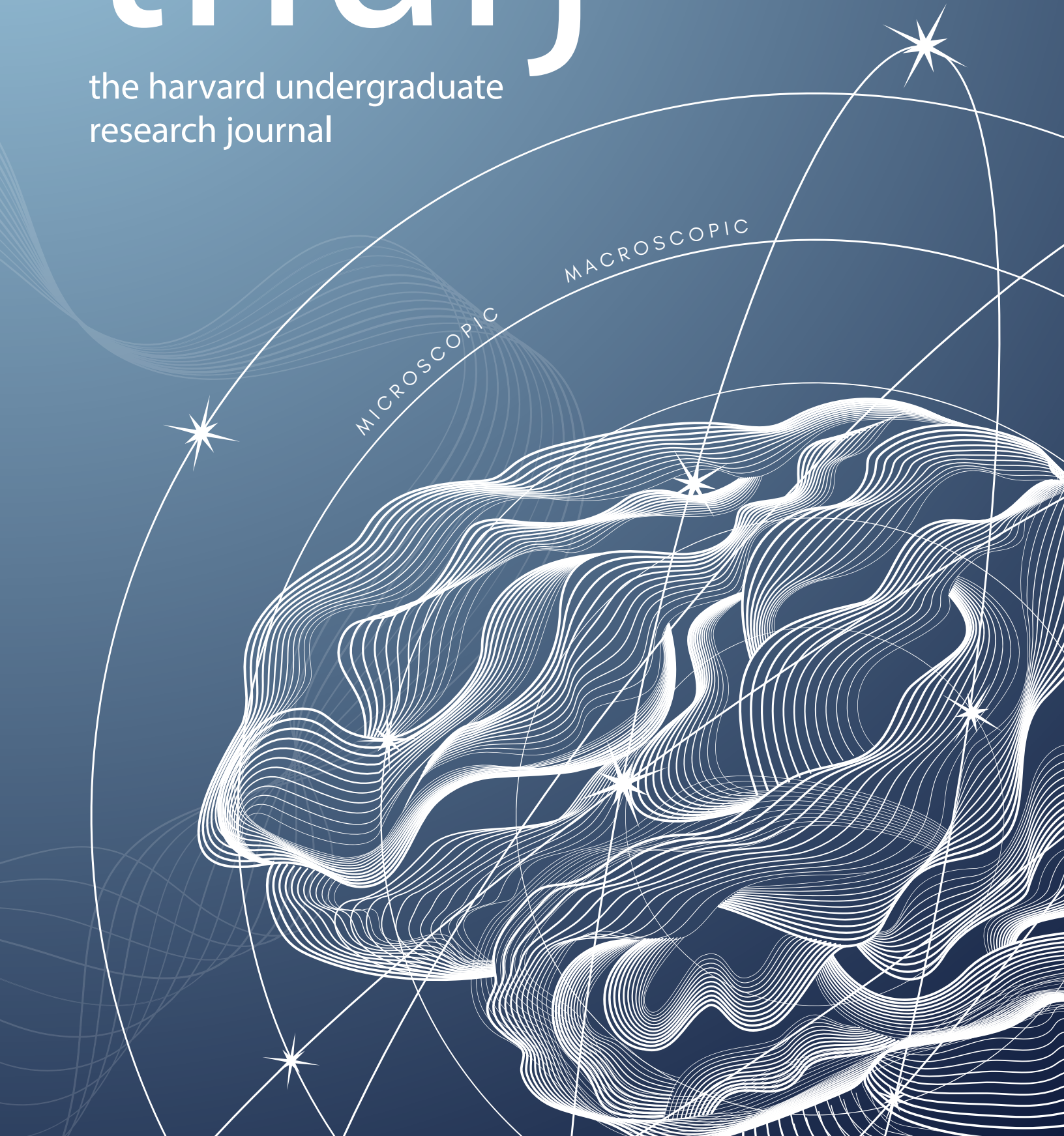
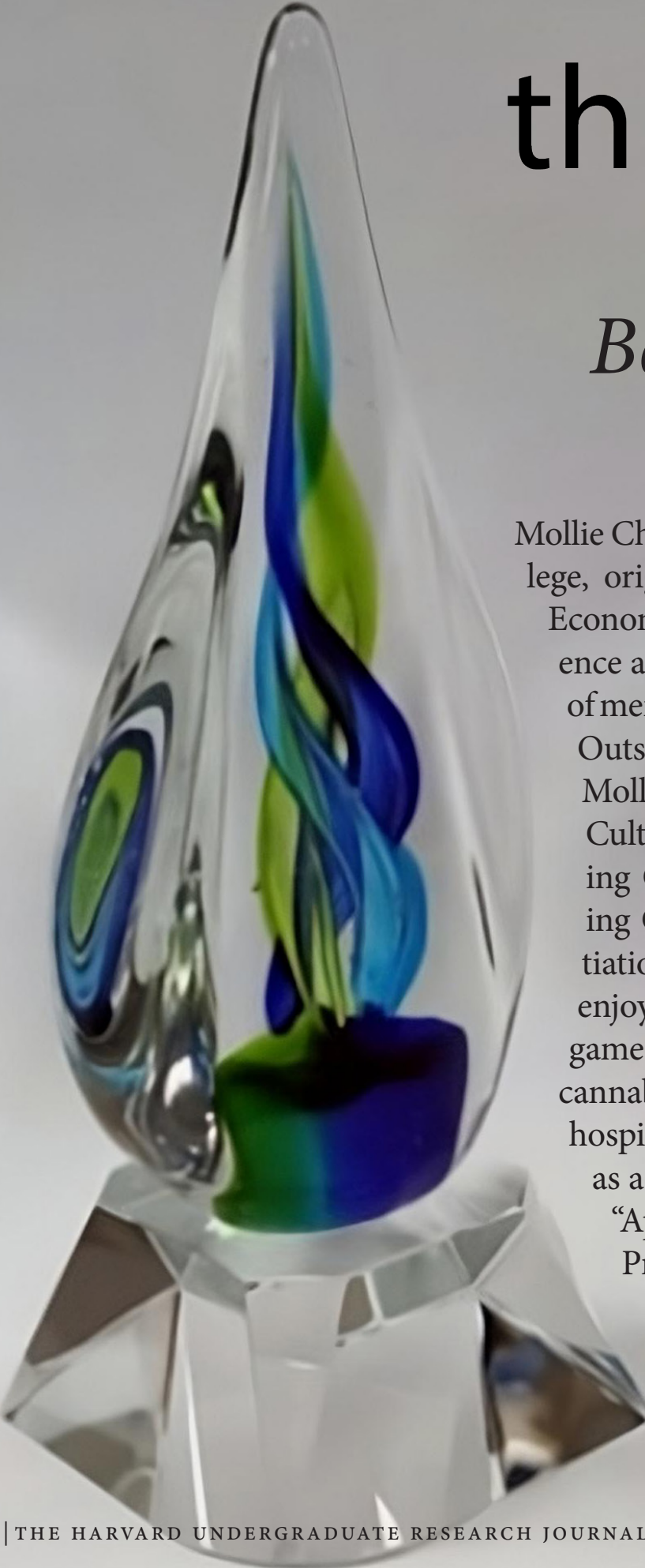


thurj

the harvard undergraduate
research journal

Spring 2025
Vol. 15, Issue 2





thurj the harvard
undergraduate
research journal

Spring 2025 Best Manuscript

Mollie Cheng

Mollie Cheng '26 is a junior at Harvard College, originally from Taiwan. She studies Economics with a secondary in Neuroscience and is interested in the intersection of mental health, economics, and the law. Outside of academics and coursework, Mollie is involved in the Taiwanese Cultural Society, Indigo Peer Counseling Group, Harvard College Consulting Group, and Undergraduate Negotiation Club. In her free time, Mollie enjoys ceramics, painting, and board games. Mollie's manuscript regarding cannabis legalization and mental health hospitalization outcomes was produced as a final research paper for Econ 970: "Applying Economic Theory in the Practice of Law."

Submit your research to

thurj

The Harvard Undergraduate Research Journal

THURJ is the *only* campus publication that showcases peer-reviewed undergraduate student research.

Circulation to dorms, departments, libraries, and scientific institutions.

Submit your research, manuscript or thesis!

Reference www.thurj.org/submit for submission guidelines.

THURJ accepts submissions that reflect original research in all disciplines!

\$200 for Top Manuscript!

For more info, including our latest issues, visit www.thurj.org.

To contact the Executive Board, contact thurjall@gmail.com or visit www.thurj.org/board.

May 2025

Dear Harvard Community,

We are thrilled to present the Spring 2025 issue of *The Harvard Undergraduate Research Journal* (THURJ), a student-run, biannual publication committed to showcasing and celebrating exceptional research from Harvard undergraduates of all disciplines and interests.

Since its founding in 2007, THURJ has served as a platform for student scholarship. Like many student initiatives, however, THURJ faced an abrupt halt in 2021 amid the COVID-19 pandemic. Over the past two years, we are proud to share that THURJ has not only been revived but has flourished. With over 80 active members, a rapidly growing online readership, and an increasingly selective submission process, THURJ now stands among the largest and most competitive undergraduate research journals in the world.

Yet, THURJ’s mission extends far beyond publication alone. At our core, we are a community driven by a shared commitment to expanding access to research and fostering intellectual curiosity and enthusiasm among the next generation of scholars, especially among those lacking traditional opportunities to engage in research. In that spirit, over the past year, we have significantly expanded the THURJ Research Competition, where we inspire, mentor, and encourage high school and middle school students to pursue independent research. With over 200 participants from across the globe, we are deeply proud of the impact THURJ members make on the future generation of scientists, innovators, and philosophers. Our members have also launched initiatives ranging from DOI integration and Harvard Library collaborations to organizing national Undergraduate Research Journal Summits and panels with leading researchers—efforts that reflect our desire to redefine what student-led research engagement can look like. It’s with this passion that our members have worked tirelessly to make an impact on the field in any way possible, beyond their general organizational responsibilities, something that has been a true inspiration.

Therefore, first and foremost, we would like to extend our sincere gratitude to each of the members of THURJ for the passion and dedication they display in making the work at THURJ possible. We also thank our faculty reviewers, whose generous insights remain a cornerstone of the THURJ editorial process. Next, we thank our wonderful faculty advisor, Dr. Andrew Berry, an individual who shares our deep passion for promoting youth research. Finally, we would like to extend our immense gratitude to the Office of the Dean of Science Education, the Office of the Dean of Harvard Medical School, and the Office of the Dean of the School of Engineering and Applied Sciences, whose incredible generosity has enabled us to continue our work with THURJ. It is because of the incredible support we have received that we at THURJ are thrilled to present to the Harvard community the Spring 2025 publication of *The Harvard Undergraduate Research Journal*! In this edition, we feature six cutting-edge, insightful works of undergraduate research and six fascinating pieces of commentary on the state of research at Harvard and beyond. We hope you enjoy reading and learning from these pieces as much as we did!

Sincerely,



Hugh Hankenson
Editors-in-Chief



Aditya Tummala

Table of Contents

Research

- 9

Examining the Mental Health Impact of the Legalization of Recreational Cannabis: A Comparative Analysis of Massachusetts and Rhode Island
Mollie Cheng '26
- 17

Mental Health In Crisis: A Case-Study Analysis of Syria
Maryam Guerrab '25
- 25

EXPLAIN THIS, PRUNER! The Effect of Zero-Order Pruning on LLM Explainability and Curvature
Joseph Bejjani '26, *Camilo Brown-Pinilla* '26, *David Ettel* '26
- 37

Examining Period Poverty and Menstrual Equity in Nepal
Alissar Dalloul '27
- 43

Ramanujan-Nagell Equation and Elliptic Curves
Lale Baylar '28 and *Karin Lund* (Dartmouth College '29)
- 50

Nullius in Verba: Artists, Corpses and Empiricism in Post-Enlightenment England and Beyond
Taylor Larson '25

Features

- 59

Are We Entering the Dark Ages of Science?
Leah Lourenco '26
- 62

The Neuroscience of Time Perception and the Ethics of Altering It
David Kim '27
- 66

Keeping Scientists Accountable: Combating Fraudulent Research with Post-Publication Scrutiny
Antonino Libarno '28
- 70

Rethinking Alzheimer’s: A New Autoimmune Perspective
Hiteyjit Singh Gujral '27 (Visiting Undergraduate Student)
- 74

Aristotle’s Genome: From the Origins of Form to Modern Code
Krishna S. Rajagopal '28
- 79

A Modern Primer on Consciousness for a Neurological Lens
Avery Mizrahi '28

Executive Board

Editors-In-Chief

Hugh Hankenson ’26
Aditya Tummala ’26

Co-Managing Editors of Content

Saketh Sundar ’27
Theo Tobel ’27

Co-Business Managers

Avi Agarwal ’27
Natasha Kulviwat ’28

Co-Managing Editors of Design

Catherine Feng ’27
Natalie Zhang ’27

Co-Managing Editors of Peer Review

David An ’27
April Keyes ’26

Co-Technology Managers

Yewon Lee ’26
Himani Yarlagadda ’27

Co-Internal Relations Managers

Kow Simpson ’26
Maria Xu ’27

Co-External Relations Managers

Charles Bratton ’27
Pranathi Ganti ’28

Associate Board

Associate Editors of Peer Review

Audrey Limb ’27
Marvel Hanna ’27
Tianna Tout-Puissant ’27
Gianna Tout-Puissant ’27
Grace Zhang ’27
Danielle Im ’27

External Relations Associate

Adan Eftekhari ’28

Associate Editors of Content

Héctor Andrés Martínez Luna ’27
Joanne Ji ’27
Sophie Gao ’28

Business Associates

Neo Hou ’27
Virgil Guo ’28
Ryan Whalen ’28

Associate Editors of Design

Iris Sung ’27
Lale Baylar ’28

Technology Associates

James Pelaez ’27
Hanjing Wang ’28

Internal Relations Associate

Sadeea Morshed ’28
Tiffany Huang ’26

General Board

Designers

Bridgette Pehrson ’27
Nishi Patel ’28

Original Content Writers

Leah Lourenco ’26
David Kim ’27
Daleyvon Knight ’27
Lara Rahman ’28
Avery Mizrahi ’28
Manya Gupta ’28
Hiteyjit Singh Gujral ’27
Krishna S. Rajagopal ’28
Antonino Libarnes ’28

Peer Reviewers

Mizan Abraha ’27
Parth Nikumbh ’28
Nina Khera ’27
Ava Pakravan ’28
Sunny Shi ’26
Jiajia Zhang ’27
Wyatt Jensen ’27
Umaama Hussain ’27
Gavriela Kalish-Schur ’28
Yash Ravipati ’28
Kevin Lim ’28
Elizabeth Norris ’27
Leah Lourenco ’26

Daleyvon Knight ’27
Manya Gupta ’28
Vanessa Norris ’26
Akmal Hashad ’27
Ella Borgman ’27
Alexa Fein ’28
Elena Ferrari ’28
Akmal Hashad ’27
Zhixiao Yip ’27
Stephanie Dragoi ’28
Eric Gong ’27
Hiteyjit Singh Gujral ’27
Clifford Stow ’26
Antonino Libarnes ’28

Faculty Advisory Board

Andrew Berry, Ph.D.
Assistant Head Tutor, Integrative Biology, Lecturer in Organismic and Evolutionary Biology

Hopi Hoekstra, Ph.D.
Edgerley Family Dean of the Faculty of Arts and Sciences, C.Y. Chan Professor of Arts and Sciences, Professor of Organismic and Evolutionary Biology and of Molecular and Cellular Biology, Alexander Agassiz Professor of Zoology in the MCZ

Robin Kelsey, Ph.D., J.D.
Dean of Arts and Humanities, Faculty of Arts and Sciences
Shirley Carter Burden Professor of Photography History of Photography and American Art

L. Mahadevan, Ph.D.
Lola England de Valpine Professor of Applied Mathematics, of Organismic and Evolutionary Biology, and of Physics

Christopher Stubbs, Ph.D.
Samuel C. Moncher Professor of Physics and of Astronomy; Dean of Science in the FAS

Martha Whitehead, M.L.S.
Vice President for the Harvard Library and University Librarian
Roy E. Larsen Librarian for the Faculty of Arts and Sciences

Faculty Review Board

Austin Conner, Ph.D.
Postdoctoral Fellow and Lecturer, Department of Mathematics

Max Weinreich, Ph.D.
Postdoctoral Fellow and Lecturer, Department of Mathematics

Aaron Landesman, Ph.D.
Postdorctoral Fellow, Department of Mathematics

Matlin Gilman, MPH
Ph.D. Candidate, Population Health Sciences

Felipe Pereda, Ph.D.
Director of Graduate Studies, Department of History of Art and Architecture
Fernando Zóbel de Ayala Professor of Spanish Art

Sarah S. Richardson, Ph.D.
Aramont Professor of the History of Science
Professor of Studies of Women, Gender, and Sexuality

Eram Alam, Ph.D.
Assistant Professor of the History of Science

Bohan Li
Ph.D. Candidate in Health Policy Management, Harvard Business School

Licensing

The Harvard Undergraduate Research Journal (THURJ) is an open-access publication that publishes under the CC BY-NC-ND license. Authors retain copyright, granting THURJ the right of first publication. The license permits non-commercial sharing of articles as long as appropriate credit is provided, but prohibits derivative works. Authors may also establish separate non-exclusive agreements for further distribution, with acknowledgment of THURJ as the original publisher.

Copyright © 2025 The Harvard Undergraduate Research Journal

About Us

The Harvard Undergraduate Research Journal (THURJ) showcases peer-reviewed undergraduate student research from all academic disciplines. As a biannual publication, THURJ familiarizes students with the research publication process. This process not only stimulates faculty student collaboration and provides students with valuable feedback on their research, but also promotes collaboration between the College and Harvard’s many graduate and professional schools. In addition to publishing original student research papers, THURJ keeps the Harvard community updated on and provides an important forum for discourse on the cutting-edge research that impacts our world today.

About the Cover

Kaitlyn Zhou ’25

The cover art explores the dynamic relationship between the microscopic and macroscopic. The image draws on the intricate structure of the human brain, where individual neurons form vast, interconnected networks. The flowing lines suggest both synaptic pathways and planetary orbits to invoke a visual parallel between neural activity and cosmic systems. Just as microscopic signals give rise to complex thought and behavior, large-scale patterns in society and science emerge from countless small interactions. Thus, the design reflects research as a constant movement across scale: zooming in to uncover detail, zooming out to find meaning.

Research

Examining the Mental Health Impact of the Legalization of Recreational Cannabis: A Comparative Analysis of Massachusetts and Rhode Island

Mollie Cheng
Harvard College '26

As mental health gains prominence in public health discourse, scientific literature on the relationship between recreational cannabis use and mental health outcomes remains mixed. How does the legalization of recreational cannabis affect mental health-related hospitalizations, and how do these outcomes vary across diagnoses and subpopulations? While cannabis use has been linked to disorders such as psychosis and depression, it has also demonstrated therapeutic and stress-relieving effects. Despite ongoing debate, much of the existing scientific and economic literature is inconclusive, outdated, or lacking nuance. This study compiles results across seven mental health classification groups, finding that the legalization of recreational marijuana—particularly the start of sales in 2019—was associated with a highly statistically significant reduction of approximately 75 psychosis-related hospitalizations per 100,000 people in Massachusetts relative to Rhode Island ($p < 0.01$). Using a differences-in-differences regression, this study observes parallel pre-treatment trends that provide a strong preliminary case for further causal research on the mental health impact of cannabis legalization.

