stimulus window. Warmer colors indicate stronger positive contributions to decoding performance, whereas cooler colors indicate weaker contributions. The strongest activity was concentrated over occipital and parieto-occipital regions, suggesting a major role for early visual cortical processing in animacy decoding.

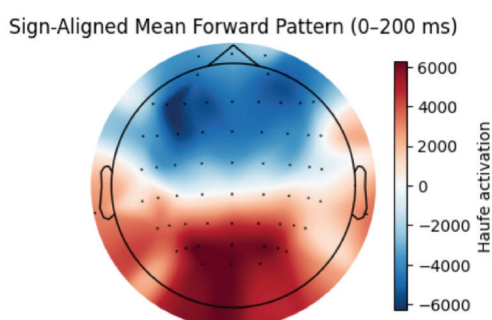


Figure 4: Sign-aligned group group-average Haufe forward model for animacy decoding within the 0 0-200 ms post post-stimulus window. Individual participant patterns were aligned before averaging to improve consistency across subjects. The resulting scalp map remained strongly posterior, demonstrating that animacy animacy-related activity was highly consistent across participants.

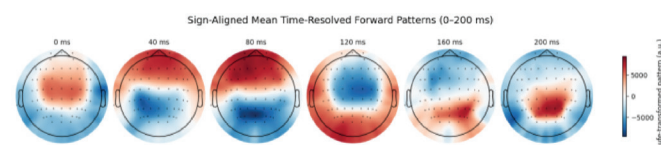


Figure 5: Time-resolved group group-average Haufe forward models shown at selected time points between 0 and 200 ms after stimulus onset. Topographic maps illustrate the evolving spatial distribution of animacy animacy-related neural activity over time. Posterior sensor patterns became increasingly prominent between approximately 120 and 200 ms, indicating that animacy animacy-related information emerged primarily within occipital and parieto-occipital regions during early visual processing.

The time-resolved forward maps supported this posterior focus. At 0, 40, and 80ms, the strongest positive weights appeared over frontal sites. However, when the time window reached 120, 160, and 200 ms, the weights shifted to posterior sites, peaking between 100 and 200 ms (Figure 5). This spatial pattern is consistent with early visual cortical activity. It suggests that animacy information is represented in distributed patterns over occipital regions around 100ms after stimulus onset.

Occipital ERP/ROIs for animate versus inanimate objects:

Event-related potentials over occipital and parieto-occipital electrodes showed clear differences between animate and inanimate objects. Using an occipital ROI (O1, O2, Oz, PO7, PO8), separate ERPs were calculated for each condition for animate and inanimate trials and then averaged across each participant. Both group-average waveforms exhibited typical early visual components, including a P1 around 100 ms and an N170 around 170 ms (Figure 6-7).

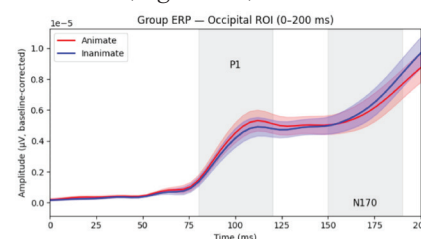


Figure 6: Group-average event event-related potentials (ERPs) from an occipital region of interest (O1, O2, Oz, PO7, PO8) for animate and inanimate stimuli within the 0 0-200 ms window. The colored regions represent ± 1 SEM across participants, and the shaded time intervals indicate the P1 and N170 components. Differences between animate and inanimate responses were most apparent between approximately 100 and 200 ms, suggesting early category category-sensitive visual processing.

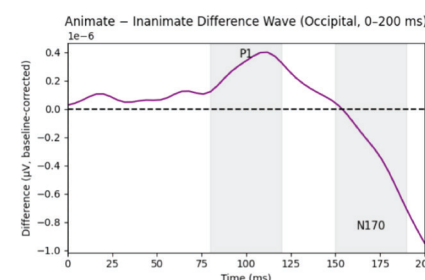


Figure 7: Group-average difference waveform (animate - inanimate) from the occipital region of interest within the 0 0-200 ms window. Shaded areas mark the P1 and N170 time ranges, showing early ERP differences related to animacy. The divergence from zero during the N170 interval suggests that animacy animacy-related neural differences emerge during early stages of visual processing.

The animate condition showed slightly lower amplitudes than the inanimate condition between 150 and 200 ms in the Occipital ROI, reflecting early category-sensitive processing (Figure 6). The group difference wave (animate minus inanimate) diverged from zero in this window (Figure 7), suggesting that related differences show up in early visual activity rather than only emerging during later cognitive stages. Although decoding was the focus in this study, these ERP differences support the view that animacy-related differences are reflected in early visual responses.

EEG-behavior relationships and correlations:

Participants overall performed well on the animacy task, with an average accuracy of approximately 0.84 (range = 0.77–0.91). This indicated that they could reliably distinguish animate from inanimate images. Median reaction times for correct trials were approximately 1040 ms, with values ranging from 0.69 to 1.30 seconds across all the participants (Figure 8–9).

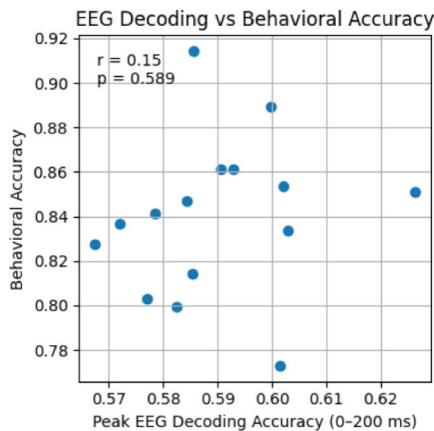


Figure 8: Relationship between peak EEG decoding accuracy (0–200 ms) and behavioral accuracy in the animacy classification task across participants. Each point represents one participant, with the correlation coefficient and p-value shown. Although participants with stronger decoding tended to show slightly higher behavioral accuracy, the relationship was not statistically significant.