Introduction

Since California’s legalization of medical marijuana in 1996, 24 states plus Washington, D.C. have legalized recreational marijuana as of January 2024. Public support has grown in tandem, with nearly 90% of U.S. adults supporting some form of legal marijuana use (Pew Research Center 2024). Alongside this policy shift, researchers studied legalization’s effects on crime, public safety, and increasingly, public health. Yet, whether recreational legalization worsens or alleviates mental health outcomes remains unclear. While the CDC links cannabis use to higher risks of psychosis, schizophrenia, and suicide (National Academies 2017), marijuana is also associated with benefits such as pain and stress relief (APA 2018).

Using a Differences-in-Differences methodology, this study evaluates whether Massachusetts’ legalization of recreational marijuana—enacted in 2016 and implemented through retail sales in 2019—affected mental health-related hospitalizations compared to Rhode Island, which did not legalize recreational use during the study period (2008–2021). To claim causality, parallel trends before the treatment (legalization) should be observed between the control group (Rhode Island) and treatment group (Massachusetts). Recreational marijuana laws (RMLs) are studied over medical marijuana laws (MMLs) for their broader impact and unique relevance for studying public health impacts. Hospitalization data offer an additional perspective to existing survey-based studies.

Results indicate that the 2019 retail rollout of recreational cannabis corresponded to a significant decrease of ~75 psychosis hospitalizations per 100,000 residents in Massachusetts compared to in Rhode Island. These preliminary results highlight the need for more robust research into possible causal effects of legalization.

This paper includes a literature review, description of the research design and identification strategy, results with robustness checks, discussion of methodological limitations, and policy implications.

Scope and Limitations of Research

Massachusetts and Rhode Island are adjacent states with broadly comparable healthcare infrastructure, age distributions, and socioeconomic characteristics, which allow for more confidence in attributing outcome differences to legalization policy. However, given the limited number of aggregate observations, findings should be interpreted with caution. Standard errors clustered at the state level could reduce the statistical significance of observed effects, and this small sample size limits the strength of causal claims.

This analysis is restricted to hospitalization data, which captures only the most severe mental health outcomes. While this ensures objectivity and data quality, it excludes less severe cases and may be an imperfect proxy for mental health. Additionally, co-use with other substances and delayed mental health responses are not accounted for. Potential spillover effects further complicate interpretation; Rhode Island residents could have accessed marijuana in Massachusetts, potentially biasing estimates downward. The presence of illegal markets before and after legalization may also attenuate treatment effects.

Given these constraints, results should be viewed as preliminary and exploratory. The evidence suggests a possible association between legalization and reduced hospitalization rates for psychoses, but more robust research, with larger samples and stronger identification strategies, is needed to draw firm conclusions.

Literature Review

In this literature review, a range of studies within the past ten years are reviewed to inform this study. This section synthesizes (1) cannabis’s effects on mental health from a medical perspective, and (2) existing work that examines the impacts of legalization on mental health outcomes. Gaps in the literature are noted throughout.

Cannabis’s medical effects on mental health are still hotly debated. On one hand, some claim that it helps to relieve stress. Studies conducted by the American Psychological Association found that mental health patients using cannabis for medical purposes had “largely improved cognitive performance, reduced clinical symptoms and anxiety-related symptoms as well as a reduced use of conventional medications, including opioids, benzodiazepines, and other mood stabilizers and antidepressants” (APA 2018, p. 2). On the other hand, researchers observe that cannabis use increases the severity and probability of developing mental health disorders. Marconi et al. demonstrates that higher levels of cannabis use increase the risk of psychotic outcomes. Although “a causal link cannot be unequivocally established, there is sufficient evidence to justify harm reduction prevention programs” (Marconi 2016, p. 8).

Reflecting these mixed medical interpretations, existing literature on legalization and mental health has also proven quite mixed. Most studies focus on MMLs, which are limited in accessibility, and often use extreme outcome variables like suicide rates. Bartos et al. (2020) and Anderson et al. (2014) found a “strong negative relationship” between MMLs and suicide. However, suicide remains an imperfect proxy for mental health. Grucza et al. (2015) challenged these findings using individual-level data, finding no statistical relationship after controlling for covariates.

Singer et al. (2020) expanded the scope to RMLs, concluding that “adverse mental health outcomes do not follow cannabis liberalization at the state level,” though the benefits may primarily affect young populations. Yet, Borbely et al. (2022) found no average effect of RMLs on mental health. This study relied on the outcome variable “days of bad mental health” in the past thirty days—data collected through surveys conducted via landlines and cell phones. This methodology introduces selection bias, as it excludes individuals without regular phone access and relies on subjective, self-reported measures. Similarly, Elser et al. (2023) found no significant aggregate change in psychosis-related outcomes, but noted statistically significant increases in certain subgroups. However, Elser’s dataset includes only insured individuals, which may skew representativeness over time.

Recent contributions highlight persistent gaps. With the rise of RMLs, Anderson and Grucza’s findings are due for an update. Many studies still rely on extreme or narrow outcomes. Beyond suicides, various other outcomes could be of interest, from self-reported mood to medically verifiable hospitalization data. Furthermore, the use of suicide as a proxy assumes that all diagnoses are equally affected by legalization. To obtain more targeted results, work remains to be done in differentiating legalization effects on specific disorders. Perhaps these limitations contribute to differing significance levels across studies, as incorrectly aggregating variables tends to bias results toward insignificance as divergent effects cancel each other out.

Given inconclusive and contradictory findings, outdated scopes, extreme proxies, and unrepresentative sampling methods, this paper aims to (1) add to the growing literature on the public health implications of RMLs and (2) highlight differential effects of cannabis legislation on different diagnoses and subgroups. Improving upon Borbely’s study, objective hospitalization data rather than survey data is used. This study segments interpretation of mental health outcomes to allow for more targeted policy implications, using hospitalization data drawn from a

representative sample—regardless of payer status. There is little existing scholarship that quantitatively assesses RML effects on nuanced subgroups to this extent.

Research Methods and Data Analysis

This section outlines the analytical approach used in the study. It begins by justifying the use of a Differences-in-Differences (DID) design and its key assumptions, followed by an evaluation of Massachusetts and Rhode Island as comparable case studies. The section then describes the treatment, control, and outcome variables, along with their data sources. Finally, it presents the tested regressions and corresponding hypotheses. The following section discusses the results, their implications, and key technical limitations.

a) DID Regression Design

To estimate the impact of recreational marijuana legalization on mental health outcomes, this study draws on the methodology of Differences-in-Differences (DID). This approach compares changes in hospitalization rates over time between a “treatment” group—Massachusetts, which legalized recreational use, and a “control” group—Rhode Island, which did not. By tracking trends before and after legalization in both states, the DID method helps isolate the effect of the policy, assuming that both states followed similar trends before the law changed.

b) Comparing Massachusetts (Treatment Group) and Rhode Island (Control Group)

First, it would be necessary to examine a variety of demographic characteristics of Massachusetts and Rhode Island to evaluate whether they are “true comparisons.” Although it is difficult, if not impossible, to find groups with identical characteristics outside of RMLs, Massachusetts and Rhode Island proved to be quite similar in political leanings, age distribution, healthcare provision, and proportion of college students despite some differences in racial demographics and religion.

Politically, both Massachusetts and Rhode Island lean very Democratic. House and Senate representatives from both Massachusetts and Rhode Island have been Democratic-affiliated since 1992 (Ballotpedia 2024). This might suggest similarities in state-specific policies, despite inevitable nuances that should ideally be further controlled for. This paper assumes that this can help (to some extent) with controlling for potential policy-related confounders, such as Covid-19 response and statewide economic policies.

Massachusetts and Rhode Island have similar age distributions as of the most recent 2021 Census Data. Rhode Island’s population skews slightly older, but the relative distribution seems quite similar with percentages differing by less than 1 percent. Massachusetts has a higher median household income of \$99,858 ± \$1,355, while Rhode Island’s median falls at \$84,972 ± \$2,566, but both states share similar poverty rates.

Importantly regarding healthcare provision and quality, both states expanded Medicaid under the Affordable Care Act, providing similar healthcare access and mental health treatment availability. They rank second and third for best healthcare in the United States, behind Minnesota (Forbes 2023, para. 10-11). Massachusetts has the second-highest number of primary care physicians (23.34 per 10,000 state residents) and the lowest percentage of

residents who lack health insurance coverage (2.50%). Rhode Island boasts the highest number of primary care physicians (25.89 per 10,000 state residents) and the fourth-lowest percentage of residents who lack health insurance coverage (4.34%). This is an important criterion, as similar insurance adoption and healthcare availability minimizes institutional confounding variables (e.g., underreporting of diagnoses due to a lack of hospitals).

However, Massachusetts and Rhode Island have varied racial composition. Compared to Massachusetts, Rhode Island has a slightly higher, slightly lower, and notably higher proportion of white, black, and Hispanic residents, respectively as of 2021 Census data. Therefore, controlling for race and ethnicity would help make the results more robust. This study controlled for % White, % Black, and % Hispanic in the regressions.

In conclusion, while Massachusetts and Rhode Island may differ in terms of race, they are overall good comparisons. While other unconsidered confounding factors may exist, this study attempts to control for race to minimize demographics-induced differences.

c) Variables and Data

This subsection discusses the treatment, control, and outcome variables in depth. This study’s outcome variable is “rate of hospitalization discharges” for each diagnosis, as specified below. The treatment variables are (1) 2017, the first effective year of RML implementation in Massachusetts (RMLs were passed in late-2016) and (2) 2019, the start of sale year for recreational marijuana in Massachusetts (start of sale began in late-2018). The regressions account for different racial/ethnic demographics by including control variables for the percentage of white, black, and Hispanic/Latino people based on data from the U.S. Census Bureau’s official American Community Survey (ACS), an annual demographics survey program. All data sources are officially released data from the U.S. government.

The two treatment variables are binary indicators of 1) whether RMLs were effectively legalized in Massachusetts and 2) whether recreational cannabis is being sold in Massachusetts. Drawing from state legislation archives, the regressions accounted for a staggered legalization timeline in Massachusetts, with legalization entering effect on December 15, 2016, and recreational sales starting on November 20, 2018. Due to both changes occurring towards the end of the year, treatment points are set at 2017 (effective legalization date) and 2019 (start of recreational sales). Rhode Island passed MMLs on January 3, 2006, and only passed RMLs on May 25, 2022, after the hospitalization dataset time frame of 2008-2021. As stated before, this study operates on the informed assumption that MMLs are not widely accessible, as strict medical requirements and physician recommendations are required to access medical marijuana in Massachusetts and Rhode Island (refer to section “Scope and Limitations of Research”). Therefore, this paper assumes that Rhode Island’s earlier passage of MMLs had negligible effects on mental health hospitalization rates and that RMLs, once introduced, become the main source of cannabis consumption for both recreational and dual users. This mitigates challenges of differentiating medical, recreational, and dual use – RMLs are much more widely accessible, regardless of ability to obtain medical marijuana.

The outcome variable is segmented by diagnosis to illuminate differential effects between conditions. To improve upon existing

literature that uses subjective survey data and unquantifiable personal outcome variables such as “bad mental health days” (Borbely et al (2022), this paper measures “hospitalization discharge rates per 100,000 people,” as the final outcome variable to compare the different population sizes of Massachusetts and Rhode Island. Raw hospitalization discharge numbers were drawn from the U.S. Department of Health and Human Services before being combined with corresponding state population data from the U.S. Census Bureau’s ACS. Integrating these two data sources, hospitalization discharge rates were calculated as such: **Hospitalization Discharge Rate = Total Hospitalization Discharge Number/ Total Population**, per year, per diagnosis.

Hospitalization data was retrieved from the U.S. Agency for Healthcare Research and Quality (AHRQ), an official database maintained by the U.S. Department of Health and Human Services. The database’s Healthcare Cost and Utilization Project (HCUPnet) aggregates longitudinal data on hospital inpatient stays. State-specific data contains all inpatient care records in forty-two participating states, including inpatient discharge records.

This dataset is comprehensive and credible. Data on Massachusetts was provided by the Massachusetts Center for Health Information and Analysis (an agency of the Commonwealth of Massachusetts), and data on Rhode Island was provided by the Rhode Island Department of Health. Information is recorded “regardless of payer.” This study selected only diagnoses numbers listed under the “Mental Diseases & Disorders” category on the official U.S. Department of Health and Human Services website (US. Department of Health and Human Services).

Adopting the Department of Health and Human Service’s “Mental Diseases & Disorders” classification, this study analyzes hospitalization discharge rates for Diagnosis Related Groups (DRGs) of:

- Acute Adjustment Reaction & Psychosocial Dysfunction;
- Depressive Neuroses;
- Neuroses Except Depressive
- Disorders of Personality and Impulse Control;
- Organic Disturbances & Intellectual Disability;
- Psychoses; and
- Behavioral and Developmental Disorders.

Each of these DRGs are “umbrella” classifications for more specific symptoms. It is a system used by Medicare to categorize “all patients, regardless of payer.” and determine how much a hospital is paid for treating them. Medicare’s payment to the hospital under the MS-DRG system is calculated based on the patient’s “principal diagnosis, up to 24 additional diagnoses, and up to 25 procedures performed during the stay” (Centers for Medicare and Medicaid Services 2024, para. 7). Patients are unlikely to be “double-counted,” or sorted into two DRGs at once, as patients are only reimbursed for the primary DRG they are sorted into for accurate reimbursement of hospital fees.

d) Hypothesis Testing and Regression

Finally, hypothesis testing empirically assesses the impact of RMLs on mental health outcomes between the two states. The two test hypotheses are:

H₀ (null): Recreational legalization *does not* statistically significantly affect hospitalization rates of X diagnosis, where X corresponds to each of the diagnoses listed above.

H_a (alternative): Recreational legalization *does* statistically significantly affect hospitalization rates of X diagnosis.

Lastly, four regressions for each diagnosis were ran:

- Regression (1) examines the impact of the 2017 treatment, without controlling for race.
- Regression (2) examines the impact of the 2019 treatment, without controlling for race.
- Regression (3) examines the impact of the 2017 treatment, controlling for race.
- Regression (4) examines the impact of the 2019 treatment, controlling for race.

Results and Discussion

a) Primary Analysis per Diagnosis: Baseline Regression and Controls

This section discusses the study’s results. Borders of each regression table are color-coded for interpretive convenience, with green indicating statistical significance ($p < 0.05$) in all four regressions and parallel pre-treatment trends, orange indicating statistical significance for one or more regressions and no parallel pre-treatment trends, and gray indicating no statistical significance in any regression and no parallel pre-treatment trends. All figures are included in the Appendix. For the following analyses, a coefficient is deemed “statistically significant” at a p-value of 0.05 or lower.

Results for Neuroses Except Depressive; Disorders of Personality and Impulse Control; and Behavioral and Developmental Disorders were not statistically significant for any of the regressions. Results for Acute Adjustment Reaction & Psychosocial Dysfunction; Depressive Neuroses; Organic Disturbances & Intellectual Disability; and Psychoses were statistically significant for one or more regressions, though all failed to meet the parallel trends condition. This makes a weak case for causation.

For Psychoses, however, the parallel trend assumption seems to hold after 2014. As shown in Figure 5, the distance between the black and red lines appears consistent across time until after the first post-2017 treatment. This fulfills a key criterion for causal analysis through DID regressions. Furthermore, outcomes for psychoses are statistically significant in all but one regression. Most notably, post-2019 recreation sales corresponded to a statistically significant decrease in 60.04 hospital discharges per 100,000 people at the $p < 0.05$ level and an even larger statistically significant decrease in 74.92 hospital discharges per 100,000 people at the $p < 0.01$ level after controlling for race. The increase in statistical significance after controlling for potential confounders seems to suggest that the legalization of recreational cannabis possibly caused a decrease in about 75 Psychoses hospitalizations per 100,000 individuals after recreational sales started in 2019.

One potential explanation of this relationship is that access to recreational marijuana may alleviate symptoms in individuals with psychotic disorders, perhaps due to self-medication effects or reduced reliance on substances with higher psychosis risks, such as alcohol or synthetic drugs (American Psychological Association 2018, p. 2). Surprisingly, however, some medical research has found a positive association between cannabis use and symptoms of psychosis. A literature review published in 2020 finds that “the scientific literature indicates that psychotic illness arises more frequently in cannabis users compared to non-users, and cannabis users have an earlier onset of psychotic illness compared to

non-users. Cannabis use was also associated with increased relapse rates, more hospitalizations and pronounced positive symptoms in psychotic patients” (Hasan 2020, p. 1).

b) Discussion of Results for Psychoses

Parallel trends after 2014 for Psychoses could be associated, at least in part, with the implementation of the Affordable Care Act (ACA), which was signed into law on March 23, 2010, but not fully implemented until January 1, 2014. The ACA expanded Medicaid eligibility and prohibited insurance companies from denying coverage due to preexisting conditions, increasing accessibility of health insurance (US Department of Health and Human Services 2022, para. 2). Very importantly, the ACA “requires coverage of mental health and substance use disorder services as one of ten essential health benefit (EHB) categories in non-grandfathered individual and small group [insurance] plans” (US Department of Health and Human Services 2014, para. 43). Additionally, the ACA is widely accredited for strengthening the Mental Health Parity and Addiction Equity Act (MHPAEA) of 2008, a previously unenforced federal law that prohibited health insurers from “imposing less favorable benefit limitations on [mental health or substance use disorder (MH/SUD)] benefits” (Centers for Medicare and Medicaid Services 2024, para. 1). For these two reasons, it is plausible that the full implementation of the ACA in 2014 1) increased affordability of healthcare in general, 2) increased insurance coverage rates of MH/SUD, and 3) enforced equal allocation of benefits among mental and physical healthcare insurance. These effects could explain the parallel trends in Massachusetts and Rhode Island starting from 2014, as the ACA democratized insurance coverage and healthcare access nationally. The ACA could have 1) generated greater awareness of mental healthcare and 2) decreased both actual and perceived healthcare costs for everyone regardless of payer or state of residency.

Qualitative parallel trends after 2014, paired with strong and statistically significant corroboration, especially after the ACA, suggests a potential negative causal effect. Yet, more research should be conducted to test this result alongside different control variables beyond race/ethnicity. As discussed in the following “Limitations in Methodology and Data” section, using relatively few control variables due to few data points limits confidence that this relationship is truly causal. As more comprehensive datasets are discovered and more controls are tested in future studies, the results of this study could perhaps serve as a step towards incrementally proving causation.

Limitations in Methodology and Data

This study makes several assumptions, as mentioned throughout this paper. 1) It assumes that hospitalization discharges are a close proxy for true hospitalization counts. 2) While Massachusetts and Rhode Island share many demographic similarities, they are not identical comparisons. It also assumes that similar political leanings between Massachusetts and Rhode Island minimize policy-related differences, 3) that Rhode Island’s earlier passage of MMLs had negligible effects on mental health hospitalization rates, and 4) that RMLs, once introduced, become the main source of consumption for both recreational and dual users. These assumptions, if challenged, could challenge the generalizability of these results.

There are several technical limitations to this study that highlight opportunities for further research. Most notably, hospitalization data is only publicly available at the state-year level. This raises challenges regarding statistical power and the feasibility of the results, especially given the relatively large estimated effects. Perhaps clustering standard errors at the state level would reduce the significance of the findings, warranting cautious interpretation.

The study controls for race but does not include other potential confounders due to data limitations. The absence of more granular hospitalization and census data (e.g., monthly or quarterly) restricts the ability to include additional controls without further diminishing statistical power. Ideally, the analysis would incorporate controls for variables such as age, education, urbanization, religion, profession, and healthcare access, as well as factors like insurance coverage and state-level healthcare policies. This limitation underscores the importance of experts with granular hospitalization data access to study this topic. Without richer data or more frequent observations, the conclusions of this study remain tentative.

Lastly, it is argued by some that cannabis acts as a “gateway” substance, so the results of this study could be informed by further studies examining the effect of recreational legalization on other mental health-associated predictors, such as alcohol, opioid, and tobacco consumption.

Conclusions and Policy Implications

By employing credible and objective state-level longitudinal hospitalization data, this paper expands upon emerging literature regarding RMLs as opposed to MMLs. It also quantitatively evaluates outcomes segmented by diagnosis and demographic group, offering a more nuanced investigation of mental health effects as caused by RMLs. Ultimately, results show a statistically significant decrease in psychosis hospitalizations both immediately after RMLs became effective in 2017 and after recreational sales began in 2019. While still a long way from definitively proving causation, these findings suggest that more research should be done to further refine this model and inch towards investigating true causal effects.

Thus, recreational cannabis as it affects psychosis should continue to be researched. Studying alternative outcomes (e.g., therapy visit rates) could complement existing findings and provide a broader view of how RMLs affect psychosis-related healthcare.

Policymakers could consider utilizing recreational cannabis as a supplementary tool for mental health interventions, particularly for conditions like psychosis. They could also bolster drug education programs, increase access to mental health resources to encourage greater utilization. Policymakers should fund more careful tracking of the therapeutic effects of cannabis, contributing to an ever-growing conversation between medicine, policy, and public health.

Mental health is often a secondary consideration in drug policy debates. The findings of this paper suggest that mental health should play a more central role. Given the profound impact of mental health on individuals and communities, integrating it into policy frameworks is essential.

Appendix:

Figure 1 displays how hospitalization rates have changed over time in both Massachusetts and Rhode Island. “Legalization Status” on the y-axis is a binary variable, with 0 indicating *not recreationally*

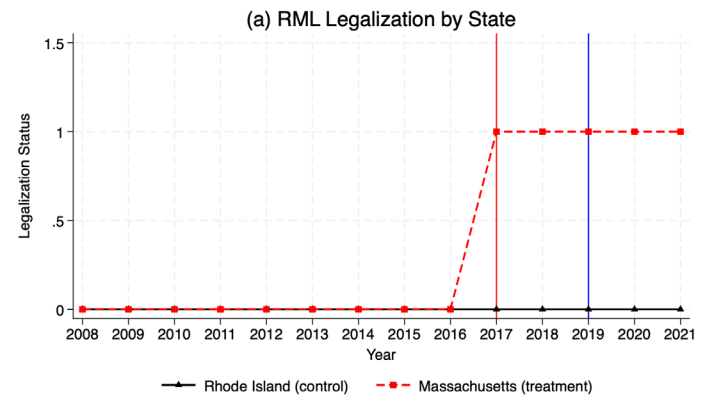


Figure 1. Treatment Timeline: Legalization Status

legalized and 1 indicating *recreationally legalized*. The red vertical line indicates the first full effective implementation year of Massachusetts RMLs (2017). The blue vertical line indicates the first full year of legal recreational sales in Massachusetts (2019). Massachusetts’s (treatment group) timeline is indicated by the red dotted line, and Rhode Island’s (control group) timeline is indicated by the black solid line.

In Figure 2b (and Figures 3b-8b below), the regression results are presented in four columns: Columns 1 and 3 show regressions for the period post-2017 (after RML implementation), with and without control variables, while Columns 2 and 4 display regressions for the

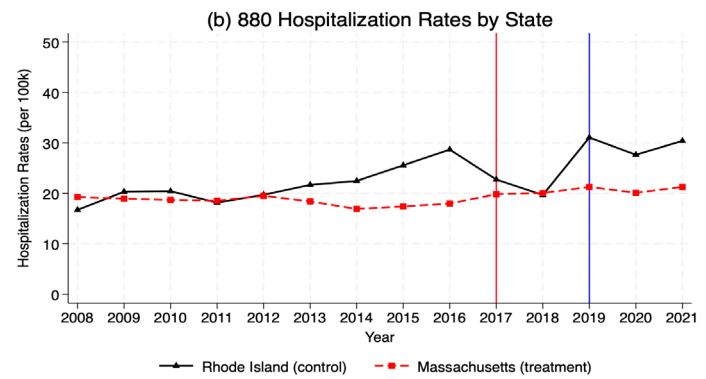


Figure 2a. Acute Adjustment Reaction & Psychosocial Dysfunction

VARIABLES	(1)	(2)	(3)	(4)
	No Controls	No Controls	Control Race	Control Race
Massachusetts	-3.117** (1.285)		-2.632 (3.247)	
PostTreatment	4.780* (2.476)		-1.771 (3.169)	
DID	-2.680 (2.511)		-0.616 (3.810)	
Massachusetts2		-2.779** (1.091)		-2.617 (1.526)
PostTreatment2		8.235*** (1.392)		6.176** (2.724)
DID2		-6.041*** (1.467)		-4.507* (2.469)
percwhite			-76.03** (30.59)	-18.57 (14.40)
percblack			141.9 (230.5)	42.42 (63.08)
perchispanic			116.6* (56.03)	50.04 (47.76)
Constant	21.52*** (1.252)	21.47*** (1.047)	59.18*** (13.29)	27.45*** (8.164)
Observations	28	28	22	22
R-squared	0.487	0.687	0.699	0.753

Figure 2b. Acute Adjustment Reaction & Psychosocial Dysfunction

period post-2019 (after the start of legal recreational sales), also with and without controls. The variable “Massachusetts” represents the treatment group, where 1 denotes Massachusetts and 0 represents Rhode Island. “PostTreatment” is a binary variable indicating the years following the implementation of the RMLs (2017) and the start of recreational sales (2019). The “DID” interaction term measures the treatment effect of recreational legalization. Additionally, the control variables “percwhite”, “percblack”, and “perchispanic” account for the racial and ethnic composition of Massachusetts and Rhode Island.

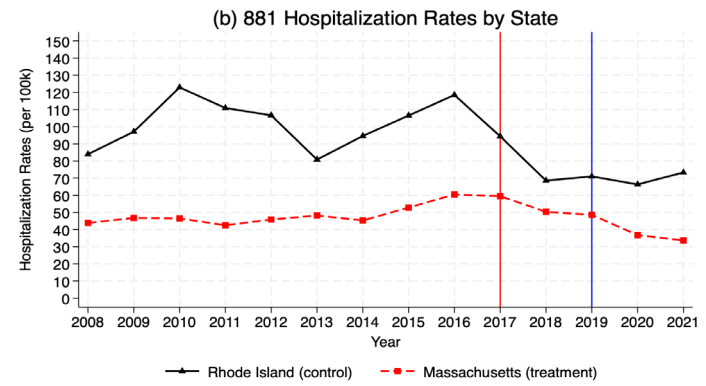


Figure 3a. Depressive Neuroses

VARIABLES	(1) No Controls	(2) No Controls	(3) Control Race	(4) Control Race
Massachusetts	-54.39*** (5.264)		-35.19*** (10.17)	
PostTreatment	-27.72*** (6.930)		-29.36** (10.72)	
DID	25.45*** (8.499)		41.58*** (8.297)	
Massachusetts2		-49.33*** (5.462)		-51.97*** (14.55)
PostTreatment2		-28.41*** (5.440)		-21.63 (15.81)
DID2		18.79** (7.008)		16.01 (15.63)
percwhite			253.6** (109.4)	60.41 (142.4)
percblack			-1,042* (536.8)	-97.59 (494.1)
perchispanic			333.4* (175.0)	-113.1 (348.8)
Constant	102.4*** (4.922)	98.62*** (5.134)	-79.42 (54.67)	73.55 (104.0)
Observations	28	28	22	22
R-squared	0.854	0.836	0.898	0.826

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 3b. Depressive Neuroses

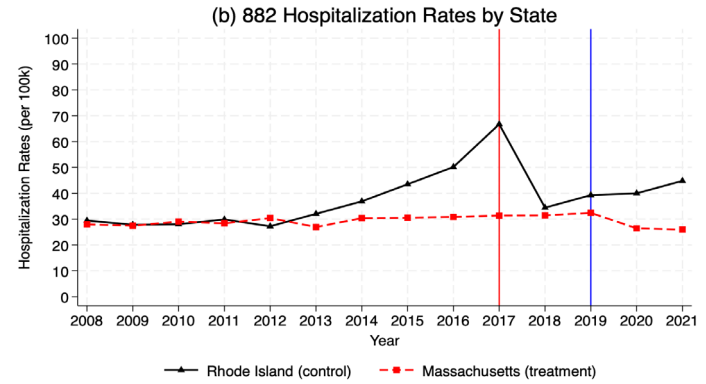


Figure 4a. Neuroses Except Depressive

VARIABLES	(1) No Controls	(2) No Controls	(3) Control Race	(4) Control Race
Massachusetts	-4.785* (2.785)		11.78 (7.840)	
PostTreatment	11.16* (6.120)		6.155 (13.05)	
DID	-10.73 (6.281)		0.331 (7.408)	
Massachusetts2		-7.399* (3.834)		16.69 (11.49)
PostTreatment2		4.434 (4.100)		-7.268 (8.536)
DID2		-5.679 (4.517)		13.81 (13.13)
percwhite			119.1 (128.1)	79.90 (105.1)
percblack			-703.8 (478.9)	-693.2 (460.2)
perchispanic			340.7 (196.8)	555.0* (262.8)
Constant	33.89*** (2.740)	36.92*** (3.801)	-63.16 (79.20)	-60.12 (90.52)
Observations	28	28	22	22
R-squared	0.430	0.266	0.550	0.568

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 4b. Neuroses Except Depressive

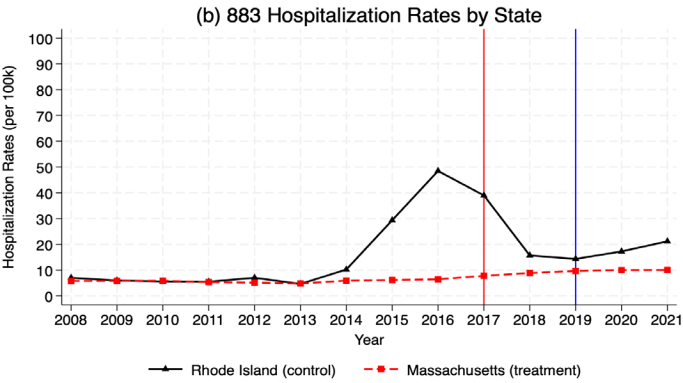


Figure 5a. Disorders of Personality and Impulse Control

VARIABLES	(1) No Controls	(2) No Controls	(3) Control Race	(4) Control Race
Massachusetts	-8.050 (5.134)		10.20 (12.77)	
PostTreatment	7.725 (6.737)		-1.822 (16.18)	
DID	-4.146 (6.752)		8.483 (11.40)	
Massachusetts2		-10.04** (4.825)		13.41 (12.06)
PostTreatment2		1.359 (5.117)		-17.01 (12.47)
DID2		2.380 (5.132)		23.42 (14.49)
percwhite			74.39 (139.2)	5.945 (112.6)
percblack			-745.0 (587.2)	-681.9 (472.7)
perchispanic			435.3 (366.1)	598.7** (261.7)
Constant	13.77** (5.131)	16.23*** (4.811)	-56.74 (101.6)	-27.37 (95.77)
Observations	28	28	22	22
R-squared	0.279	0.216	0.398	0.489

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 5b. Disorders of Personality and Impulse Control

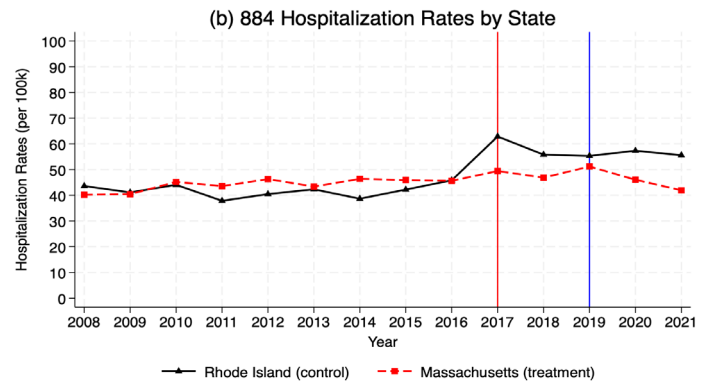


Figure 6a. Organic Disturbances & Intellectual Disability

VARIABLES	(1) No Controls	(2) No Controls	(3) Control Race	(4) Control Race
Massachusetts	2.317* (1.193)		10.40*** (2.114)	
PostTreatment	15.58*** (1.618)		14.62*** (2.872)	
DID	-12.60*** (2.371)		-9.451*** (2.413)	
Massachusetts2		-0.138 (2.516)		17.73** (7.030)
PostTreatment2		11.09*** (2.430)		3.685 (4.442)
DID2		-9.548** (3.496)		3.453 (7.447)
percwhite			70.76** (28.94)	77.84 (67.04)
percblack			-225.1* (127.5)	-331.0 (280.7)
perchispanic			133.1** (48.90)	405.7** (147.5)
Constant	41.79*** (0.870)	44.98*** (2.366)	-19.66 (15.99)	-53.65 (56.51)
Observations	28	28	22	22
R-squared	0.817	0.319	0.893	0.623

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 6b. Organic Disturbances & Intellectual Disability

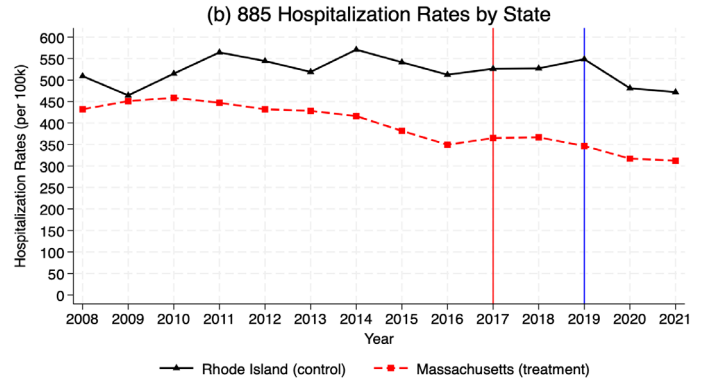


Figure 7a. Psychoses

VARIABLES	(1) No Controls	(2) No Controls	(3) Control Race	(4) Control Race
Massachusetts	-104.9*** (16.31)		-131.2*** (29.35)	
PostTreatment	-15.75 (18.01)		21.16 (25.08)	
DID	-64.51** (24.32)		-50.95* (27.47)	
Massachusetts2		-115.0*** (15.08)		-122.3*** (24.23)
PostTreatment2		-26.24 (23.16)		78.99*** (15.87)
DID2		-60.04** (27.75)		-74.92*** (13.91)
percwhite			762.9** (339.1)	1,312*** (203.2)
percblack			-238.6 (2,104)	-1,535* (854.0)
perchispanic			-680.8 (465.2)	-887.8** (413.1)
Constant	526.7*** (11.05)	526.7*** (9.044)	27.51 (152.1)	-306.4** (130.9)
Observations	28	28	22	22
R-squared	0.842	0.828	0.930	0.950

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 7b. Psychoses

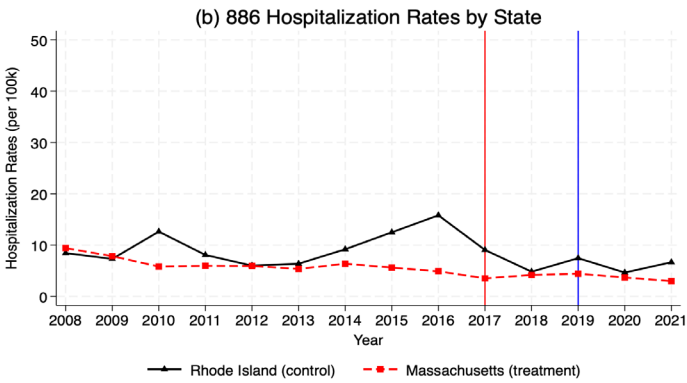


Figure 8a. Behavioral and Developmental Disorders

VARIABLES	(1) No Controls	(2) No Controls	(3) Control Race	(4) Control Race
Massachusetts	-3.229** (1.226)		-0.709 (2.772)	
PostTreatment	-3.063** (1.384)		-4.395 (2.764)	
DID	0.462 (1.484)		3.286 (2.419)	
Massachusetts2		-3.202*** (1.143)		-1.537 (2.273)
PostTreatment2		-2.854** (1.262)		-0.633 (3.003)
DID2		0.648 (1.407)		1.453 (2.708)
percwhite			18.30 (21.33)	36.50** (12.81)
percblack			-136.4 (131.4)	-160.7** (74.01)
perchispanic			75.51 (62.55)	14.48 (58.29)
Constant	9.595*** (1.129)	9.112*** (1.025)	-6.430 (13.55)	-12.05 (8.305)
Observations	28	28	22	22
R-squared	0.496	0.407	0.611	0.510

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 8b. Behavioral and Developmental Disorders

References

American Addiction Centers. (n.d.). The real gateway drug. Retrieved December 11, 2024, from <https://americanaddictioncenters.org/the-real-gateway-drug>

American Psychological Association. (2018, December). The highs and lows of marijuana. Monitor on Psychology, 49(11). Retrieved from <https://www.apa.org/monitor/2018/12/marijuana>

Anderson, D. M., Rees, D. I., & Sabia, J. J. (2014). Medical marijuana laws and suicides by gender and age. American Journal of Public Health, 104(12), 2369–2376.

Anderson, D. M., & Rees, D. I. (2023). The public health effects of legalizing marijuana. Journal of Economic Literature, 61(1), 86–143. <https://www.aeaweb.org/articles?id=10.1257/jel.20211635>

Anderson, P. J. (2015, May 12). R.I. the Most Catholic State. Providence Journal. <https://www.providencejournal.com/story/lifestyle/faith/2015/05/12/r-i-most-catholic-state/34573347007/>

Ballotpedia. (n.d.). Party control of Rhode Island state government. Retrieved November 27, 2024, from https://ballotpedia.org/Party_control_of_Rhode_Island_state_government

Bartos, B. J., Kubrin, C. E., Newark, C., & McCleary, R. (2020). Medical marijuana laws and suicide. Archives of Suicide Research, 24(2), 204–217.

Borbely, C., & Anderson, D. M. (2023). Marijuana legalization and mental health: Evidence from longitudinal data. Journal of Political Economy, 131(5), 1203–1254. <https://doi.org/10.1086/721267>

Centers for Disease Control and Prevention. (n.d.). Cannabis and mental health. Retrieved November 18, 2024, from <https://www.cdc.gov/cannabis/health-effects/mental-health.html>

Centers for Medicare & Medicaid Services. (n.d.). ICD-10-CM code P00.22: Newborn affected by maternal use of marijuana. Retrieved November 27, 2024, from https://www.cms.gov/icd10manual/version33-fullcode-cms/fullcode_cms/P0022.html

Centers for Medicare & Medicaid Services. (n.d.). ICD-10-CM code P00.22. Definitions Manual. Retrieved December 11, 2024, from https://www.cms.gov/icd10m/version372-fullcode-cms/fullcode_cms/P0022.html

Centers for Medicare & Medicaid Services. (n.d.). Mental Health Parity and Addiction Equity Act. Retrieved November 27, 2024, from <https://www.cms.gov/marketplace/private-health-insurance/mental-health-parity-addiction-equity>

Centers for Medicare & Medicaid Services. (n.d.). ICD-10-CM code. Retrieved December 11, 2024, from https://www.cms.gov/icd10m/version372-fullcode-cms/fullcode_cms/P0330.html

Cerdá, M., Mauro, C., Hamilton, A., Levy, N. S., Santaella-Tenorio, J., Hasin, D., Wall, M. M., Keyes, K. M., & Martins, S. S. (2020). Association between recreational marijuana legalization in the United States and changes in marijuana use and cannabis use disorder from 2008 to 2016. JAMA Psychiatry, 77(2), 165–171.