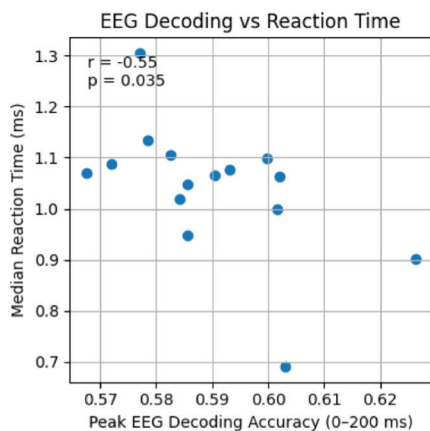


Figure 9: Relationship between peak EEG decoding accuracy (0–200 ms) and median reaction time for correct trials across participants. Each point represents one participant, with the correlation coefficient and p-value reported. Participants with stronger early animacy decoding generally responded more quickly, resulting in a significant negative correlation between decoding performance and reaction time.

To test whether early decoding was behaviorally relevant, I correlated peak decoding accuracy within 0–200 ms with each participant's behavioral accuracy and median reaction time. Peak decoding accuracies across the 15 participants ranged from around 0.57 to 0.63 (Standard deviation = 0.016). Peak EEG decoding accuracy showed a non-significant positive correlation with behavioral accuracy (Pearson $r = 0.15$, $p = 0.589$) (Figure 8). Additionally, the decoding accuracy showed a significant negative correlation with median reaction time (Pearson $r = -0.55$, $p = 0.035$) (Figure 9). These results suggest that participants who had stronger early decoding also tended to respond faster in the animacy task. However, because the sample size was quite small and several behavioral correlations were tested, this finding should be interpreted cautiously.

This study looked at whether animacy information can be decoded from EEG signals during the earliest stages of visual processing, and also tried to clarify when and where in the scalp this information shows up. I hypothesized that animate objects would elicit stronger or earlier neural responses than inanimate objects within the first 200 milliseconds after the stimulus onset. The results supported this hypothesis. Animacy-related information could be decoded from EEG activity beginning around 100 ms, with the best classification occurring between 100 and 200 ms. These findings replicate previous evidence showing that the human visual system can distinguish animacy quickly during early visual processing, while also extending this to a much larger and more naturalistic EEG dataset. The early timing of animacy decoding observed in this study is consistent with previous EEG and MEG research, which shows that object and category information can be extracted from neural signals shortly after stimulus onset.^{1,4} The observed timeframe aligns with known early visual ERP components, such as the P1 and N170, which have been linked to category-sensitive processing, including face processing.⁵ Although the study did not focus just on faces, the broader distinction between animate and inanimate objects appears to be reflected in similar early time windows, supporting the idea that animacy is a fundamental organizing factor of visual representations.

The spatial results from the Haufe-transformed forward-model analyses further support early visual involvement. The strongest activity appeared over occipital and parieto-occipital areas, consistent with early visual processing. These group-average forward models align with past findings that object information is represented across distributed occipital regions.^{2,3} In line with ERP results, these results indicate that animacy is not solely represented in high-level temporal regions, as shown by fMRI studies, but is already present in early visual responses measurable at the scalp.^{6,7}

One important implication of these findings is that animacy-related differences can still be detected even across highly diverse and naturalistic stimulus sets, which suggests that these early category signals are not just limited to more controlled lab image sets. Rather than relying entirely on later semantic processing, the results suggest that broad categorical differences may already be partly represented during early perceptual processing. Because animate objects are often behaviorally rel-

evant, being able to quickly tell them apart from inanimate stimuli could be advantageous during natural scene perception.

Interestingly, early decoding was also associated with behavior. Participants who showed stronger early decoding typically responded more quickly during the animacy task. This suggests that stronger early animacy decoding may be associated with faster visual detections. Similar connections between early neural decoding and behavior have been reported in prior work on object recognition and category processing, and the present findings extend this evidence to large-scale, naturalistic image datasets.⁸⁻¹⁰

Although this study has many strengths, there are several limitations to consider. First, while the NOD-EEG dataset includes a large and diverse set of images, this analysis focused on just 15 participants of similar age due to data quality issues. Because the sample size was quite small and mostly made up of young adults, the findings may not fully generalize to other populations or different age groups. Second, while decoding analyses reveal when information is present in the EEG signal, they do not identify its source within the brain. Lastly, the analysis was limited to the 0-200 ms window, so additional timeframes were not explored in detail. Future studies that combine EEG with MEG source localization could help identify the cortical sources of these early animacy signals, and extending the analyses to later time periods may also help reveal how early perceptual encoding connects to more complex processing.

■ Conclusion

This study shows that animacy is encoded very early in the human visual system and can be decoded from EEG signals within the first 200 milliseconds after image onset. Using large-scale naturalistic EEG data and time-resolved multivariate decoding, I found that animate and inanimate objects become distinguishable in neural activity within that early time window. Decoding accuracy peaked between 100 and 200 ms, and Haufe-transformed forward models showed consistent activity across posterior scalp regions, indicating early involvement of the visual cortex. ERP results also showed clear category-related differences in early brain responses, whereas behavioral results indicated that participants with stronger early decoding tended to respond more quickly during the tasks.

Altogether, these results support the conclusion that animacy is represented during the earliest stages of human cortical processing. By combining decoding, ERP, and behavioral analyses with large-scale naturalistic stimuli, this project helps connect controlled lab results with more realistic visual experiences. For future work, it would be interesting to examine finer distinctions within animacy, study differences across individuals or age groups, and combine EEG with other neuroimaging methods to gain deeper insight into the neural processes that support rapid perception.

■ Acknowledgments

I want to thank my mentor, Ioannis Delis, for his guidance throughout this project. I am especially grateful for his feed-

back on the paper's structure, the design of the analysis, and the interpretation of the results, all of which improved the study. I am also grateful to the creators of the Naturalistic Object Dataset EEG (NOD-EEG) for providing open access to their data, which enabled the study of large-scale patterns in visual object representations. This work would not have been possible without these open neuroscience resources and tools. All analyses were performed using open-source software, including Python, MNE, NumPy, SciPy, scikit-learn, and Matplotlib, and were run in Google Colab. ChatGPT was used to assist with writing and debugging code and to provide explanations. As the author, I attest that the ideas, graphics, and writing in this paper are entirely my own.

■ References

1. Grootswagers, T.; Wardle, S. G.; Carlson, T. A. Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time-Series Neuroimaging Data. *J. Cogn. Neurosci.* **2017**, *29* (4), 677–697. https://doi.org/10.1162/jocn_a_01068.
2. Jozwik, K. M.; Kriegeskorte, N.; Storrs, K. R.; Mur, M. Disentangling Five Dimensions of Animacy in Human Brain and Behavior. *Commun. Biol.* **2022**, *5*, 1247. <https://doi.org/10.1038/s42003-022-04194-y>.
3. Rogers, T. T.; Devereux, B. J.; Hodges, J. R.; Patterson, K. Evidence for a Deep, Distributed and Dynamic Code for Animacy in Human Ventral Anterior Temporal Cortex. *eLife* **2021**, *10*, e66276. <https://doi.org/10.7554/eLife.66276>.
4. Grootswagers, T.; Robinson, A. K.; Carlson, T. A. The Representational Dynamics of Visual Objects in Rapid Serial Visual Processing Streams. **2019**, *188*, 668–679. <https://doi.org/10.1016/j.neuroimage.2018.12.046>.
5. Eimer, M. The Face-Sensitivity of the N170 Component. *Front. Hum. Neurosci.* **2011**, *5*, 119. <https://doi.org/10.3389/fnhum.2011.00119>.
6. Wang, R.; Cichy, R. M.; Albada, S. J.; Oliva, A.; Kaiser, D. Mid-Level Feature Differences Support Early Animacy and Object Size Distinctions: Evidence from Electroencephalography Decoding. *J. Cogn. Neurosci.* **2022**, *34* (9), 1670–1680. https://doi.org/10.1162/jocn_a_01883.
7. Haufe, S.; Meinecke, F.; Görgen, K.; Dähne, S.; Haynes, J.-D.; Blankertz, B.; Bießmann, F. On the Interpretation of Weight Vectors of Linear Models in Multivariate Neuroimaging. *NeuroImage* **2014**, *87*, 96–110. <https://doi.org/10.1016/j.neuroimage.2013.10.067>.
8. Gifford, A. T.; Dwivedi, K.; Roig, G.; Cichy, R. M. A Large and Rich EEG Dataset for Modeling Human Visual Object Recognition. *NeuroImage* **2022**, *264*, 119754. <https://doi.org/10.1016/j.neuroimage.2022.119754>.
9. Zhang, G.; Li, S.; Li, Y.; Wang, Y.; Liu, H.; Wang, L.; Song, Y.; Liu, Z. A Large-Scale MEG and EEG Dataset for Object Recognition in Naturalistic Scenes. *Sci. Data* **2025**, *12*, 857. <https://doi.org/10.1038/s41597-025-05174-7>.
10. Grootswagers, T.; Zhou, I.; Robinson, A.; Hebart, M.; Carlson, T. Human EEG Recordings for 1,854 Concepts Presented in Rapid Serial Visual Presentation Streams. *Sci. Data* **2022**, *9*, 3. <https://doi.org/10.1038/s41597-021-01102-7>.

■ Author

Stewart S. Lee is a recent graduate of Stratford High School in Houston, Texas. He will attend Harvard College as a mem-

ber of the class of 2030 and plans to study neuroscience. His academic interests include visual perception, EEG decoding, and large-scale neural datasets.

U.S. Federal Reserve Monetary Policy: Systematic or Discretionary Based on Real-Time Evidence

Zachary Wong

Walter Payton College Prep, 1034 N Wells St, Chicago, Illinois, 60610, USA; zachary.e.wong@gmail.com

ABSTRACT: In this paper, we evaluate the extent to which the U.S. Federal Reserve interest rate policy actions represent systematic responses to macroeconomic conditions versus discretionary or inertial responses when analysis is limited to real-time information. We utilize vintage-aligned datasets to constrain models to information available to the Federal Open Market Committee at the time of each decision. Rather than treat policy action as one continuous decision function, we decomposed each decision into two distinct components: direction, representing a hike or cut; and magnitude, representing the extent to which the Federal Reserve is moving in that direction. From that split, we use a variety of methods (machine learning, Taylor Rule, etc.) to model each sub-decision, applying expanding-window backtesting to maintain real-time fidelity. We found that while the direction component is highly systematic, with simple rules performing accurate predictions on par or better than machine learning, the magnitude was less rule-like, implying flexibility. Based on counterfactual simulations from a stylized New Keynesian model, these flexibilities result in small deviations during stable periods, but are meaningful during periods of stress. We conclude that monetary policy is largely rule-based in direction, but discretion plays a more important role in magnitude, especially during crisis periods.

KEYWORDS: Economics, Macroeconomics, Monetary Policy, Real-time Data, Machine Learning.

■ Introduction

The Federal Reserve's policy decisions have a significant impact on metrics such as inflation, employment, and financial stability. To that extent, it has been researched thoroughly, with some claiming that financial policy is often a discretionary judgment under uncertainty and based on a wide host of factors, while others point to it following more systematic patterns or rules (e.g., Taylor rule, a simple formula relating interest rates and factors such as output gap and inflation).¹

However, most studies on systematic monetary policy use revised ex post data that policymakers of the past could not have known with such accuracy. Data revisions often occur months or years after they were relevant to policymakers, directly leading to hindsight bias.^{3,4} This led us to hypothesize that apparent policy "mistakes" may disappear under real-time data constraints. Using real-time data allows for clean inference about decision structure, and predictability under such constraints is evidence of systematic behavior.