Cleveland Clinic. (2023). Adjustment disorder. Retrieved December 11, 2024, from <https://my.clevelandclinic.org/health/diseases/21760-adjustment-disorder>

Commonwealth Beacon. (2022). Could Massachusetts become a cannabis research hub? Retrieved December 11, 2024, from <https://commonwealthbeacon.org/marijuana/could-massachusetts-become-a-cannabis-research-hub/>

Elser, Holly. (2023). State Cannabis Legalization and Psychosis-Related Health Care Utilization. Jama Network, 1-13.

Forbes Advisor. (n.d.). Best and worst states for healthcare. Retrieved November 18, 2024, from <https://www.forbes.com/advisor/health-insurance/best-worst-states-for-healthcare/>

Gruzca, R. A., Hur, M., Agrawal, A., Krauss, M. J., Plunk, A. D., Cavazos-Rehg, P. A., Chaloupka, F. J., & Bierut, L. J. (2015). A reexamination of medical marijuana policies in relation to suicide risk. Drug and Alcohol Dependence, 152, 68–72.

Hansen, Miller, & Weber. Washington State University. (2018). Behavioral Risk Factor Surveillance System (BRFSS) overview. Washington State University. <https://s3.wp.wsu.edu/uploads/sites/286/2020/01/BRFSS-Paper.pdf>

Hasan, Alkomiet. (2020). Cannabis use and psychosis: a review of reviews. <https://pubmed.ncbi.nlm.nih.gov/31563981/>

Investopedia. (2024). P-value. Retrieved December 11, 2024, from <https://www.investopedia.com/terms/p/p-value.asp>

National Library of Medicine. ArcView Market Research & New Frontier. (2014). Prevalence of Marijuana Use at College Entry and Risk Factors for Initiation During Freshman Year. Frontiers in Psychiatry, 5, 1–11. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4098711/>

Mar, Javier. (2023). Incidence of mental disorders in the general population aged 1–30 years disaggregated by gender and socioeconomic status. <https://pmc.ncbi.nlm.nih.gov/articles/PMC9872752/>

Marconi, A. (2016). Meta-analysis of the Association Between the Level of Cannabis Use and Risk of Psychosis. Schizophrenia Bulletin, 42(5), 1262–1269. <https://academic.oup.com/schizophreniabulletin/article/42/5/1262/2413827>

Marijuana Policy Project. (n.d.). Massachusetts marijuana laws. Retrieved November 18, 2024, from <https://www.mpp.org/states/massachusetts/>

Massachusetts Cannabis Control Commission. (2024). Ten things new patients should know. Retrieved December 11, 2024, from <https://masscannabiscontrol.com/new-patients/ten-things-new-patients-should-know/>

Miron, J. A., & Ekins, E. (2021). The budgetary effects of ending drug prohibition. Cato Institute. <https://www.cato.org/sites/cato.org/files/2021-01/PA908.pdf>

National Academies of Sciences, Engineering, and Medicine. The health effects of cannabis and cannabinoids: the current state of evidence and recommendations for research. Washington, DC: The National Academies Press; 2017. <https://nap.nationalacademies.org/catalog/24625/the-health-effects-of-cannabis-and-cannabinoids-the-current-state>.

Office of the Governor, Rhode Island. (n.d.). Governor McKee signs legislation legalizing and safely regulating recreational marijuana. Retrieved November 18, 2024, from <https://governor.ri.gov/press-releases/governor-mckee-signs-legislation-legalizing-and-safely-regulating-recreational>

Panchal, Nirmita. Kaiser Family Foundation. (2023). Racial and ethnic disparities in mental health care: Findings from the KFF survey of racism, discrimination, and health. <https://www.kff.org/racial-equity-and-health-policy/issue-brief/racial-and-ethnic-disparities-in-mental-health-care-findings-from-the-kff-survey-of-racism-discrimination-and-health/>

Pew Research Center. (2024, April 10). Facts about marijuana. Retrieved from <https://www.pewresearch.org/short-reads/2024/04/10/facts-about-marijuana/>

Pew Research Center. (2024, March 26). Americans’ views of marijuana legalization: Deep divisions on whether marijuana should be fully legal for medical and recreational use. Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2024/03/PP_2024.3.26_marijuana_REPORT.pdf

PewResearch Center. (2021, May 26). Religious Americans are less likely to endorse legal marijuana for recreational use. Pew Research Center. <https://www.pewresearch.org/short-reads/2021/05/26/religious-americans-are-less-likely-to-endorse-legal-marijuana-for-recreational-use/>

Rhode Island Department of Health. (n.d.). Medical marijuana patient requirements. Retrieved December 11, 2024, from <https://health.ri.gov/publications/requirements/MedicalMarijuanaPatients.pdf>

Seattle University. (n.d.). Suicide myths. Retrieved December 11, 2024, from <https://www.seattleu.edu/life-at-seattle-u/health-wellness/caps/suicide-and-crisis-information/suicide-myths/>

Secretary of the Commonwealth of Massachusetts. (n.d.). Registered voter enrollment statistics. Retrieved November 27, 2024, from <https://www.sec.state.ma.us/divisions/elections/research-and-statistics/registered-voter-enrollment.htm>

Singer, J. A., Rich, J. J., Capodilupo, R., & Schemenaur, M. (2020). Effect of cannabis liberalization on suicide and mental illness following recreational access: A state-level longitudinal analysis in the USA.

Rhode Island Department of State. (n.d.). Registered voter statistics. Retrieved November 27, 2024, from <https://datahub.sos.ri.gov/RegisteredVoter.aspx>

Ridley, M., Rao, G., Schilbach, F., & Patel, V. (2022). Poverty, depression, and anxiety: Causal evidence and mechanisms. Retrieved from <https://economics.mit.edu/sites/default/files/2022-09/poverty-depression-anxiety-science.pdf>

U.S. Census Bureau. (n.d.). Explore Census data. Retrieved November 18, 2024, from <https://www.census.gov/>

U.S. Department of Health and Human Services. (n.d.). About the Affordable Care Act. Retrieved November 27, 2024, from <https://www.hhs.gov/healthcare/about-the-aca/index.html>

U.S. Department of Health & Human Services. (n.d.). Affordable Care Act implementation FAQs (Set 18). <https://www.hhs.gov/guidance/document/affordable-care-act-implementation-faqs-set-18>

World Health Organization. (n.d.). Age-standardized suicide rates (per 100,000 population). Retrieved November 27, 2024, from <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/845>

Mental Health In Crisis: A Case-Study Analysis of Syria

Maryam Guerrab
Harvard College '25

The following paper implements a biosocial approach to answer 1) How has the ongoing conflict in Syria shaped the mental health crisis among internally displaced persons (IDPs)? and 2) What structural, political, and cultural barriers prevent effective mental health care delivery in low-resource and volatile settings? Using a framework which emphasizes the importance of social factors in shaping health outcomes, the study draws from reports from international organizations (e.g. WHO, UNHCR, and the IMC), academic literature, and statistical indices (e.g. Disability-Adjusted Life Years), to critique neoliberal ideologies, western-centered approaches to care, and bureaucratic constraints that limit impactful mental health intervention. The findings demonstrate how systemic failures—funding misallocation, lack of culturally adaptive tools, and disconnects between local and international actors—exacerbate psychological distress and perpetuate health inequities for Syrians, IDPs, and others in conflict zones around the world.

Introduction

Armed conflict destabilizes every dimension of public health, causing death, physical injury, and destroying vital health infrastructure (Murray et al., 2002). The most invisible toll of conflict is on mental well-being. Two billion people live in conflict-affected areas. The rates of those suffering from psychological distress or with a diagnosable mental illness such as schizophrenia, depression, or post-traumatic stress disorder (PTSD) are incredibly high (Archie, 2020; World Health Organization, 2022). However, 80% of people in conflict zones with mental disorders do not receive mental health services (Bayard et al., 2019). The disconnect between urgent need and inadequate response persists despite decades of global recognition by institutions and medical professionals on the importance of mental health care in humanitarian settings (Jenson, 1996).

In Syria, the mental health outcomes of internally displaced populations (IDPs) have only worsened since the onset of the country’s Civil War in 2011 (The Editors, 2023). While the war and subsequent displacement contributed to the onset of psychological harm, the continued deterioration of mental well-being is attributed to structural factors that limit access to effective care. I argue that the lack of effective mental health care provision to IDPs in Syria is due to entrenched ideas of scarcity, the dominance of Western mental health paradigms which neglect cultural context, and rigid bureaucratic practices that limit collaboration between international and local partners. These problems are not unique to Syria and can be used to explain broader patterns of mental health care delivery in conflict and low-resource settings.

To best understand the interdisciplinary factors contributing to the state of mental health in Syria and locations across the world, a biosocial approach is necessary. A biosocial framework assumes that health outcomes are inextricably linked to the social contexts they exist within. Human biology, the experience of illness, and medical interventions are shaped by social factors such as inequality, poverty, education, and housing. (Farmer, 2013). For instance, one cannot understand the spread of the Covid-19 pandemic in the United States only by analyzing the rate of infection or the strength of vaccine resistance for

different mutations of the disease. The impact that geography, poverty, education, and culture had on vaccination rates of different populations in the nation significantly shaped disease outcomes (Albrecht, 2022). Similarly, in Syria, mental health outcomes are not simply a result of trauma exposure, but also of global power dynamics rooted in capitalism.

This paper seeks to understand how social, political, and economic factors have interacted to create the current state of mental health in Syria and how structural barriers limit culturally responsive and sustainable mental health treatments in conflict zones. Through a case study of Syria, the paper hopes to uncover how global health care systems reproduce inequity and how these challenges can be addressed to advance mental health outcomes globally.

Context In Syria

After a series of pro-democracy protests erupted in 2011 demanding a change to the dictatorial rule of the Assad regime, the government responded violently by cracking down on protests, silencing the media, and torturing dissidents. In response, militia groups formed to oppose the government and a civil war began that persists until this day (The Editors, 2023). Although millions of Syrians have been forced to flee, 23 million currently still reside in the nation; 6.7 million people are internally displaced, and 14.6 million are in need of humanitarian assistance (Syria, 2023). The violence of the civil war, dire lack of resources, and the protracted nature of the conflict have contributed to abysmal mental health outcomes in this population.

Data from the United Nations High Commissioner for Refugees (UNHCR) and the International Medical Corps (IMC) show that 61% of Syrians have severe emotional disorders, with that figure growing over time (Hijazi & Weissbecker, 2023). A survey conducted by the Syrian American Medical Society (SAMS) found that 37.7% of those surveyed had been turned completely inactive because of significant distress sometime in the past two weeks of the survey date. 14.4% of respondents experienced such intense feelings of hopelessness they would rather not continue to live. In children between the ages of 8

and 15, 60% were expected to have one probable psychological disorder; 35.1% of children were expected to have PTSD, and a high proportion of them, depression and anxiety. Despite the protracted nature of the conflict in this country and the wealth of knowledge in the field of mental health treatment, there were no significant improvements in mental health outcomes for this population (Hamza & Hicks, 2021; Institute of Health, 2019).

One metric used to measure the impact of individual illness is Disability-Adjusted Life Years (DALYs). DALYs quantify the loss in health through both premature mortality and morbidity. In Syria, mental health disorders in the population contributed to a DALY increase from 1574.6 to 1621 from 2011 to 2019 (Institute of Health, 2019). Many might expect that declining mental health outcomes are to be expected in a population that has experienced war for so long. However, while this may be the case, it does not fully encompass the historically entrenched causes of abysmal mental health outcomes.

Throughout the literature, it has widely been identified that to effectively address mental health in conflict zones, providers should ideally be from the local community; the care that is given should be adaptable to the rapidly changing context; and the care should be community-driven, combining both scientific and traditional methods (Hamza & Hicks, 2021). However, this solution is not implemented. Reinforced norms of scarcity, a lack of culturally competent health care provision, and challenges arising from bureaucratic difficulties contribute to a disconnect between local and international organizations providing mental health care, leading to worse mental health outcomes.

The Beginning of Humanitarian Mental Health

Although occurring more than three decades apart and in vastly different contexts, the Spitak earthquake in Armenia and the Syrian Civil War reflect the impact that trauma and displacement have on mental health. The field of humanitarian mental health as we know it today was not established until the devastating aftermath of the 1988 earthquake in Armenia. By examining the psychological impact these events had, we can trace the evolution—and persistent shortcomings—of humanitarian mental health care delivery.

The Spitak earthquake, a 6.8 on the Richter scale, led to the death of 25,000 people, injury of more than 20,000, and displaced 360,000 (Kiureghian, 2025). Humanitarian organizations such as Mediciens Sans Frontiers (MSF) and Mediciens du Monde (MDM) responded to the natural disaster by providing the resources and staff necessary to treat the physical wounds of the victims. However, no organization considered addressing the psychiatric distress Armenians were experiencing, despite the tremendous psychological impact the natural disaster had. Researchers from UCLA found that 50% of those observed suffered from PTSD, 28% from depression, and 26% from anxiety. In response, Armenians from across the globe flocked to their homeland to help their kin through this time of hardship by connecting with the survivors of the earthquake and leveraging their shared history and culture. This was the first time an official evaluation of trauma neurosis and PTSD was administered in response to a humanitarian crisis (Fassin et al., 2011).

Over the following years, the vocabulary and tools to treat

the mental health of those in humanitarian zones developed; when the Yugoslav Wars began in Europe two years later, the implementation of mental health programs was a priority (Fassin et al., 2011). In 1994, the first World Health Organization Regional Model on Mental Health was implemented in Sarajevo. This model, a coordinated set of mental health projects meant to serve certain geographic areas and populations of 400,000, emphasized 1) the collaboration between intergovernmental, non-governmental, and national mental health organizations, 2) establishing a regional coordination center 3) training of local professions and 4) promoting a community-oriented mental health approach—objectives still agreed upon by researchers to be effective to address psychological trauma (Jenson, 1996). By 1995, nearly 200 psychosocial projects were in progress in Bosnia and Herzegovina and Croatia, seeking to address the widespread psychological impacts of trauma. A study conducted on the war-afflicted populations in Croatia found that more than 30% of inhabitants accessed specialized mental health services, with evidence suggesting that general practitioners also served as mental health support for much more of the population. The persistence of mental health disorders in this population for fifteen years led the researchers to conclude that while psychological support is essential, other forms of informal, community-based social support are also necessary (Francisković et al, 2008).

In Syria, much like in Armenia and Yugoslavia, the impact of trauma is not merely physical, but deeply psychological. However, in the case of Armenia and Yugoslavia, mental health interventions in humanitarian contexts were only beginning to be developed, a potential explanation for the shortcomings of effective treatments; in Syria, despite the growing recognition and evolution of culturally sensitive and trauma-informed mental health tools, humanitarian systems have failed to effectively integrate into emergency responses. Understanding this historical pattern is critical to addressing the ongoing mental health crisis occurring globally.

Mental Health Apportionment in Syria

Currently, mental health services in Syria are apportioned by both INGOs and local health care organizations; however, there is little collaboration between the two entities, leading to ineffective health provision (Hamza & Hicks, 2021; Hijazi & Weissbecker, 2023). International agencies, led by the International Medical Corps (IMC) and UNHCR, provide psychological first aid services, basic counseling for individuals, groups, and families, pharmacological and non-pharmacological management of mental disorders, and limited psychotherapy services. A study from the IMC found that in the areas IMC and other partners operated, 60.1% of women and 69% of children were receiving mental health treatments. However, despite the INGOs’ presence since the start of the war, mental health outcomes have remained stagnant or worsened. The IMC has identified that challenges in mental health care provision stem from lack of funding, instability, and government restrictions. Local NGOs in the area face unique challenges as well (Hijazi & Weissbecker, 2023).

Research conducted on mental health apportionment describes the difficulties experienced by organizations and networks such as SAMS who work locally in the area (Hamza

& Hicks, 2021). Due to limited mental health staff before the onset of the war, it was impossible for small organizations to establish sustainable mental health programs in the midst of conflict (Assalman et al., 2008; Hamza & Hicks, 2021). This reality led to the formation of significant barriers to overcome at the start of the war that continue to impede these organizations’ capacity to address the mental health need in Syria.

INGOs and local organizations contribute to Syrian mental health provision in different ways. From the perspective of local agencies, large, institutionalized humanitarian organizations are more focused on treating physical injuries that are immediately life-threatening rather than mental health concerns. Even when mental health concerns are addressed by these organizations, the intervention is ineffective because of the lack of investment into appropriate culturally relevant training. While local organizations are more equipped to provide culturally competent treatment, they do not have the same reach as large INGOs because of limited resources (Hamza & Hicks, 2021). While the scarcity of resources experienced by local organizations is understandable, the scarcity mindset adopted by large INGOs who have millions, if not billions, in their budget, is not.

Scarcity

The rise of neoliberalist ideals, an economic ideology characterized by the privatization of social services to counteract socialist and communist models in the post-Soviet era, in the 1980s fundamentally altered the global health agenda by international NGOs (INGOs). Powerful international agencies, such as the World Health Organization (WHO), recognized the inability of low- and middle-income countries to provide necessary health care for their citizens and began to shift the responsibility of financially securing health care to individuals. These agencies believed that encouraging individuals to cover their personal medical expenses would encourage ideals of a democratic economy based on the tenets of capitalism. This system was deeply rooted in and reflected colonialist ideas—incredibly wealthy former colonial countries who control these agencies would rather not finance treatments for the expensive diseases afflicting those in poorer nations. There would be a negligible return of investment; investing in non-Western lives was simply not worth it. The revolutionary changes to the field of global health allowed neoliberalist ideas to permeate the agendas of INGOs, contributing to the stagnation or worsening health outcomes across the globe (Keshavjee, 2014).

Constructed ideas of scarcity—resource constraints not attributed to a lack of materials but a consequence of political and economic decisions—present within large INGOs are used to justify their failures in adequately addressing the mental health challenges (Schrecker, 2013). By claiming there are insufficient funds and resources, organizations like the WHO, the UNHCR, and the IMC are without fault when significant areas of humanitarian need are not effectively addressed (Hijazi & Weissbecker, 2023). This lack of funding has contributed to the increase in DALYs for mental health illnesses in Syria despite the passage of more than a decade since the start of the conflict, time when the existing knowledge to effectively address mental health could have been leveraged to develop impactful programs (The Editors, 2023; Institute of Health, 2019)

Funding and Inappropriate Funding Priorities

INGOs such as the IMC and the UNHCR provide substantial funding to Mental Health and Psychosocial Support (MHPSS) programs in Syria; however, the allocation of this funding does not align with the needs of the population. Although it was difficult to acquire financial data on a country-level scale, financial data on the regional scale provides key insights into the situation. The IMC allocates 28% (\$40.7 million) of its operating budget to programs in the Middle East (Financial Reports, 2022). The UNHCR dedicates 36% (1.638 billion) of its regional operating budget to the Middle East and North Africa, more than any other area (Funding UNHCR, 2023). Despite this significant investment, the inadequate implementation of community-driven, culturally informed, bottom-up approaches have yielded limited improvements to mental health outcomes (Hamza & Hicks, 2021; Institute of Health, 2019).