Machine learning has been shown to have success in macro forecasting and high-dimensional relationships, which are potentially useful in predicting complex economic relationships and policy analysis for decisions like the Federal Reserve's interest rates.^{5,6} However, as monetary policy is low-frequency in our sample, often without changes for years, the data available for model training is extremely limited. Additionally, as we later show, factors like institutional inertia play a large role in the decisions they make, making machine learning forecast accuracy alone insufficient for policy interpretation.

Rather than investigate whether algorithmic policy can "beat" the Federal Reserve in decision-making, we ask what

predictability or lack thereof says about the observed policy structure of the US Reserve system. While economic research often treats policy as a single continuous determination, that choice may overly simplify distinct sources of systematic behavior and discretion. Thus, we break each decision down into direction and magnitude components and evaluate each independently. While we acknowledge the small sample size inherent to institutional monetary policies such as the FOMC decision structure and the resulting limitations of the study, they are not faults of methodology but unavoidable in the analysis of central bank decision-making. We use these changes to compare machine learning models to structural benchmarks, as well as review the consistency of policy decisions based on economic regimes and build illustrative counterfactual simulations using stylized New Keynesian models. Our models and methodology were designed to allow us to distinguish true discretionary behaviors from noise generated by factors such as data revisions and institutional inertia. Based on these analyses, we assess the macroeconomic implications of systematic versus discretionary policy paths.

In this manner, we focus on three main contributions: First, we eliminate the hindsight bias common in policy analysis by using strict real-time information. Second, we break down interest rate decisions into direction and magnitude, allowing us to highlight behavioral differences between the two. Lastly, we show that discretionary deviations from systematic rules are often limited but become meaningful during crises.

■ Data and Real-Time Information Structure

Real-Time Macroeconomic Data:

The primary macroeconomic predictors we used to predict the Federal Reserve's interest rates were taken from the Federal Reserve Bank of Philadelphia's Real-Time Data Set for Macroeconomists (RTDSM).^{4,16} These predictors include inflation, core inflation, consumption growth, labor market utilization, and output indicators. This dataset was necessary to maintain the scope of information the models received, limited to the latest data release available before each FOMC decision. Additionally, we accounted for consumer sentiment and expectations using the University of Michigan inflation expectations and consumer sentiment index.²⁰ All of this data was forward-filled to decision dates, and through these methods, we prevent models from using revised ex post information, which is essential to prevent models from exploiting information unavailable to policymakers of the time.³

Policy Decisions:

The specific metric of policy we attempt to predict is the midpoint of the Federal funds rate target ranges, obtained from FRED and FOMC documentation.^{17,19} We record every identified date that the Federal Reserve changes this midpoint, labeling that a decision. In this manner, we constructed two datasets. The primary set we analyze is a decision-event dataset that only includes meetings with these rate changes, focusing only on active decisions. The secondary dataset is a meeting-level dataset that includes all of the holds in the dataset and thus enables 3-class classification (hike, hold, cut). Because we limit the decision-event dataset to only hikes and cuts, we can avoid inflated accuracy from predicting "no change," as that is the vast majority of the dataset. For that reason, this paper is based on the primary dataset entirely, with the secondary only existing for robustness.

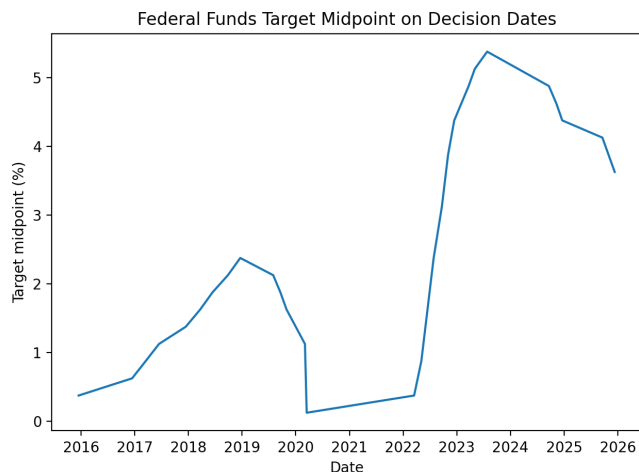


Figure 1: A visualization of the changes in the Federal Funds Rate over time.

Figure 1 illustrates that changes in the Federal Funds Rate tend to happen in clear cycles, not isolated and random movements. Periods of repeated increases or decreases suggest that the policy decisions move consistently in the same direction during a tightening or easing cycle, pointing to the idea of policy inertia.

■ Methods

Decomposition of Policy Decisions:

Policy decisions are decomposed into two distinct components: direction, a measure of whether the policy decision is a rate hike or cut, and magnitude, measuring the extent of rate change in percentage points. This reflects one of our core assumptions that the decision to move rates and the choice of how much to move them may be governed by different core logics and levels of discretion. This aligns with the documented behavior of policymakers sequentially discussing whether to adjust policy before deciding how much, supported by historical accounts of FOMC deliberations.¹⁵

Models:

We utilize two machine learning models to predict the direction of policy: a logistic regression model and a random forest classifier. We also utilize two machine learning models to predict the magnitude of policy: a ridge regression model and a random forest regression model. We utilize a Taylor-type rule to model economic decisions and interest rates based on a rigid, systematic formula.^{1,2} Expanding-window backtests were used to train and test models in the face of our limited dataset, as it mimics real-time learning while avoiding look-ahead bias.

Baselines:

We use two simple baseline comparators to add context to performance: an always hike model and an always cut model. Both of these perform the same policy action for every decision. The primary baseline used, however, is the "Last Decision" baseline, which predicts that each decision will be the same as the previous action (e.g., if the last decision was hike, predict hike), serving as a momentum baseline that tests the predictive power of persistence in our sample.

Evaluation Metrics:

To evaluate the effectiveness of each model, two metrics are used. For direction, the first metric is the standard accuracy, measuring how often each decision is predicted correctly. The second metric is balanced accuracy, where cumulative hike and cut predictions are weighed equally. The first metric for magnitude is Mean Absolute Error (MAE), or the sum of the absolute value of the prediction's residuals. Second is the Root Mean Squared Error (RMSE), which sums the squares of the prediction residuals.

■ Results

Directional Predictability:

We concluded that direction is highly predictable and systematic in the Federal Reserve's decision-making. The strongest predictive models we tested for directional accuracy were logistic regression and the Taylor-rule benchmark. As seen in Table 1, the Taylor Rule benchmark met or exceeded the machine learning models used. Additionally, it can be seen that the momentum baseline performs very well, which adheres to the concept of institutional inertia by indicating that policy momentum is central to policy direction in our sample. As the success shown would be unlikely if directional poli-

cy decisions were primarily discretionary, we have interpreted these results to mean that direction is governed by stable rules and inertia, with extra flexibility from machine learning models adding little further insight.

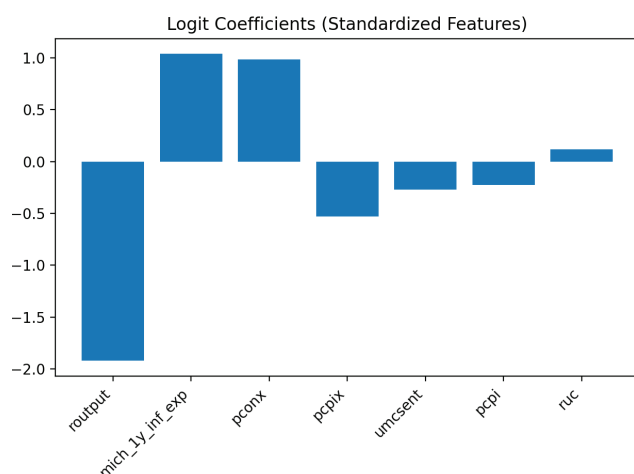


Figure 2: The relative importance of feature variables to the logistic regression model, with the output gap being the most important.

Figure 2 shows that the logistic regression model places the greatest importance on the output gap as a factor for predicting the direction of policy changes. This suggests that real economic activity plays a central role in the Federal Reserve's decision of whether to raise or lower interest rates.

Table 1: Directional prediction accuracy on decision events. The Taylor rule benchmark achieves the highest accuracy, indicating strong systematic behavior in policy direction.

Model	Accuracy	Balanced Accuracy	Notes
Always Hike	0.579	0.500	Naïve baseline
Always Cut	0.421	0.500	Naïve baseline
Last Decision	0.895	0.892	Momentum baseline
Logistic Regression	0.842	0.830	ML, interpretable
Random Forest	0.737	0.739	Nonlinear ML
Taylor Rule (Benchmark)	0.947	0.955	Structural benchmark

The strong direction-predicting performance of both the Taylor Rule Benchmark and the Last Decision Baseline suggests that the direction of the Federal Reserve's rate changes is not an impromptu judgment, but reflects a combination of macroeconomic rule-following and institutional policy momentum. This supports a view that even when the Federal Reserve exercises its power, the basic choice of hiking or cutting is highly constrained by clear and systematic factors like observable economic conditions and the current policy regime.

Magnitude Predictability:

In contrast, we found that magnitude is harder to predict with rule-based or systematic approaches than direction. As seen in Table 2, Ridge regression performs the best overall,

with superior accuracy to the more rigid and methodical Taylor Rule. As the Taylor rule's predictions are too smooth for the erraticism of decisions in magnitude, it underpredicts large moves even when accurate in direction. This underprediction is consistent with the selective discretion we identified in the magnitude of policy adjustments being episodic and concentrated during periods of uncertainty and stress. Additionally, the random forest regression model demonstrates limited additional improvement.

This loss in predictability seems to suggest a level of flexibility in how the Federal Reserve decides the extent of an adjustment, especially during periods of stress such as the COVID-19 pandemic, resulting in emergency easing followed by a post-pandemic inflation tightening cycle. We interpret this divergence as evidence that discretion matters less for whether the Federal Reserve hikes or cuts rates, and more for how aggressively it moves in that direction. Overall, we conclude that while directional decisions are likely largely systematic, magnitude decisions may be influenced more by selective discretion.

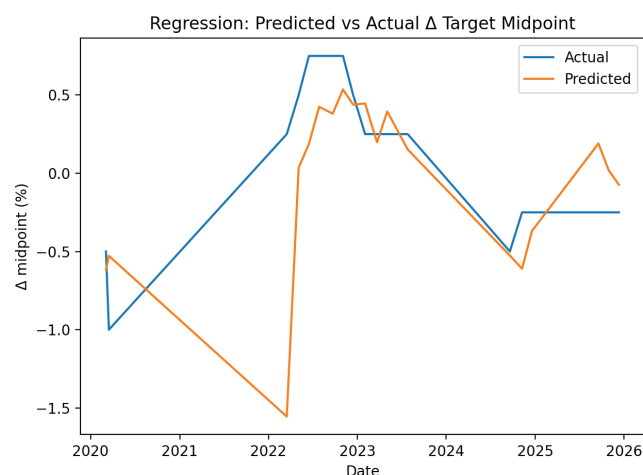


Figure 3: The discrepancy between regression-predicted rates and actual target rates.

Figure 3 shows that the predicted rate changes follow the general movements of actual policy, but fail to predict the larger deviations. This suggests that while the model can capture the broad direction of policy quite well, it fails to predict the exact size of rate adjustments, particularly during highly unusual or stressful periods where the most drastic changes occur.

Table 2: Magnitude prediction performance (Δ policy rate). Ridge Regression and Random Forest Regression outperform the Taylor Rule benchmark, suggesting less rule-like behavior.

Model	MAE	RMSE
Taylor Rule (Benchmark)	0.41	0.79
Ridge Regression	0.33	0.50
Random Forest (Regression)	0.37	0.47

The weaker magnitude-predicting performance of the Taylor Rule benchmark suggests that while systematic rules seem to capture the direction of Federal Reserve decisions well, they fail to accurately capture how aggressively it moves. This

distinction is important because policy magnitude is where central bankers may respond to financial stress, uncertainty, market expectations, or other crisis-specific risks that are complex and difficult to encode in a simple policy rule.