INGOs have the financial resources and infrastructure to implement extensive, wide-reaching and sustainable interventions. In areas where the IMC operates, the plurality of women and children use mental health resources the organization provides. These organizations also have the capacity to build extensive resource networks, train individuals, assess programs, and garner international attention. However, these organizations adopt top-down, western-centric approaches that neglect local values and valuable non-clinical local resources which are known to positively impact mental well-being.

Instead of using their budgets to empower local actors in providing culturally relevant care, INGOs train mental health professionals and implement programs based on the WHO Mental Health Gap Action Programme (mhGAP) Guidelines for mental, neurological and substance use disorders (Hijazi & Weissbecker, 2023). For the most part, these guidelines only suggest that variations in cultural context should be considered. The most extensive references to cultural sensitivity are for dementia. Here, the guidelines highlight prior research which indicates that interventions that are adapted for cultural context, rather than being developed with the target culture in mind, are less effective. Furthermore, regarding the treatment for drug dependence, the guideline highlights that spiritually oriented interventions and peer support are culturally sensitive and reduce treatment stigma. Nowhere in the guidelines does it provide any instruction for how a culturally sensitive approach should be implemented (Mental Health, 2025). An assessment of the interventions enacted by INGOs reveals that these guidelines contribute to programmatic rigidity, and most “culturally sensitive adaptations” simply constitute language translation. A bottom-up approach to mental health care provision would enhance the efficacy and suitability of mental health interventions and research (Hamza & Hicks, 2021).

The progress that can be made toward improving Syrian mental health is being impeded by the inadequate use of funds. Currently, the IMC working group does not facilitate (1) communal spaces for members to organize and provide support to each other, (2) traditional, spiritual, or religious supports, including communal healing practices, (3) community worker mental health referrals and follow-ups, (4) teacher- and school-based psychosocial support, (5) non-pharmacological care, or (6) research (Hijazi & Weissbecker, 2023). An analysis of a book published by a Syrian-American medical neuropsychologist

who worked in Syria for nine years delivering clinical care and developing medical programs in active combat zones and besieged areas to IDPs and refugees, has denoted the importance for care to be community-driven, drawn from shared cultural, religious, and traditional experiences, and provided continuously (Hamza & Hicks, 2021).

By not funding interventions that prioritize the cultural and communal foundations upon which effective mental health provision in Syria must be built, the rampant trauma the population has and continues to experience is not being addressed. Although these organizations possess positive intentions to reduce harm, gaps in care due to a scarcity mindset stemming from neoliberalism ideologies comes at the expense of Syrian mental health and reinforces systems of structural violence.

Lack of Culturally Appropriate Care & Culturally Adaptive Tools

Although large INGOs profess to strive to incorporate local context in their work and engage closely with community partners, significant strides need to be made until both local and international entities work cohesively together. INGOs concede that the burden of mental health care falls primarily on their shoulders, with little to no mention of smaller local nonprofits providing similar services (Hijazi & Weissbecker, 2023). However, the lack of cooperation between local partners and large INGOs is concerning and impedes the progress of psychiatric outcomes. While large INGOs have the ability to administer care to more people, organizations operating locally are more effective in the culturally competent care they provide (Hamza & Hicks, 2021).

Overcoming the stigma surrounding mental health in Syria is a significant issue INGOs must overcome. Culturally, people with mental illness are considered unstable and weak; individuals who other members of society seek to distance themselves from (Badawi et al., 2022). Men are taught to suppress the outward expression of emotion and women are fearful of discussing mental health issues for fear of being deemed unsuitable for marriage (Hassan et al., 2015; Badawi et al., 2022). Across religious traditions, faith serves as an important source of resilience (Hamza & Hicks, 2021). Islamic teachings, the most common religion in Syria, profess that those who surrender to God accept their fate as the will of God; their hardship is an opportunity to strengthen their devotion to Him (Hassan et al., 2015). One narration describes a woman who lost five of her children, lived in a refugee camp, and worked for one dollar a day. When asked how she was doing, she had a peaceful look on her face and praised God (Hamza & Hicks, 2021), reflecting her belief that this world is transient and only a test from God (Hassan et al., 2015). Struggling to progress past trauma may reflect a form of spiritual and moral weakness, leading to both self and community stigmatization. These cultural and religious nuances are better understood by fellow community members than by Western physicians (Hamza & Hicks, 2021).

SAMS stresses the importance of training non-professionals and refugee peers to address the mental health concerns in Syria. There is greater trust between the patient and provider when care comes from a community-member instead of an implanted “Western Doctor.” Feelings of mistrust have been cultivated in the local population since Syrians feel INGOs do

not consult the community on its needs. A bottom-up approach to mental health care provision would be more effective at addressing patient concerns and better received (Hamza & Hicks, 2021).

Local organizations staffed by community members are better able to recognize unique ways mental health challenges present themselves. The usual symptoms that a Western practitioner would expect to see in various mental health conditions may be significantly different in Syria where physical complaints, such as “I’m so tired,” may reflect deeper psychological issues (Hamza & Hicks, 2021). For instance, the top five expressions of PTSD among people from different backgrounds vary widely. The most common symptoms of PTSD in order of frequency for individuals from North America, Europe, and Australia are negative emotional state, detachment from others, fear, lack of social support, alienation/isolation, and intrusive distressing memories. For those from the Middle East and North Africa, the most common symptoms are negative emotional state, fear, alienation/isolation, detachment from others, inability to experience positive emotions, and a sense of loss (Michalopoulos & Meinhart, 2018). As is evident, there are both similarities and differences between expressions of trauma across cultures and local providers are better equipped to recognize these unique expressions (Hamza & Hicks, 2021).

Cultural differences also impact which practices are best when providing care. A study found Syrians were willing to engage in a discussion about their concerns not when a mental health professional claimed that they wanted to “support” them or to provide them with therapy, but when in the midst of a conversation, a provider asked about the “challenges” they were facing in their lives. This is reflective of the fact that Syrians have limited prior exposure to mental health care and may consider seeking mental health treatment both stigmatizing and incomprehensible (Hamza & Hicks, 2021). Further, as a deeply religious society, Syrian populations find strength and resilience through a higher power. This therapeutic practice is not provided by Western INGOs, reflecting a rigidity in treatments rooted in Western methodology that does not effectively leverage cultural knowledge to provide the most appropriate care, leading to worse outcomes (Hamza & Hicks, 2021; Hijazi & Weissbecker, 2023).

Lack of Accurate Data

In addition to the fact that transplanting western models of care leads to inadequate mental health provision in Syria, the centrality of the western perspective, medical terminology, and tools also leads to inadequate research results for diverse populations.

First, few sources of data detail the state of mental health in Syria. From the data available, the integrity of select findings are questionable. For example, in 2019, the Institute of Health Metrics and Evaluation published data on the estimated number of DALYs due to different categories of mental illness. It is evident that the data was centered around Western nations. For all countries, the only mental disorders evaluated are depressive disorders, anxiety disorders, bipolar disorder, schizophrenia, or eating disorders. While these are the most prevalent conditions prevalent in western, industrialized nations, this is not the case for nations experiencing war and conflict, as Syria is

(Michalopoulos & Meinhart, 2018).

Second, there is a significant issue in terms of how the DALYs are assessed. While the DALY is the internationally accepted measure of the impact of mental health concerns, it is important to interrogate this metric. As is argued in Reimagining Global Health: An Introduction, the DALY was defined based on western research methodologies (Farmer et al., 2013). Thus, the use of it as a basis of evaluating the impact of mental health in a nation must be questioned. According to the data, the measured burden of disease in Syria for the measured mental disorders (1621 DALYs per 100,000) is smaller than that of Australia (2062 DALYs per 100,000), the United States (1796 DALYs per 100,000), and the United Kingdom (1625.4 DALYs per 100,000) (Institute of Health, 2019). Considering Syrians are actively engaged in a civil war and the Australian, American, and British populations are not currently experiencing the same death, destruction, and persecution, the integrity of the DALY must be questioned. These metrics do little to account for the full scope of suffering and how the suffering of an individual has peripheral impacts on the larger community (Farmer et al., 2013).

One reason the data presented by traditionally reliable sources is not accurate in the Syrian context is because of a lack of cultural competence. For instance, foreign mental health professionals use non-rigorously translated English-language surveys (Hamza & Hicks, 2021). The intricacies of mental health experiences cannot be assessed effectively using something such as Google Translate, since some concepts may not exist or may be expressed differently in the Arabic language and Syrian dialect.

Further, as seen by the programs and initiatives provided for by the INGOs, research is not prioritized (Hijazi & Weissbecker, 2023). Without the proper research infrastructure, there cannot be a robust evaluation of mental health needs or the effectiveness of interventions. This limits the ability of INGOs to adapt treatments or advocate for programs with no evidence to justify the use of funds, contributing to the default implementation of Western models of care which are ineffective and culturally misaligned.

In fact, the WHO, whose work centers around research and leading innovation in the field of global health, only spends 1.14% of its budget in the Eastern Mediterranean region (Mazumdar, 2020). This incredibly limited investment at least partially explains why there exists a significant gap in research on Syria. This creates a vacuum where the state of health cannot be adequately assessed, nor can methods be developed to improve it.

Bureaucracy

The impact that large humanitarian organizations have is constrained by the rigid adherence to bureaucratic practices that restrict effective service delivery (InterAgency, 2022). One of the most pressing examples is the reliance of INGOs on coordinating aid provision with the national government; however, the national government is the entity primarily responsible for enacting violence against the Syrian people (The Editors, 2023; Kayyali, 2019). For instance, the WHO regularly works with national health ministries to implement programs, operating under the assumption that national programming will reach the

most people in this way because of the government’s ability to resource, distribute, and plan projects on a large scale (World Health Organization, 2023). This strategy is not effective in Syria—through legitimizing relations with the government responsible for the trauma they are trying to treat, INGOs are undermining their legitimacy with the Syrian people. Further, individuals in need of assistance in rebel-controlled or besieged territories outside of Damascus cannot access any aid because the state refuses to operate in these locations (Hijazi & Weissbecker, 2023).

Bureaucratic restrictions also restrain the services INGOs provide because of stringent approval processes dictating the types of programs offered, the number of staff, and the locations which organizations are allowed to work in (Hijazi & Weissbecker, 2023; Kayyali, 2019). If unsatisfied with the project proposal or the organization, the government arbitrarily refuses to grant INGOs permission to operate. Thus, INGOs succumb to the requirements of the government and implement programs that are based on the government’s desires and not on an unbiased appraisal of the population’s needs (Kayyali, 2019). The programming that is enacted neglects vulnerable populations in the north of the country, an area not under government control (Hijazi & Weissbecker, 2023). However, to explain the lack of health care provision in these areas, the IMC has cited “security concerns” and “limited capacity” (Hijazi & Weissbecker, 2023). These explanations mask the deeper issue of government interference which restricts INGOs ability to serve those most in need.

In Syria, INGOs can only operate if they partner with government-approved, humanitarian actors. These actors include human rights abusers who are not vetted properly by INGOs. In one instance, a local organization founded by a member of the National Defense Forces, information the UN was aware of, was paid by a UN agency to enact a project. Months later, it was revealed the organization used the funds for other purposes and was forging their reports to the UN. Even when projects are implemented by local, pre-approved actors, information on the program’s beneficiaries is accessible by security forces, a breach of patient confidentiality that threatens the lives of those who seek medical care (Kayyali, 2019).

Local, independent organizations are better suited to address the gap in service provision created by this system. Organizations like SAMS and SEMA (Syrian Expatriates Medical Association) can flexibly adapt to changing circumstances and are able to provide mental health interventions in besieged areas (Hamza & Hicks, 2021). Yet global bureaucratic processes prioritize funding the work of larger, more recognized INGOs, sidelining and delegitimizing the work of smaller organizations. Despite filling a critical service gap, organizations like SAMS are not given the money, resources, or support they need to expand operations and maximize their impact (Hamza & Hicks, 2021; Hijazi & Weissbecker, 2023).

These various forms of bureaucratic rigidity, including (1) policies which force INGO collaboration with the government, (2) policies that dictate which programs are implemented and locations receive care, (3) regulations that force INGOs to collaborate with pre-approved organizations, and (4) global administrative biases against independent, local organizations create an inequitable health care delivery system in Syria that

undermine the values of neutrality and equality humanitarian organizations claim to uphold (HRW).

Recommendations

To address the systemic barriers to effective Syrian mental health care provision highlighted in this paper, I propose the following recommendations:

1) Fund programs which prioritize community-based, culturally competent care

From the inception of humanitarian mental health, global leaders have identified the importance of culturally relevant interventions. From the coordinated mental health approach used in Croatia during the Yugoslav wars to the WHO mhGAP guidelines, culturally relevant interventions are necessary to effectively treat mental health issues in Syria. However, INGOs are failing to provide traditional, religious, and spiritual support, communal healing spaces, school-based psychosocial support, and mental health follow-up services. In light of this, I advocate for:

a) Fund community-led initiatives:

Interventions that are created with the target-population in mind, as opposed to being adapted for cultural-relevance, are more effective. INGOs and local organizations should involve Syrian communities and front-line mental health providers to 1) design mental health programs and 2) determine how budgets are spent. Including the input of Syrians in program development and implementation is integral to creating effective services and will allow service providers to better understand unique expressions of trauma within the population.

b) Diversify Funding Recipients:

Local organizations such as SAMS may have less institutional capacity than INGOs such as the IMC, however, they are able to reach often-neglected populations in volatile areas and have greater community trust. Further, local organizations engage with non-clinical interventions and invest in training trusted community members, often with limited health background, to provide support services to those in the local area. It is therefore necessary to reorient global funding schemes to provide significant resources to these local organizations, highlighting their role in filling in critical gaps in care.

c) Incorporate Spiritual Practices in Care Models:

Although Western mental health frameworks do not emphasize spiritual and religiously informed practices, faith is central to the lives of many Syrians, serving as a source of support and resilience. Thus, it is crucial to 1) develop workshops for implanted physicians which incorporate both scientific and religious practices, 2) train non-professional community members (who are already versed in the spiritual tradition) in providing psychosocial support, and 3) create treatment programs which leverage the knowledge and cultural background of the two aforementioned stakeholders to provide informed and comprehensive care.

d) Address Research Gaps:

With little investment in mental health research in the

region, it is difficult to develop evidence-informed approaches, measure programmatic efficacy, or adapt treatments effectively and responsibly. The established metric, DALYs, were developed with a western audience in mind, and are ineffective in properly assessing the extent of the mental health burden in Syria. Increases in funding to create, and not simply adapt, measurement tools with the intended population in mind are imperative.

2) Increase Collaborative Efforts Between INGOs and Local Partners

To provide the most effective care, the resources from INGOs must be coordinated with local organizations such as SAMS to provide care that is culturally relevant, well-received, and impactful. Although the Assad regime is no longer in power, the future of Syria remains uncertain (Abdulrahim, 2025). In this volatile time, commitments to organizational change can be most effective.

a) INGOs must conduct independent needs assessments and program implementation:

Humanitarians who have worked in the crisis in Sudan, the Congo, and Yemen have described the unique hardships INGOs face to have a substantial impact in Syria due to the government’s restrictions (Kayyali, 2019). To improve mental health outcomes, INGOs must negotiate with any future government to conduct independent needs assessments and rigorous investigations on any local partners, as well as retain the ability to choose which partners to engage with, which programs to implement, the receiving population, and the dedication to protecting patient information. Dictatorial regimes require the presence of INGOs to provide basic services, especially in conflict settings (Heiss, 2017). INGOs must leverage their importance to negotiate terms which prioritize the health of all Syrians.

b) Establish Clear Coordination of Care Models:

Coordination of Care Models, a structured approach to ensure effective communication and collaboration among healthcare providers and patients to facilitate seamless communication and improve patient models, are integral between INGOs and local organizations (Duan-Porter, 2022). Considering the finite aid Syria receives, this model allows the various organizations in the nation to 1) engage in mutual learning, 2) coordinate which services they are best equipped to provide, and 3) limit extraneous resource use which provide services to some and neglect others.

3) Advocate for International Change on the Structural Level

The broad-sweeping changes required to restructure entire care systems cannot occur without a fundamental shift in global consciousness. Localized interventions are helpful, and improvements to the global systems that have perpetuated inequities in the apportionment of Syrian mental health care are necessary. In order to address constructed ideas of scarcity, the marginalization of non-Western and culturally impactful interventions, and bureaucratic exclusion, the following is required:

a) Interrogate Aid Models Rooted in Neoliberalism:

INGO care models prioritize cost-efficiency, individual responsibility, and short-term impact. An operational agenda rooted in neoliberalism has allowed for the permissibility of stagnant or worsening mental health outcomes in Syria under the guise of “unsustainability” or “financial constraints.” Deconstructing deeply rooted neoliberalism ideologies is needed to increase financial investment into the nation to implement long-term, culturally informed care.

b) Develop Interventions Centering Biosocial Frameworks:

Viewing the mental health burden Syrians experience only as a health issue resulting from the trauma of war is impeding meaningful outcome improvements. Legitimizing the impact that power dynamics and state violence, economic precarity and a lack of agency, social stigma, and other factors have can better inform interventions. By better understanding how to develop relevant measurement metrics, include cultural practices in treatment, and leverage religious resilience, care in the Syrian context and beyond can drastically improve. Adopting biosocial frameworks in the creation of interventions is necessary to address the social and medical roots of illness.

An Example from Rwanda

A successful community-based mental health intervention developed in Rwanda to address the trauma from the Rwandan genocide lends support to the aforementioned recommendations. In 1994, a harrowing genocide against the Tutsis occurred in Rwanda, completely upending the nation’s social structure. Genocide survivors experienced nearly three times the rate of depressive disorders (35% v. 12%) and eight times the rate of PTSD (27.5% v. 3.6%) as compared to the general population in addition to feelings of fear and mistrust, isolation, and collective shame. Similar to the case of Syria, limited funding, a lack of adequately trained providers, and stigma impeded the mental health outcomes in Rwanda. In response, the Ubuntu Center for Peace (UCP) established a Community-Based Social Healing Model (CBSH) to address both the trauma from the war and continued social stressors the community faced.

The CBSH model acknowledges that western frameworks of mental health treatment largely overlook “communal, sociocultural, and historical factors” integral to addressing Rwandan mental health. The comprehensive model included three interventions: 1) Breath-Body-Mind (BBM) is a practice of breathing techniques known to shift state’s of fear and anger to one of safety and connectedness, shown to be effective in other conflict zones in South Sudan and Uganda; 2) creating communal safe spaces for narrative sharing and active listening rooted in Rwandan cultural rituals (e.g. singing, dancing, and drumming); 3) at the end of the intervention, empowering participants with the leadership skills necessary for independent socioeconomic activities, allowing them to assume agency over their own healing in a long-term and community-rooted solution. These interventions were implemented in small groups (approximately 20 participants) in 54 different rural communities. The implementation was led by Community Healing Assistants, community leaders who were provided with six weeks of training and supervision by trained psychologists. The program’s design—led by community members in

partnership with trained medical professions, leveraging cultural practices in community settings, and providing long-term interventions empowering participants—represents one example of a biosocial model that can have significant impact in addressing the mental health needs in Syria and beyond (Jansen et al., 2024). At the time of publication, the study was still ongoing; however, analysis of a community-based pilot program previously run by the UCP significantly decreased the prevalence of depression, anxiety, and PTSD (Ubuntu, 2023). In Syria, if INGOs were to work with local organizations and effectively implement a CBSH model, similar progress may be achieved in Syria.