Taylor-Rule Benchmark:

When evaluated on our primary decision-event date dataset, the Taylor-rule benchmark performed with exceptionally high directional accuracy but performed relatively poorly in accurately predicting policy magnitude. As seen in Table 3, the Taylor rule's estimated coefficients indicate that policy responses are significantly impacted by output gap and core inflation, while the unemployment rate had minimal influence on predicted decisions. While not true structural policy reaction parameters, these coefficients give descriptive insights into deeper systematic decision-making within the realm of interest rate decision-making in the Federal Reserve.

Table 3: Estimated Taylor rule coefficients. Output gap and core inflation are weighted significantly compared to the unemployment rate.

Variable	Coefficient
Output gap	-0.96
Core inflation	0.86
Unemployment rate	-0.01

The strength of the output gap and core inflation coefficients reinforces evidence that the Taylor Rule benchmark is largely driven by these two major macroeconomic variables. The degree to which the coefficients of these variables are greater than the unemployment rate supports the idea that the Federal Reserve seems to act in a systematic way in response to transparent, broad macroeconomic fundamentals.

Regime and Cycle Analysis:

To further the notion of institutional inertia within federal interest rate policymaking, we split samples into time segments aligning with known policy cycles, such as the post-zero-lower-bound liftoff, as seen in Table 4. Within these regimes, policy behavior is nearly deterministic, with consistent direction. Errors only occur during transitory periods between regimes, which are inherently outside of the norm, suggesting that discretion is episodic in its impact. This uniformity suggests that discretion is episodic and used mainly during turning points. This also lends credibility to our momentum-based model success, as outside of outlying turning points, direction is almost entirely conforming to existing regime momentum. Overall, we find that regime persistence and institutional inertia play a central role in the Federal Reserve's interest rate decision-making.

Table 4: Model performance across monetary policy regimes. Remarkable direction consistency is held within regimes.

Regime	Period	N	% Hikes	Logit Accuracy	Baseline (Last Decision)
ZLB liftoff & hikes	2015–2018	9	100%	0.78	1.00
2019 easing	2019	3	0%	1.00	1.00
COVID emergency	2020	2	0%	1.00	1.00
Inflation hikes	2022–2023	11	100%	1.00	1.00
2024–2025 easing	2024–2025	6	0%	1.00	1.00

The regime results suggest that a great deal of the apparent predictability in monetary policy comes from persistence within policy cycles. Inflationary pressure, labor-market conditions, and output trends may lead the Fed into tightening and easing cycles, which continue into subsequent decisions in the same direction until another major regime shift occurs. This suggests that discretion may be most important not for ordinary decisions within a regime, but during crucial moments when the Fed decides to shift regimes.

Interpretability:

As seen from Table 3 and Table 5, the logistic and Taylor coefficients are extremely visible and interpretable, as the importance of each variable is evident in its coefficient. Our findings align with economic intuition of the influence of inflation and activity measures. The feature importance comparators of the random forest model also confirm these findings, as seen in Figure 4. By these models' success and interpretability, we conclude that policy is not truly opaque, and economic behavior is interpretable.

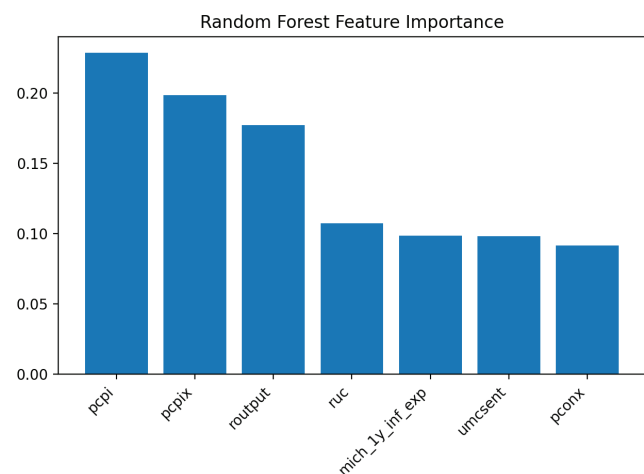


Figure 4: The relative importance of the features of the Random Forest model highlights the strength of the Consumer Price Index.

Figure 4 shows that CPI inflation, core CPI inflation, and the output gap are some of the most important prediction factors in the Random Forest model. This points to the idea that inflation and economic activity are key to the systematic portion of the Federal Reserve's decision-making.

Table 5: Logistic regression coefficients for policy direction. The output gap has the strongest predictive power in the logistic regression model.

Predictor	Coefficient	Coefficient Magnitude
Output Gap	-1.92	1.92
1-year inflation expectations	1.04	1.04
Core consumption inflation	0.98	0.98
Core CPI inflation	-0.53	0.53
Consumer Sentiment	-0.27	0.27
CPI inflation	-0.23	0.23
Unemployment rate	0.12	0.12

The output gap has the largest coefficient by magnitude, followed by inflation expectations and core consumption inflation. This table describes which variables were most important for predicting policy direction for our logistic regression model, and suggests that a mix of real economic activity and inflation-based measures plays a key role in the systematic component of Federal Reserve rate decisions.

Counterfactual Analysis:

The framework of our counterfactual evaluation to test the macroeconomic effects of systematic and discretionary policy is a stylized New Keynesian model.^{7,8} This model allows for a descriptive comparison between the actual policy path and a path implied by the Taylor-rule benchmark. While purely illustrative, this analysis aligns with our previous findings, suggesting that while the two paths only have small differences during stable periods, there are significant deviations during periods of stress. With this model as the basis for comparison, we find that discretion matters when concentrated, but over long periods of time, policy decisions of similar directional foundations through stable regimes result in largely similar outcomes.

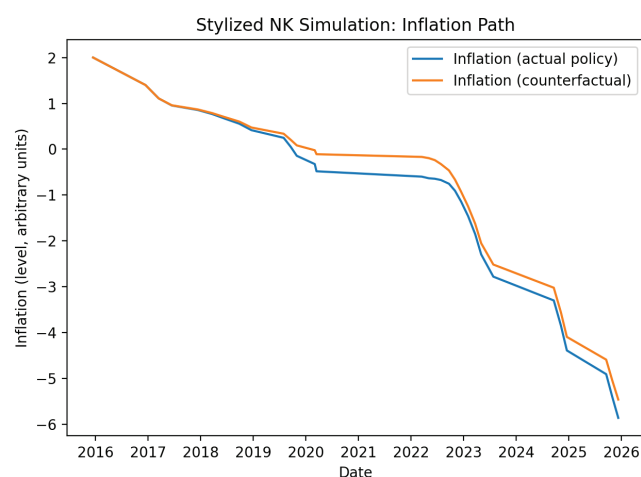


Figure 5: The predicted differences between actual policy and modeled policy over time demonstrate clear coherence during stable periods but larger differences during larger changes.

Figure 5 shows that actual and modeled policy paths derived from the stylized New Keynesian simulation are relatively similar during stable periods, but larger differences occur during periods of stress, causing greater divergence. This supports the conclusion that while systematic rules can explain much

of normal policy behavior, discretion becomes more important during crises and major regime shifts.

■ Discussion

These results reveal a nuanced view of Federal Reserve monetary policy. The strong performance of the Taylor Rule benchmark supports classic rules-based views of monetary policy associated with Taylor and later with Clarida, Galí, and Gertler, arguing that central bank behavior can be described through systematic responses to inflation and economic activity. As this paper utilizes real-time data, it extends Orphanides' critique of ex post economic analysis by claiming that real-time uncertainty matters,³ as well as prior research emphasis on the systematic nature of U.S. monetary policy.^{1,2} The relevant question is not whether policy appears systematic after future data revisions supply information unavailable to policymakers, but whether it remains predictable while restrained to data available at the time. Under these constraints, machine learning matches but does not beat rules in directional prediction modeling, highlighting the predictive strength of simple and interpretable rules, but the magnitude of rate changes is not nearly as well captured. This paper supports the view that directional monetary policy is well captured by rules, but that these rules may fail to fully capture the magnitude of policy adjustments.

A possible explanation for these findings is that the Federal Reserve's decisions are neither entirely mechanical nor discretionary. Simple policy rules may perform well in predicting the direction of decisions because they are highly related to the broad reaction function of the Federal Reserve. When inflationary pressure rises, or economic slack falls, the Fed will naturally tend to tighten, and in opposite conditions, it will tend to ease. From the success of the Last-Decision Baseline, it can be inferred that subsequent decisions tend to continue in the same direction until a major regime shift occurs. Machine learning offering limited improvement may suggest that much of the Fed's directional behavior is already largely captured by transparent macroeconomic relationships and institutional continuity.

On the other hand, simple rules' relatively weak ability to capture the magnitude of changes serves as possible evidence pointing to discretion playing a much larger role in the size and timing of adjustments as compared to the basic direction of policy. This reveals that central banks may behave simply at a high level while relying more on judgment when calibrating precise policy actions. This is sensible, as during crises, the Federal Reserve must judge whether or not to implement a more significant move or not for reasons ranging from the stabilization of markets, rapidly changing expectations, or even preservation of credibility. This offers a potential explanation for why the Taylor-rule benchmark correctly predicts the direction of policy, but underpredicts large moves.

These findings contribute to the rules-versus-discretion debate by suggesting that the dichotomy between rules and discretion may be less absolute than is often assumed. While the Federal Reserve appears to follow a rule-like pattern in general directional decisions, it may still preserve discretion

to determine the degree to which it should respond to economic conditions. In that sense, policy rules are shown to be useful not because they are a replacement for central bankers, but because they offer insights into understanding and evaluating policy behavior. This paper thus supports a middle ground view: rules dealing with broad macroeconomic indicators hold strong explanatory power over the systematic portion of monetary policy, but discretion in the size of an adjustment remains an important factor for uncertainty, institutional judgment, and conditions not covered by standard macroeconomic variables.

Future research could be put into comparison of the Federal Reserve's decision frameworks with those of other central banks, as well as more in-depth simulations of an automated central bank's effects on societal metrics.

■ Limitations

This study has several major limitations. Firstly, the dataset is extremely limited because FOMC decisions are sparse, and many meetings do not result in a change in target rate, which further limits the usable data points that prediction models can train and test on. This significant restriction of the amount of data for the models used limits the strength of statistical claims for accuracy. Secondly, the analysis only focuses on the selected set of macroeconomic indicators available to the public, but does not include qualitative information available to policymakers, such as market functioning, geopolitical risks, financial stability concerns, and internal deliberations. The lack of all potential factors in consideration makes it difficult to ascertain whether or not decisions are truly systematic or discretionary. Thirdly, the New Keynesian counterfactual simulations are not structural estimates of the full economy, but are merely illustrative. All in all, these major limitations mean that results should be treated as evidence about policy predictability and decision structure, and not a definitive model of how the Federal Reserve makes decisions.

■ Conclusion

This paper finds that the Federal Reserve's interest rate decisions are composed largely of two sub-decisions: the direction of rate change, which we thus refer to as "direction", and the extent of rate change, which we call "magnitude". To that end, based on our real-time data constraints, direction is largely rule-based while magnitude is comparatively discretionary and flexible, especially during periods of crisis or other turning points in economic regimes. This suggests that market expectations may be anchored in systematic directionality even during uncertainty around policy intensity. Machine learning models tested in this paper were able to match but not surpass structural benchmarks once real-time information constraints and regime persistence were imposed. This finding contributes to the macroeconomic policy rules-versus-discretion debate by adding the nuance of the components of policy action and their respective adherence to systematic decision-making, rather than overturning existing understandings. Because the sample size of active FOMC decisions is small and excludes potential qualitative internal deliberations, these findings should be in-

terpreted as evidence of predictability and structure in Federal Reserve decision-making rather than a complete model.