Structural Violence and Concluding Remarks

Enforced ideas of scarcity, the lack of culturally adaptive care, and bureaucratic rigidity that limits the scope of health care provision—create an environment of structural violence; Although INGOs provide psychological first aid, basic counseling, and various forms of pharmacological and nonpharmacological treatments and psychotherapy services, Syrians are still denied access to quality mental health care because of societal factors, leading to disastrous consequences for their well-being (Farmer, 2010; Hijazi & Weissbecker, 2023). These figures paint a picture of suffering which can be curtailed if the known methods of treatment were properly implemented in Syria, but this is not the case. Although the geopolitical context in Syria has shifted, the same mental health concerns remain.

As displayed in this paper, to better understand the current inequities that exist in Syria, a biosocial approach is necessary. On the forefront, it might seem that the present mental health challenges experienced by the population are attributable to the ongoing trauma of war. However, it is impossible to deny the significant, if not primary role, that other societal, political, economic, historical, and cultural factors play.

Through understanding the impact that ideas of scarcity, western-centric models of care which display a lack of cultural competency, and bureaucratic rigidity have, one can understand the barriers that exist to providing adequate mental health care to the Syrian population impacted by war. This understanding is integral in addressing the mental health outcomes in similar conflict zones throughout the world—Gaza, Sudan, Ukraine, Congo, Haiti, and dozens of other nations worldwide (Council, 2025). Only through understanding the issues at hand, can strides be made towards reducing the mental health burden experienced by this population and the needs of Syrians addressed in a way that is most supportive to their emotional betterment.

References

Abdulrahim, R. (2025, March 9). Syria’s interim president calls for unity amid fresh fighting. *The New York Times*. <https://www.nytimes.com/2025/03/09/world/middleeast/syria-latakia-clashes.html>

Archie, A. (2022, March 31). World is seeing the greatest number of conflicts since the end of WWII, U.N. says. *NPR*. <https://www.npr.org/2022/03/31/1089884798/united-nations-conflict-covid-19-ukraine-myanmar-sudan-syria-yemen>

Albrecht D. E. (2022). COVID-19 in Rural America: Impacts of Politics and Disadvantage. *Rural sociology*, 87(1), 94–118. <https://doi.org/10.1111/ruso.12404>

Assalman, I., Alkhalil, M., & Curtice, M. (2008). Mental health in the Syrian Arab

Republic. *International psychiatry : bulletin of the Board of International Affairs of the Royal College of Psychiatrists*, 5(3), 64–66.

Bawadi, H., Al-Hamdan, Z., Khader, Y., & Aldalaykeh, M. (2022). Barriers to the use of mental health services by Syrian refugees in Jordan: A qualitative study. *Eastern Mediterranean Health Journal*, 28(3), 197–203. <https://doi.org/10.26719/emhj.22.030>

Council on Foreign Relations. (2025). Global Conflict Tracker. *Council on Foreign Relations*. <https://www.cfr.org/global-conflict-tracker>

Duan-Porter W, Ullman K, Majeski B, et al. (2020) Care Coordination Models and Tools: A Systematic Review and Key Informant Interviews. *Department of Veterans Affairs (US)*; <https://www.ncbi.nlm.nih.gov/books/NBK566155/>

The Editors of Encyclopaedia Britannica. (2023, December 8). Syrian Civil War. *Encyclopædia Britannica*. <https://www.britannica.com/event/Syrian-Civil-War/Civil-war>

Farmer, Paul. 2010 Partner to the Poor. *Berkeley: University of California Press*. Chapter 1: On Suffering and Structural Violence

Farmer, P., Kleinman, A., Kim, J., & Basílico, M. (Eds.). 2013. Reimagining Global Health: An Introduction. Berkeley: University of California Press.

Fassin, D., & Rechtman, R. (2011). Humanitarian Psychiatry. In *The Empire of Trauma: An Inquiry into the condition of victimhood* (pp. 163–188). Essay, W. Ross MacDonald School Resource Services Library.

Financial Reports. International Medical Corps. (2022). <https://international-medicalcorps.org/who-we-are/accountability-financials/financial-reports/>

Franciskovic, T., Tovilovic, Z., Sukovic, Z., Stevanovic, A., Ajdukovic, D., Kraljevic, R., Bogic, M., & Priebe, S. (2008). Health care and community-based interventions for war-traumatized people in Croatia: community-based study of service use and mental health. *Croatian medical journal*, 49(4), 483–490. <https://doi.org/10.3325/cmj.2008.4.483>

Funding UNHCR’s programmes. *UNHCR*. (2023). <https://reporting.unhcr.org/global-appeal-2024/funding-unhcrs-programmes>

Hamza, M. K., & Hicks, M. H. (2021). Implementation of mental health services in conflict and post-conflict zones: Lessons from Syria. *Avicenna journal of medicine*, 11(1), 8–14. https://doi-org.ezp-prod1.hul.harvard.edu/10.4103/ajm.ajm_141_20

Hassan, G, Kirmayer, LJ, MekkiBerrada A., Quosh, C., el Chammay, R., Deville-Stoetzel, J.B., Youssef, A., Jefee-Bahloul, H., Barkeel-Oteo, A., Coutts, A., Song, S. & Ventevogel, P. (2015) Culture, Context and the Mental Health and Psychosocial Wellbeing of Syrians: A Review for Mental Health and Psychosocial Support staff Working with Syrians Affected by Armed Conflict. *UNHCR*. <https://www.escap.eu/uploads/Refugee/13-mental-health-syria-unhcr.pdf>

Heiss, Andrew (2017). Amicable Contempt: The Strategic Balance between Dictators and International NGOs. *Dissertation, Duke University*. Retrieved from <https://hdl.handle.net/10161/16313>.

Hijazi, Z., & Weissbecker, I. (2023). Syria Crisis: Addressing Regional Mental Health Needs and Gaps in the Context of the Syria Context. *International Medical Corps*. <https://internationalmedicalcorps.org/wp-content/uploads/2017/07/Syria-Crisis-Addressing-Mental-Health.pdf>

Institute of Health Metrics and Evaluation. (2019). Burden of Disease *From Each Category of Mental Illness*. Retrieved October 25, 2023.

InterAgency Standing Committee. (2022). UNDERSTANDING AND ADDRESSING BUREAUCRATIC AND ADMINISTRATIVE IMPEDIMENTS TO HUMANITARIAN ACTION: FRAMEWORK FOR A SYSTEM-WIDE APPROACH. https://interagencystandingcommittee.org/sites/default/files/migrated/2022-01/IASC%20Guidance%20Understanding%20and%20Addressing%20Bureaucratic%20and%20Administrative%20Impediments%20to%20Humanitarian%20Action_Framework%20for%20a%20System-wide%20Approach.pdf

Jansen, S., Niyonzima, J. B., Gerbarg, P., Brown, R. P., Nsengiyumva, A., Niyonsenga, J., & Nsabimana, E. (2024). Evaluating effects of community-based

social healing model on ubuntu, mental health and psychosocial functioning in post-genocide Rwanda: Protocol for Cluster Randomized Control Trial. *Trials*, 25(1). <https://doi.org/10.1186/s13063-024-08632-6>

Jenson, S. B. B. (1996). . Mental health under war conditions during the 1991-1995 war in the former Yugoslavia. *World Health Organization*. Retrieved 2025, from https://iris.who.int/bitstream/handle/10665/54528/WHSQ_1996_49_3-4_p213-217_eng.pdf?sequence=1.

Kakaje, A., Zohbi, R. A., Aldeen, O. H., Makki, L., Alyousbashi, A., & Alhaffar, M. B. A. (2021, January 2). Mental disorder and PTSD in Syria during wartime: A nationwide crisis - BMC psychiatry. *BioMed Central Psychiatry*. <https://bmcpsy psychiatry.biomedcentral.com/articles/10.1186/s12888-020-03002-3>

Kayyali, S. (2019). Rigging the system: Government policies co-opt aid and reconstruction funding in Syria. *Human Rights Watch*.

Keshavjee, S. (2014). Blind spot: How neoliberalism infiltrated global health. *Univ. of California Press*.

Keshavjee, S., & Farmer, P. E. (2012). Tuberculosis, Drug Resistance, and the History of Modern Medicine . *New England Journal of Medicine*. <https://www.nejm.org/doi/full/10.1056/NEJMra1205429>

Kiureghian, A. (2025, February 10). On the Edge of Life and Death: 18 hours Under the Ruins - Surviving the December 7, 1988 Magnitude 6.8 Earthquake in Northern Armenia. *UC Berkeley*. <https://peer.berkeley.edu/news/edge-life-and-death-18-hours-under-ruins-surviving-december-7-1988-magnitude-68-earthquake#>

Kleinman, Arthur. 2010 Four Social Theories for Global Health. *Lancet* 375(9725):1518-9.

Mazumdar, S. (2020, April 20). World Health Organization: What does it spend its money on?. *The Conversation*. <https://theconversation.com/world-health-organization-what-does-it-spend-its-money-on-136544>

Mental health gap action programme (mhgap) guideline for mental, neurological and substance use disorders. (2023). *World Health Organization*. 2025, <https://iris.who.int/bitstream/handle/10665/374250/9789240084278-eng.pdf?sequence=1>

Michalopoulos, L. M., & Meinhart, M. (2018, April 26). Global Posttrauma Symptoms: A Systematic Review of Qualitative Literature. *Sage Journals*. <https://journals-sagepub-com.ezp-prod1.hul.harvard.edu/doi/full/10.1177/1524838018772293>

Murray, C. J., King, G., Lopez, A. D., Tomijima, N., & Krug, E. G. (2002). Armed conflict as a public health problem. *BMJ (Clinical research ed.)*, 324(7333), 346–349. <https://doi.org/10.1136/bmj.324.7333.346>

Roberts, B., & Fuhr, D. (2019, October). Scaling up mental health interventions in conflict zones. *The Lancet Public Health*. [https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667\(19\)30179-3/fulltext](https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(19)30179-3/fulltext)

Schrecker T. (2013). Interrogating scarcity: how to think about ‘resource-scarce settings’. *Health policy and planning*, 28(4), 400–409. <https://doi-org.ezp-prod1.hul.harvard.edu/10.1093/heapol/czs071>

Syria. Central Intelligence Agency. (2023, December 6). <https://www.cia.gov/the-world-factbook/countries/syria/>

Ubuntu Center for Peace. (2023). (rep.). Program Outcomes Report. Retrieved 2025, from <https://static1.squarespace.com/static/64a0fc5548d32c04ff171621/t/6581c7f3a9d40952f83944c5/1703004194497/Program+Impact+Narrative+2022-23>.

World Health Organization. (2022, March 16). Mental health in emergencies. *World Health Organization*. <https://www.who.int/news-room/fact-sheets/detail/mental-health-in-emergencies>

World Health Organization. (2023). WHO In Syria: Mental health. *World Health Organization*. <https://www.emro.who.int/syria/priority-areas/mental-health>

EXPLAIN THIS, PRUNER!

The Effect of Zero-Order Pruning on LLM Explainability and Curvature

Joseph Bejjani, Camilo Brown-Pinilla, David Ettel
Harvard College ‘26

Large Language Models (LLMs) excel in language understanding and generation tasks but have significant memory and computation requirements. In addition, the size and complexity of LLMs pose challenges in XAI, an emerging field in ML concerned with the problem of explaining how a model arrives at its outputs. Model compression techniques such as pruning can be effective in reducing resource requirements and enabling more efficient inference in downstream tasks. However, it is not well understood if and how pruning of LLMs affects their explainability. Our work investigates this open problem. We identify faithfulness of explanations as a necessary metric in determining a model’s explainability. We then evaluate the faithfulness of SHapley Additive exPlanations (SHAP) and Integrated Gradients (IG) explanations of variously pruned and non-pruned DistilBERT and RoBERTa models trained on the IMDb and Yelp Polarity datasets for binary sentiment classification. We find that while magnitude-based pruning does not significantly affect explanation faithfulness, random pruning can degrade explainability. Furthermore, our results indicate that explainability is primarily influenced by model architecture. We investigate the underlying geometry of the models to explain our results and find that depending on pruning method and target sparsity, high-curvature regions can emerge, potentially undermining explanation faithfulness. Our code is available at <https://github.com/camilobrownpinilla/Explain-This-Pruner>.

Introduction

As our ability to compute has, and continues, to dramatically increase thanks to Moore’s law, there has been an increased interest in machine learning, the branch of computer science that allows us to learn complex patterns from data. Among the many tools stemming from this rich field, Large Language Models (LLMs) are arguably the most powerful and well known. LLMs are trained on vast amounts of text in order to understand and generate natural language. These models can answer questions, summarize information, and assist across a wide variety of language-based tasks, proving to be a useful tool to humanity at large.

As LLMs grow in size, they have become the target of model compression techniques aiming to reduce computational demands while preserving model performance. In particular, recent work has focused on pruning, the class of methods involving the removal a subset of network parameters according to precise criteria, leaving a sparse model with more manageable resource requirements and minimal accuracy degradation (Kwon et al., 2022; Sun, Liu, Bair, & Kolter, 2024; Dery et al., 2024; J. Li, Dong, & Lei, 2024; Ma, Fang, & Wang, 2023; Frantar & Alistarh, 2023; Kurtic et al., 2022).

With LLM-usage becoming more prevalent—accelerated in part by model compression techniques making them more resource-friendly—the challenge of explaining models has become more pressing (Zhao et al., 2023). That is to say, while we can train models to produce very accurate outputs, we do not know precisely how these outputs are produced. Being able to explain how an LLM arrives at a particular output has important implications for uncovering and debugging model bias, building user trust, and enabling transparency in decision-making. Accordingly, the field of eXplainable AI (XAI) has emerged, seeking to address these concerns, with recent work in the XAI literature focusing on developing and evaluating methods for explaining LLMs. Of the many explanation methods proposed, feature-based attribution

methods refer to the class of methods that seek to explain model outputs in terms of the inputs to the model (Sanyal & Ren, 2021; Volkov & Averkin, 2024; Hao, Dong, Wei, & Xu, 2021)(Enguehard, 2023; Lyu, Apidianaki, & Callison-Burch, 2024).

Despite the focus on each issue in isolation, the relationship between LLM-pruning and LLM-explainability has not received much attention in the literature. As model compression techniques improve and become more widely used before model deployment, it is important to understand if and how the explainability of the deployed model is affected.

By investigating the effect of pruning on the explainability of LLMs, our work aims to make progress in building a bridge between developments in model compression and XAI. We hypothesize that by reducing model complexity through the extraction of high-performing subnetworks, pruning yields models whose prediction function has lower curvature, increasing LLM explainability.

We test our hypothesis through experiments with two closely related encoder-only models, DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2020) and RoBERTa (Liu et al., 2019), trained on the IMDb (Maas et al., 2011) and Yelp Polarity (Zhang, Zhao, & LeCun, 2015) datasets for sequence classification. Following XAI literature, we employ SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) and Integrated Gradients (IG) (Sundararajan, Taly, & Yan, 2017), feature-attribution-based explanation methods, and identify the faithfulness of explanations as a necessary condition for a model to be explainable (Lyu et al., 2024). We choose these models, datasets, and explanation methods because of their ubiquity in the literature and in practical applications. As both models are variants of BERT, their architectural similarities allow us to isolate the effects if pruning on explainability more precisely than what would be possible with models that differ more substantially. Additionally, the relatively small size of these models (on the order

of 100 million parameters) lets us run several experiments that would otherwise be prohibitively expensive.

We select zero-order pruning methods, including unstructured and structured magnitude-based pruning and random pruning, and prune the models with each method to varying degrees of sparsity. We evaluate the faithfulness of explanations of each pruned subnetwork against the faithfulness of explanations of the unpruned network. We also evaluate each pruned subnetwork against a network of equivalent size that is trained from a random initialization—this gives insight into whether the particular pruning method affects explanation faithfulness, or if effects on explainability should be attributed primarily to the reduction in network size.

Finally, we investigate the local geometry of each model to understand how pruning affects the faithfulness of local explanation methods. We analyze network geometry through the lens of local curvature, since SHAP operates on the assumption of local linearity¹. In particular, we approximate the global average of the local curvature for training samples. We estimate the local curvatures using an approximation of the Hessian diagonals computed through a variation of Hutchinson’s trace estimator (Yao et al., 2021).

We find that magnitude-based pruning does not significantly affect explanation faithfulness, and importantly, does not hurt explainability while maintaining test accuracy comparable to an unpruned model. However, we find that Random Unstructured pruning can degrade faithfulness of explanations and argue that this occurs due to the emergence of high-curvature regions that violate linearity assumptions of the explanation methods.

Methods

In this section, we describe and motivate the models, datasets, methods, and metrics used in our experiments. We then detail our approach.

2.1 Models

We conduct experiments using DistilBERT and RoBERTa, high-performing language models based on the transformerencoder architecture (Devlin, Chang, Lee, & Toutanova, 2019)(Liu et al., 2019). These models are commonly used in related literature for benchmarking of language modeling tasks and are prevalent in practical applications for Natural Language Processing (NLP), making them suitable choices for studying the topic of this work.

2.2 Datasets

We train and evaluate our models on the IMDB (Maas et al., 2011) and Yelp Polarity (Zhang et al., 2015) datasets for the task of binary sentiment classification. These datasets are commonly used in conjunction with DistilBERT and RoBERTa in the literature for experimentation in the NLP domain.

2.3 Pruning Methods

We prune our trained models using Random Unstructured, L1 Unstructured, and L1 Structured pruning. Random Unstructured pruning removes a random subset of parameters constituting a specified percentage of the network. The latter two

are magnitude-based methods, which are the foundation of both classic and more recent pruning techniques (Han, Pool, Tran, & Dally, 2015; Sun et al., 2024). For example, pruning 80% of a network with L1 Unstructured pruning removes the smallest 80% of parameters ordered by L1-norm. In contrast, L1 Structured pruning removes entire channels with the lowest L1-norm.

2.4 Explanation Methods

We generate explanations of the model’s sentiment classifications using the feature attribution methods SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) and Integrated Gradients (IG) (Sundararajan et al., 2017). These methods assign importance scores to input features for a particular model prediction.

SHAP computes feature attributions based on Shapley values from cooperative game theory, approximating each feature’s marginal contribution to the prediction by considering different subsets of features (Lundberg & Lee, 2017). IG approximates the integral of gradients of the model’s output with respect to the input features along the straight-line path from a baseline input to the given input (Sundararajan et al., 2017). Following standard practice, we use the zero vector as the baseline input.

We follow previous work in adopting these methods for explaining language models. For example, Mosca et al. review SHAP-based methods applied to NLP tasks (Mosca, Szigeti, Tragianni, Gallagher, & Groh, 2022). Hao et al. (Hao et al., 2021) and Janizek et al. (Janizek, Sturfels, & Lee, 2020) adapt IG to explain token importance in sentiment classification.

2.5 Geometry

To assess the local geometry of the functions represented by the models of interest, we consider the average local curvature of the logit function with respect to the embeddings. For each model, we compute the local curvature around the predictions for 10% of the training set and take the average.