These results suggest future avenues of research, including similarities between the Federal Reserve and other central banks' decision-making and differences in the effects of changes in direction and magnitude on financial markets. Further exploration into different models, such as hybrid models incorporating rule-based and flexible mechanics, is also potentially interesting. Research into how Federal Reserve decision-making has changed, as well as comparisons between Federal Reserve behaviors towards different financial tools, could provide further nuance into the discussion surrounding its determinations.

■ Acknowledgments

The author thanks Dr. Sakk for guidance throughout the project and the Federal Reserve Bank of Philadelphia for providing access to real-time macroeconomic data.

■ References

1. Taylor, J. B. Discretion versus policy rules in practice. *Carnegie-Rochester Conf. Ser. Public Policy* 1993, 39, 195–214.
2. Clarida, R.; Gali, J.; Gertler, M. Monetary policy rules and macroeconomic stability: Evidence and some theory. *Q. J. Econ.* 2000, 115 (1), 147–180.
3. Orphanides, A. Monetary policy rules based on real-time data. *Am. Econ. Rev.* 2001, 91 (4), 964–985.
4. Croushore, D.; Stark, T. A real-time data set for macroeconomists. *J. Macroecon.* 2001, 23 (2), 253–264.
5. Coulombe, P. G.; Leroux, M.; Stevanovic, D.; Surprenant, S. How is machine learning useful for macroeconomic forecasting? *J. Appl. Econom.* 2020, 35 (7), 920–964.
6. Korobilis, D. *Machine Learning Macroeconometrics: A Primer*; Essex Finance Center Working Paper No. 22666; University of Essex: Colchester, U.K., 2018. <https://ssrn.com/abstract=3246473>. (accessed 2026)
7. Christiano, L. J.; Eichenbaum, M.; Evans, C. Nominal rigidities and the dynamic effects of a shock to monetary policy. *J. Polit. Econ.* 2005, 113 (1), 1–45.
8. Smets, F.; Wouters, R. Shocks and frictions in U.S. business cycles: A Bayesian DSGE approach. *Am. Econ. Rev.* 2007, 97 (3), 586–606.
9. Bernanke, B. S. *Monetary Policy and the Housing Bubble*; Speech at the American Economic Association Meeting, Atlanta, GA, January 3, 2010. <https://www.federalreserve.gov/newsevents/speech/files/bernanke20100103a.pdf> (accessed 2026).
10. Blinder, A. S.; Reis, R. Understanding the Greenspan standard. *Brookings Pap. Econ. Act.* 2005, 2005 (2), 11–96.
11. Athey, S. *The Impact of Machine Learning on Economics*; NBER Working Paper No. 24449; National Bureau of Economic Research: Cambridge, MA, 2018. <https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/impact-machine-learning-economics> (accessed 2026).
12. Orphanides, A.; van Norden, S. The unreliability of output-gap estimates in real time. *Rev. Econ. Stat.* 2002, 84 (4), 569–583.
13. Taylor, J. B. *Housing and Monetary Policy*; NBER Working Paper No. 13682; National Bureau of Economic Research: Cambridge, MA, 2007.
14. Wieland, V.; Wolters, M. The diversity of forecasts from macroeconomic models. *Econ. Policy* 2011, 26 (65), 1–54.

15. Federal Open Market Committee. *Minutes of the Federal Open Market Committee, December 15–16, 2015*; Board of Governors of the Federal Reserve System: Washington, DC, 2016. <https://www.federalreserve.gov/monetarypolicy/files/fomcminutes20151216.pdf> (accessed 2026).
16. Federal Reserve Bank of Philadelphia. *Real-Time Data Set for Macroeconomists*. <https://www.philadelphiafed.org/surveys-and-data/real-time-data-research> (accessed 2026).
17. Federal Reserve Bank of St. Louis. *Federal Reserve Economic Data (FRED)*. <https://fred.stlouisfed.org/> (accessed 2026).
18. Federal Reserve Bank of St. Louis. *ALFRED: Archival Federal Reserve Economic Data*. <https://alfred.stlouisfed.org/> (accessed 2026).
19. Federal Reserve Board. *FOMC Statements, Minutes, and Transcripts*. <https://www.federalreserve.gov/monetarypolicy/fomc-calendars.htm> (accessed 2026).
20. University of Michigan. *Survey of Consumers*. <https://data.sca.isr.umich.edu/> (accessed 2026).

■ Author

Zachary Wong is a student at Walter Payton College Prep in Chicago, Illinois. His research interests include macroeconomic policy and machine learning applications in economics. He intends to pursue economics and computer science at the university level.

■ Appendix

Random Forest Feature Importance. The Consumer Price Index holds the strongest importance in the Random Forest model.

Predictor	Importance
CPI inflation	0.229
Core CPI inflation	0.198
Output gap	0.177
Unemployment rate	0.108
Inflation expectations	0.099
Consumer sentiment	0.098
Core consumption inflation	0.092



Marine Biology Research at Bahamas

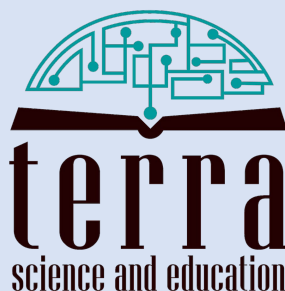
Unique and exclusive partnership with the Gerace Research Center (GRC) in San Salvador, Bahamas to offer marine biology research opportunities for high school teachers and students.

- Terra has exclusive rights to offer the program to high school teachers and students around world.
- All trips entail extensive snorkeling in Bahamian reefs as well as other scientific and cultural activities.
- Terra will schedule the program with GRC and book the flights from US to the GRC site.
- Fees include travel within the US to Island, lodging, meals, and hotels for transfers, and courses.
- For more information, please visit terraed.org/bahamas.html

Terra is a N.Y. based 501.c.3 non-profit organization dedicated for improving K-16 education

IJHSR International Journal of High School Research

is a publication of



N.Y. based 501.c.3 non-profit organization
dedicated for improving K-16 education