Computing second order derivatives for machine learning models with high parameter count is prohibitively expensive (Elsayed, Farrahi, Dangel, & Mahmood, 2024). Therefore, we adopt compute-efficient approximations of local curvature. There is a consensus in the literature that the Hessian diagonal is a good proxy for this task (Elsayed et al., 2024)(Yao et al., 2021). In particular, we compute an approximation of the diagonal of the Hessian around a given prediction using a modification of Hutchinson’s method for trace estimation, modelled after the implementation by Yao et al. (Yao et al., 2021). This method makes use of the fact that

diagonal(H) = E[v ∘ (Hv)]

where ∘ is pointwise multiplication and the expectation can be taken over a Rademacher or Gaussian distribution (Meyer & Avron, 2023). We choose the latter. The expectation is approximated by averaging over vectors v randomly drawn from the distribution.

Given the approximate Hessian diagonals, we then proceed with averaging them over a 10% subset of the training set.

We further distill the resulting average Hessian diagonal by looking at the largest absolute value achieved by its elements as well as their average absolute value, as the diagonals themselves are too large for human interpretation.

2.6 Evaluation Metrics

We use the following evaluation metrics to study the effect of pruning on model explainability.

Accuracy. We consider the accuracy of each model variation on the test dataset in order to verify the usefulness of the pruned models; in particular, we aim to investigate whether pruning affects explainability while avoiding sacrifices in model accuracy. In a practical application, a highly pruned model would not be particularly useful if it suffered from significant accuracy degradation, even if it saw improvements in explainability.

Faithfulness. Following Lyu et al., we identify faithfulness as the most important principle for evaluating an explanation (Lyu et al., 2024). Accordingly, we use the faithfulness of explanations as a metric for the explainability of our models.

Faithfulness is the degree to which an explanation accurately reflects how a model made a prediction; an unfaithful explanation, then, does not accurately describe a model’s decision-making process and therefore is not much of an explanation (Lyu et al., 2024). In the context of feature attribution explanations, faithfulness refers to the extent to which an explanation correctly captures which features of an input the model uses to generate its corresponding output (Lyu et al., 2024).

Previous work has proposed various methods for measuring the faithfulness of explanations by determining how well-aligned feature attributions are with true model behavior. Two faithfulness metrics frequently used in the XAI literature are Infidelity (INFD) (Yeh, Hsieh, Suggala, Inouye, & Ravikumar, 2019) and Faithfulness Correlation (FCor) (Bhatt, Weller, & Moura, 2020). These metrics rely on perturbing input features, measuring corresponding changes in model output, and comparing these changes to the importance scores of the perturbed features.

Yeh et al. (Yeh et al., 2019) define Infidelity as

INFD(Φ, f, x) = E_{I ~ μ_I} [(I^T Φ(f, x) - (f(x) - f(x - I)))^2]

Here, f is a black-box model, and Φ is an explanation functional. I ∈ ℝ^d is a random variable, where μ_I represents input perturbations of interest. A typical perturbation I is to replace a feature in x with some baseline value, such as 0. For a faithful explanation, we would expect the model output to change by an amount proportional to the sum of the importance scores of the perturbed features (Decker, Bhattarai, Gu, Tresp, & Buettner, 2024).

Bhatt et al. (Bhatt et al., 2020) measure faithfulness with correlation. For a model f, explanation functional Φ, input x ∈ ℝ^d, baseline value x_s, and subset size |S|, FCor defines the faithfulness of Φ to f at x as

FCor(f, Φ; x) = corr (∑_{S ∈ ([d] \setminus |S|)} Φ(f, x)_i, f(x) - f(x_{[x_s = x_s]}))

Here, corr is the Pearson correlation. Faithful explanations should have an FCor close to 1, indicating a strong positive correlation between the attribution scores given by Φ for an input x and the changes in the predictions of f under corresponding perturbations to x.

We select the FCor metric because it provides an interpretable score on a standard scale in the range [-1, 1], facilitating comparison across explanation method.

2.7 Our Approach

Model Generation. Given a model architecture and target

sparsity, we randomly initialize a model to serve as the unpruned, ‘base’ network. We then create a second model with its own random initializations and perform Random Unstructured pruning to the target sparsity. This second model serves as an independent, ‘smaller’ network.

We choose to start from random initializations to simulate the process of constructing and training a model from end to end. Additionally, starting with random initializations allows us to create and test the ‘smaller’ model; starting from pretrained weights would make the ‘smaller’ model just a pruned version of ‘base,’ rather than a smaller version with an independent set of parameters.

After the initialization of ‘base’ and ‘smaller,’ we train both for 3 epochs on the given dataset. Then, for each pruning method, we make a copy of ‘base’ and prune the parameters to the target sparsity. Following standard practice, we train each pruned model for an additional epoch to allow for accuracy recovery.

FCor Approximation. Following training and pruning, we evaluate the explainability of each model as follows. Due to resource constraints, we select 3% of the test dataset and generate an explanation for the model’s output on each test sample. For each test sample x, we compute an FCor value. We first fix |S| in Eq. 3 with a hyperparameter k. We then randomly sample 100 subsets x_s of size k from x and set them to the [MASK] token. This gives an estimate of FCor for that data point, since we do not see all ([d] \setminus |S|) subsets in the calculation, which is computationally prohibitive for long input sequences. In future work, we plan to evaluate faithfulness over more samples and with more perturbations per sample in order to achieve more accurate approximations of faithfulness.

	DistilBERT		RoBERTa	
	IMDb	Yelp	IMDb	Yelp
base	88.93	94.10	87.33	93.68
40pct	89.08	93.99	86.65	93.65
60pct	88.72	93.90	74.76	93.54
80pct	88.06	93.85	73.37	93.36

Figure 1. Test accuracies(%) across sparsities. Accuracies of pruned models are averaged across the RandUnstruct, L1Unstruct, and L1Struct methods. Accuracies of base models are averaged across the three ‘base’ models trained for each model, dataset combination.

We then take the mean of the local FCor estimates, giving a measure of the average faithfulness of explanations for that particular combination of model, dataset, explanation method, and k.

We select the [MASK] token as our baseline value for input perturbations during FCor computation because it aligns with DistilBERT’s pre-training objective of masked language modeling (Devlin et al., 2019). Moreover, the SHAP library uses the [MASK] token when perturbing inputs to estimate feature importance, so it is a natural choice for faithfulness evaluation (Lundberg & Lee, 2017).

¹ We do not discuss magnitudes of gradients in the body of this paper, as we do not expect SHAP nor IG to be influenced by them. We did, however, collect gradient-magnitude data and find that gradient magnitude is not correlated with faithfulness of explanations. The data can be found in figure 13.

k Hyperparameter. We ran experiments with different k to study whether masking different numbers of tokens at a time has an effect on FCor; in particular, we were interested in observing whether token importance scores are independent, or if explanations vary in their faithfulness when we consider groups of tokens and their summed importance. Initial tests showed negligible differences in FCor scores across $k = 1, 2, 3, 5$, suggesting that explanations are similarly faithful for each of these feature subset sizes. Future work will consider greater k to study if faithfulness is impacted. We fix $k = 3$ for our remaining experiments, due to resource constraints.

Experiments and Results

We conduct experiments on DistilBERT and base RoBERTa, trained on IMDB and Yelp Polarity. We prune each model using Random Unstructured, L1 Unstructured, and L1 Structured pruning, to sparsities of 40%, 60%, and 80%. We elect these percentages to balance computational constraints and the minimal effects initially observed when pruning under 40%. We evaluate the faithfulness of SHAP and IG explanations for each model using FCor with $k = 3$ and plot the distribution of the scores across test samples. We plot the distribution of FCor

scores for SHAP on all different models and report the average FCor scores in **Figure 2**. We find no significant pattern in our results, but note that scores were much more consistent across all DistilBERT experiments compared to RoBERTa, indicating that explainability may rely more on model architecture than sparsity or training data. To quantify the global average of a model’s local curvature, we use 10% of its training data to approximate the Hessian Diagonal for each sample using the variation of Hutchinson’s trace estimation described in 2.5. For each sample, we use 3 directional vectors drawn from a standard normal distribution, to reduce computational costs. When computing the maximum value along this diagonal, we consistently find extreme outliers among the models pruned via random unstructured pruning, indicating this method may produce regions of high curvature in the underlying model’s geometry. Further, we average over many test samples to help mitigate variance in local approximations caused by a low number of directional vectors (**Figure 4**). We note that, in this case, randomly pruned models still result in the largest values. Increasing the number of directional vectors may result in a more accurate estimation for each sample, and is left for future work.

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb 40pct	0.45	0.35	0.43	0.43	0.42	0.16	-0.07	0.33	0.27	0.03
IMDb 60pct	0.44	0.41	0.44	0.43	0.38	0.32	0.09	0.14	0.13	-0.19
IMDb 80pct	0.43	0.31	0.44	0.44	0.42	0.19	0.24	0.22	-0.07	0.30
Yelp 40 pct	0.63	0.55	0.62	0.63	0.63	0.45	0.46	0.32	0.47	0.49
Yelp 60 pct	0.64	0.60	0.59	0.65	0.61	0.47	0.32	0.50	0.54	0.34
Yelp 80 pct	0.62	0.59	0.59	0.61	0.62	0.46	-0.13	0.47	0.33	0.00

Figure 2. Average FCor Scores (max in bold, values rounded to 2 decimal places).

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb 40pct	0.81	12.68	1.02	0.80	49.68	0.36	3.86	0.44	0.42	7.86
IMDb 60pct	0.75	15.99	1.19	1.28	43.52	0.56	0.06	0.80	1.36	41.86
IMDb 80pct	0.83	739.25	1.06	1.09	184.78	0.50	0.03	0.58	0.80	0.28
Yelp 40 pct	0.96	17.11	1.20	1.11	38.36	2.26	4.95	3.29	0.67	10.67
Yelp 60 pct	1.37	24.28	1.24	1.28	13.69	1.01	0.92	1.29	0.49	6.18
Yelp 80 pct	1.53	61.92	1.23	0.80	72.80	0.73	14090.04	0.66	1.22	0.08

Figure 3. Maximum absolute value of Hessian Diagonal (min in bold, values rounded to 2 decimal places).

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb 40pct	0.07	0.08	0.09	0.07	0.08	0.03	0.04	0.03	0.02	0.06
IMDb 60pct	0.07	0.10	0.08	0.05	0.12	0.04	0.00	0.05	0.03	0.28
IMDb 80pct	0.08	0.22	0.06	0.06	0.18	0.03	0.00	0.04	0.02	0.00
Yelp 40 pct	0.05	0.04	0.05	0.04	0.04	0.04	0.03	0.07	0.03	0.03
Yelp 60 pct	0.05	0.04	0.05	0.04	0.05	0.03	0.02	0.03	0.02	0.03
Yelp 80 pct	0.05	0.06	0.04	0.03	0.05	0.03	13.72	0.02	0.02	0.00

Figure 4. Average absolute value of Hessian Diagonal (min in bold, values rounded to 2 decimal places).

Discussion

We discuss the results of our experiments, highlight key findings, and explain our findings in terms of local geometry. We observe in all our experiments that magnitude-based pruning does not significantly degrade accuracy on the test set. This agrees with the pruning literature, which has found that a prune-retrain approach can achieve high levels of sparsity with competitive accuracy (Han et al., 2015; Frankle & Carbin, 2019).

4.1 Standard IG is ill-suited for the language domain

IG assumes that (1) input features are independent, and (2) there exists a zero-information baseline from which a straight-line path integral yields faithful feature attributions (Sundararajan et al., 2017). IG was developed with the image classification domain in mind, where it was designed to operate on pixels as input features. In this domain, assumptions (1) and (2) are intuitively reasonable due to the continuous nature of image data. However, these assumptions do not hold for language models: language is inherently context dependent, invalidating (1); and (2) does not hold because points interpolated along a straight-line path from, say, a baseline zero-vector to the input embedding cannot be assumed to represent valid text data, since the word embedding space is discrete (Sanyal & Ren, 2021). Indeed, our experiments give evidence for the unfaithfulness of IG explanations for language models: across all models, pruning methods, and sparsities, the distribution of FCor scores of IG explanations are approximately normal with mean 0 (**Figures S1, S2, S3, S4**). This implies that there is no correlation between the IG-assigned token importance scores and the actual behavior of the model.

While assumption (1) is unavoidable due to the nature of natural language, variations of IG have been developed to correct (2) for language settings by considering semantically plausible non-linear paths from a baseline embedding to the input embedding (Enguehard, 2023; Sanyal & Ren, 2021). We observe significantly higher FCor scores for SHAP explanations, suggesting that the failure of assumption (2) underlies IG’s unfaithfulness. SHAP still assumes (1) (Lundberg & Lee, 2017), but the FCor scores indicate a moderate to strong positive correlation between claimed feature importances and model behavior (**Figure 2**). We leave further investigation of these claims and evaluation using improved IG methods (Enguehard, 2023; Sanyal & Ren, 2021) to future work.

4.2 Magnitude-based pruning does not affect explainability, but Random Unstructured pruning may hurt it

Our results do not show a significant effect of pruning on faithfulness of explanations for magnitude-based methods. In our experiments (**Figures 5, 6, 8**), we see that for a particular choice of architecture and dataset, the distribution of FCor scores does not vary significantly between the ‘base’ model and the models pruned with L1Unstructured and L1Structured methods. A notable exception is RoBERTa trained on IMDB (**Figure 7**), which we discuss further below. Moreover, the data do not show a consistent relationship between target sparsity and FCor score with remaining variables held constant (**Figure 2**), suggesting that changes in explanation faithfulness are primarily due to other factors, particularly model architecture and dataset.

However, we observe that explanations of Random Unstructured pruned models generally underperform in faithfulness. In particular,

Random Unstructured pruning never gives the highest FCor for a particular model, dataset, and sparsity. Moreover, it gives the lowest FCor of the pruning methods in all but 3 experiments, where RandUnstruct has the second lowest FCor by only 0.01-2 (**Figure 2**). These findings suggest that Random Unstructured pruning may negatively affect model explainability by undermining faithfulness of explanations. To understand why Random Unstructured pruning may negatively affect model explainability, we consider the effect of pruning on the local geometry of a model’s decision function.

4.3 Random Unstructured pruning creates highly curved regions

We observe that Random Unstructured pruned models have the largest Maximum Absolute Value of Hessian Diagonal (MAVHD) across the pruning methods for all but 3 experiments (**Figure 3**). Furthermore, the average MAVHD across all Random Unstructured pruned models is about 1069x and 1323x the average MAVHD for L1Unstructured pruned models and L1Structured pruned models, respectively. Note the outliers 739.25 and 14090.04 in the entries for DistilBERT-IMDb-80pct and RoBERTa-Yelp-80pct, respectively. We discuss a possible explanation below.

These findings suggest that Random Unstructured pruning destroys local linearity of the models’ underlying functions. A pruned model can be imagined as a fewer-parameter approximation of a base function. Removing weights at random has the potential to significantly modify the geometry of the function, creating regions with jagged decision boundaries and increased local curvature. By contrast, magnitude-based methods prune the weights that contribute least to the output, reducing the capacity for altering the behavior of the function. We hypothesize that, over models with large numbers of parameters (~ 100M), there is a low probability of creating a high-curvature region through random pruning, resulting in a similar average local curvature despite lower explainability due to a few diabolical regions.

The presence of such highly curved regions can be verified by considering the MAVHD. The data show that MAVHD for RandUnstruct is substantially higher than other methods in most cases, despite the average value being very similar (**Figures 3, 4**).

The data also suggest that the probability of high-curvature regions emerging depends on the target sparsity, with 80% Random Unstructured pruning resulting in extreme MAVHD values in some cases.

We also find that the ‘smaller’ models have large maximum local curvature. Recall that we create the smaller model by randomly removing weights in an initialization independent from the ‘base,’ running the same risk of creating curved regions as the RandUnstruct method. However, the smaller model trains for longer at that level of sparsity and therefore has more opportunities to smooth curved regions during training. The data reflect this, as the largest MAVHD for the smaller models is substantially less than that of the random unstructured models (14090.04 vs 184.78).

Precisely characterizing the mechanism by which random pruning produces sharply curved regions is an interesting direction for future work.

4.4 Highly curved regions make SHAP less faithful

The SHAP explanation method operates on the assumption of a locally linear model (Lundberg & Lee, 2017). A high MAVHD indicates the presence of a region with high local curvature,

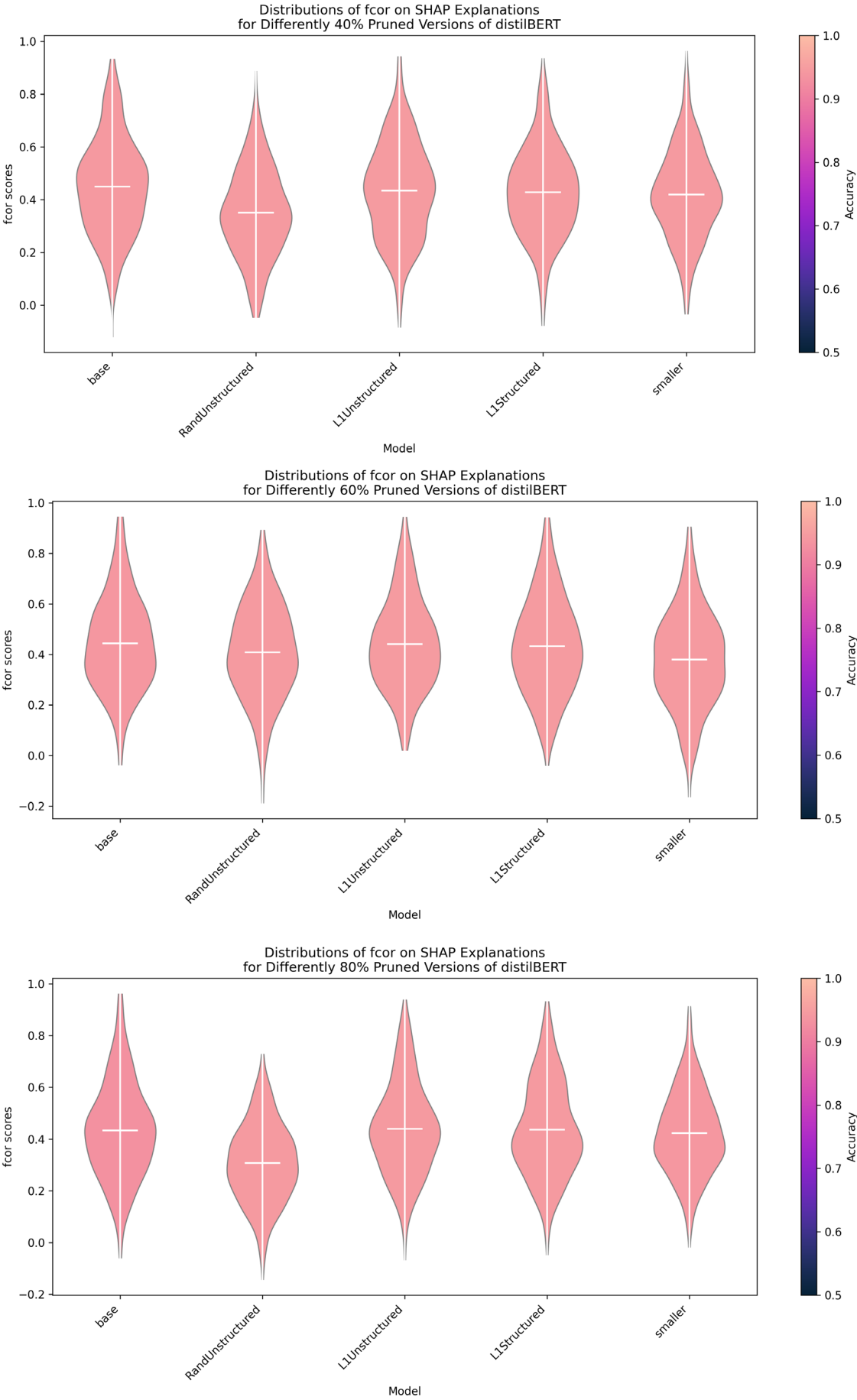


Figure 5. Distributions of FCor scores for SHAP on DistilBERT on IMDb.

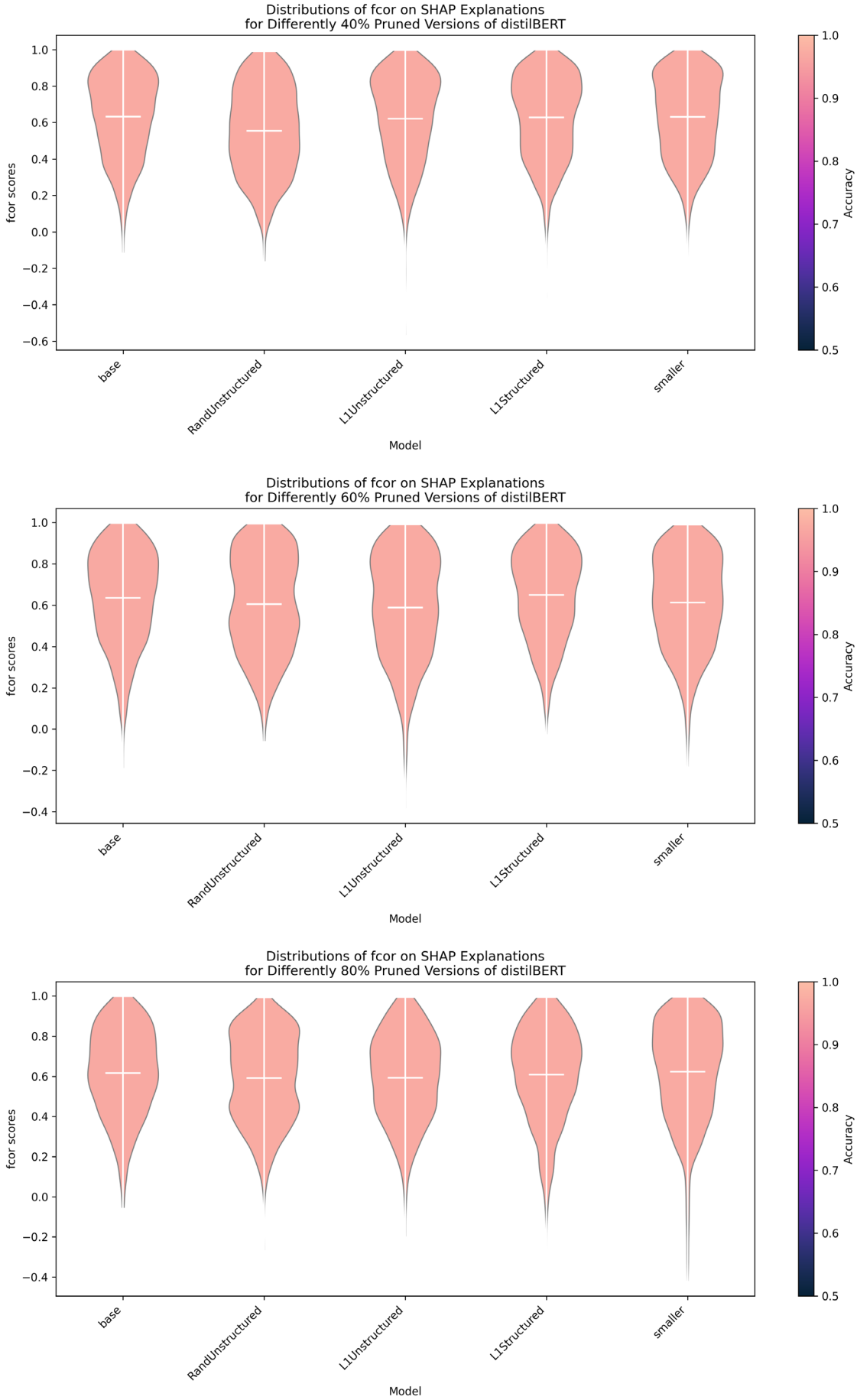


Figure 6. Distributions of FCor scores for SHAP on DistilBERT on Yelp.

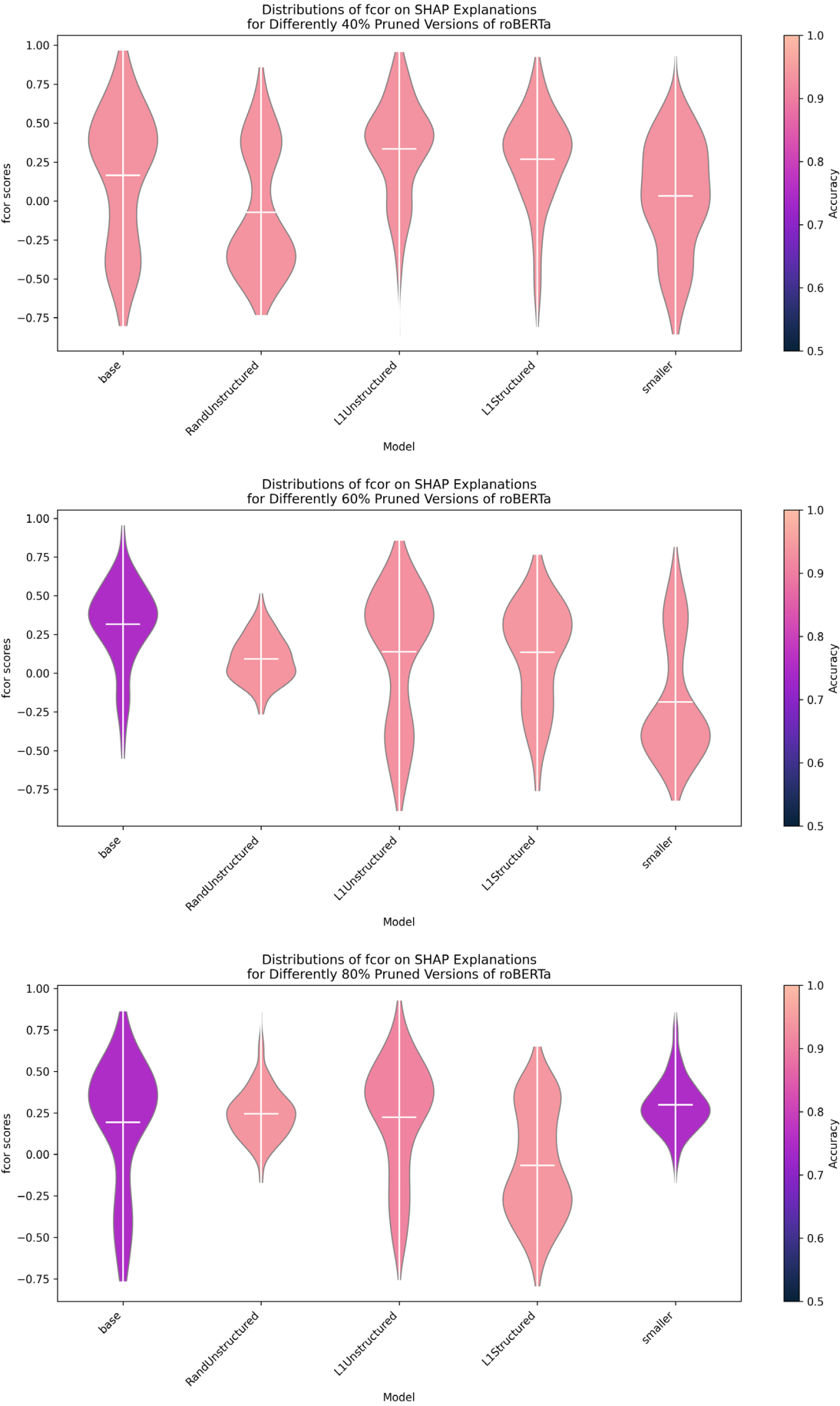


Figure 7. Distributions of FCor scores for SHAP on RoBERTa on IMDB.

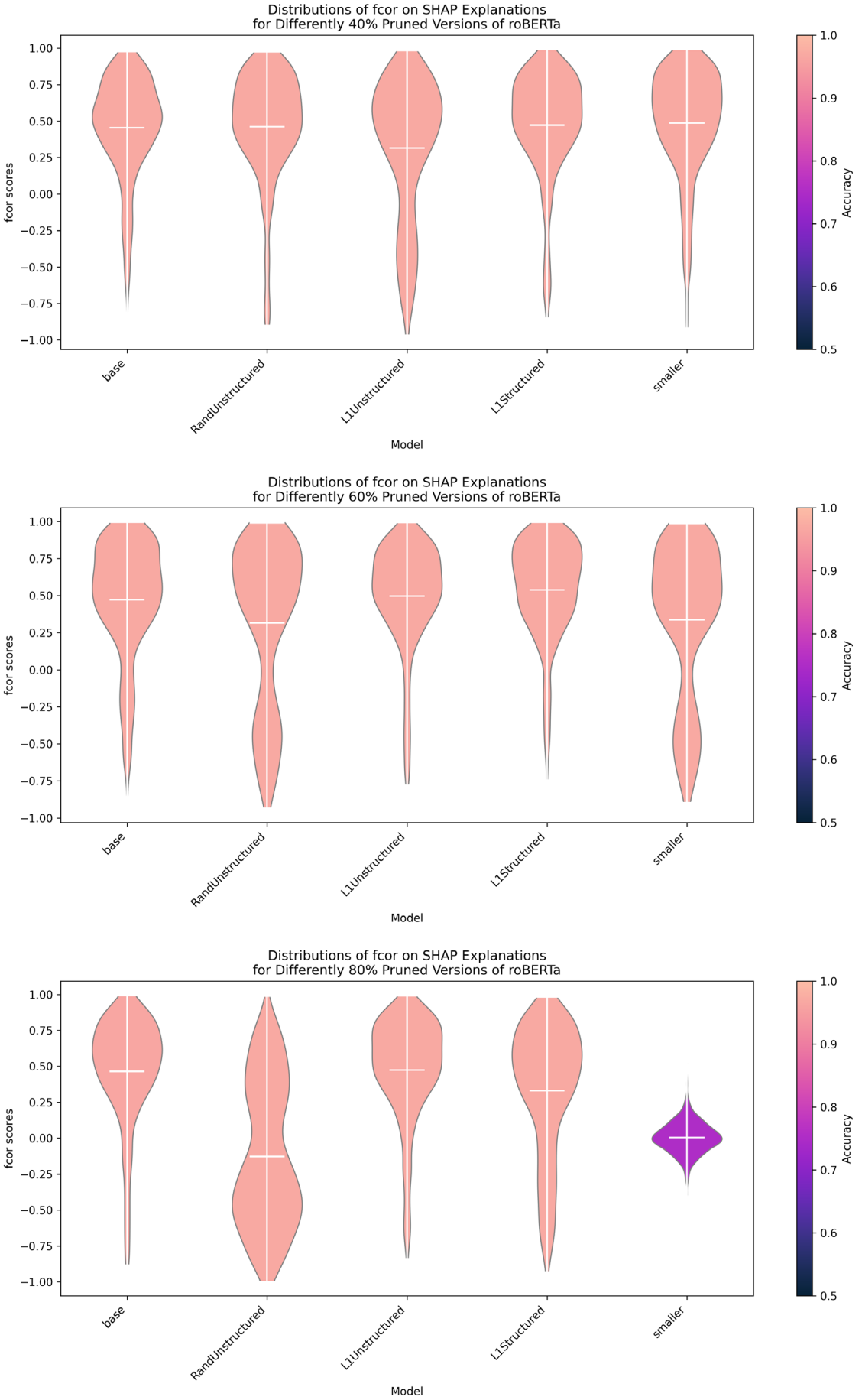


Figure 8. Distributions of FCor scores for SHAP on RoBERTa on Yelp.

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb 40pct	0.24	1.35	0.24	0.30	2.07	0.29	0.27	0.02	0.07	0.16
IMDb 60pct	0.24	0.21	0.31	0.45	1.73	0.25	0.01	0.04	0.09	0.07
IMDb 80pct	0.31	0.44	0.32	0.38	0.78	0.27	0.00	0.02	0.06	0.00
Yelp 40 pct	0.14	0.70	0.17	0.16	2.19	0.04	0.22	0.02	0.10	0.42
Yelp 60 pct	0.17	0.84	0.21	0.28	0.27	0.05	0.05	0.01	0.06	0.08
Yelp 80 pct	0.18	0.83	0.18	0.25	0.38	0.07	2.11	0.08	0.10	0.00

Figure 9. Maximum absolute value of gradient (**min** in bold, values rounded to 2 decimal places).

	DistilBERT					RoBERTa				
	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller	Base	RandUnstruct	L1Unstruct	L1Struct	Smaller
IMDb 40pct	0.03	0.03	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00
IMDb 60pct	0.03	0.03	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00
IMDb 80pct	0.02	0.04	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Yelp 40 pct	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01
Yelp 60 pct	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
Yelp 80 pct	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.10	0.03

Figure 10. Average absolute value of gradient (**min** in bold, values rounded to 2 decimal places).

implying local non-linearity; therefore, we expect high MAVHD to correspond with low FCor scores, since a model that is locally non-linear undermines the faithfulness of SHAP explanations. Indeed, we find a strong negative correlation between FCor scores and the MAVHD, with $r = -0.76$, $r = -0.13$, $r = -0.65$, and $r = -0.72$ for DistilBERT-IMDb, DistilBERT-Yelp, RoBERTa-IMDb, and RoBERTaYelp, respectively. Note the outlying weak correlation for DistilBERT trained on Yelp; we hypothesize that due to using only 3% and 10% of the training sets for FCor and Hessian computation, respectively, our FCor and Hessian approximations are not equal in their accuracy and coverage of the model’s decision landscape, potentially explaining outliers in results. Future work will address these issues.

4.5 Explainability is architecture-sensitive

We observe that explanation faithfulness is significantly impacted by model architecture. This is a very intuitive result, as one could imagine the trivial case of a constant function which is maximally explainable. However, our results give some insight into how explainability is impacted by the architecture of complex language models.

First, we observe that the distribution of FCor scores has a characteristic shape that varies primarily with architecture, holding all else constant. This suggests that the explainability of a model and its approximations (i.e. pruned models) is highly sensitive to choice of architecture (**Figures 5, 6 vs. 7, 8**).

Second, we observe more variation in FCor across pruning methods when applied to RoBERTa. This suggests that, in addition to influencing faithfulness of explanations of a model, choice of architecture also impacts a model’s sensitivity to pruning with respect to explainability. Framing this in the context of the pruning-curvature hypothesis discussed in section 4.3, it is possible that the probability of creating a curved region from a random pruning event is determined by model architecture. For example, **Figures 7 and 8** demonstrate that the distributions

of FCor for the RandUnstruct and ‘smaller’ models (which we hypothesize to be the most susceptible to the creation of high-curvature regions) tend to vary the most dramatically from the ‘base’ distribution. In contrast, the DistilBERT distributions exhibit negligible changes in their characteristic shape (**Figures 5, 6**).

To explain the differences in explainability across architectures, we observe that (1) RoBERTa has twice as many layers as DistilBERT, and (2) RoBERTa makes use of a pooling layer, which is designed to aggregate and capture all of the information contained within an encoded, variable-length input sequence and compress it into a single, fixed-length vector (Liu et al., 2019).

We hypothesize that the increased sensitivity of RoBERTa to pruning compared to DistilBERT stems from the inclusion of the pooling layer. Intuitively, the process of condensing an entire variable-length representation sequence into a fixed-length vector may result in a very dense and uninterpretable representation vector that is used directly to compute the model’s output. Future work will take a modular approach to explainability and consider the effects of particular architectural choices on a model’s explainability.

Related Work

There is some existing literature at the intersection of XAI and neural network pruning. Weber et al. study the effect of pruning on CNN explainability, finding that magnitude-based pruning methods are effective in reducing network complexity and thereby improving explainability of image classification models (Weber, Merkle, Schöttle, & Schlögl, 2023). Khalifa et al. use tree-based pruning methods to transform Random Forest models into explainable models without sacrificing accuracy (Khalifa, Abdelkader, & Elsaid, 2024).

In addition, there has been work on explainability-aware pruning methods, which seek to use explainability criteria to

determine which parameters or filters of a model to prune (Yu & Xiang, 2023; Z. Li & Song, 2024).

However, to the best of our knowledge, previous work has not investigated the effect of pruning on the explainability of LLMs. The impact of pruning on the local geometry of the network has not been well-studied. While there have been investigations on the effect of pruning on the geometry of the loss landscape or on the decision boundaries, no such work has been conducted for the impact of pruning on the local geometry of the function represented by a neural network (Cai et al., 2023; Tran, Fioretto, Kim, & Naidu, 2022).

Conclusion

This work investigates the effect of zero-order pruning methods on the explainability of DistilBERT and RoBERTa, as measured by the FCor scores of SHAP and IG explanations. We initially find that IG is ill-suited for the language domain due to the discrete nature of natural language, while SHAP gives much more faithful explanations for the sentiment analysis task.

We do not find that magnitude-based pruning affects explainability; however, it preserves accuracy as expected. In contrast, Random Unstructured pruning had a negative effect on explainability on average. We explain this finding by showing that Random Unstructured pruning can create highly curved regions in a network’s decision function, undermining SHAP faithfulness by violating the local linearity assumption. Finally, we observe that explanation faithfulness is highly dependent on model architecture, and offer an explanation based on RoBERTa’s pooling layer.

Limitations and Future Work. We experiment with relatively small language models by current standards. Future work will experiment with larger models and more varied architectures to study how the relationship between pruning and explainability is affected.

Both the IMDb and Yelp Polarity datasets used in this work represent the task of binary sentiment classification. Future work will investigate in more depth the effect of varying dataset task, size, and complexity on the trained model’s explainability. This is an especially interesting line of future work, since our results show that both accuracy and FCor increase across the board when comparing models trained on IMDb to models trained on Yelp, holding all else constant (**Figures 1, 2**). We hypothesize that the improvement in accuracy is due to the Yelp dataset’s larger size, but the effect of the choice of dataset on model explainability is unclear.

Additionally, we recognize that the paradigm in state-of-the-art language model training favors the fine-tuning of highperforming foundation models to specific tasks, contrasting with our reinitialization and train from scratch approach. Future work will investigate if our results remain consistent across training schemes.

There also remains much to explore with regard to other pruning methods. While this work selects classic, magnitude based methods as a starting point for investigating the effect of pruning on explainability, recent work has developed pruning methods tailored for LLMs, including structured and higher-order methods (Kwon et al., 2022; Sun et al., 2024; Dery et al., 2024; J. Li et al., 2024; Ma et al., 2023; Frantar & Alistarh, 2023; Kurtic et al., 2022). These methods may vary in their effect on

the evaluation metrics. Additionally, our sparsity levels are not exhaustive, and there remains much to learn on how pruning below 40% and beyond 80% affects model accuracy, explanation faithfulness, and network geometry.

Future work will investigate the effect of pruning on other metrics in the explainability literature such as robustness (Chen, Subhash, Havasi, Pan, & Doshi-Velez, 2024).

Furthermore, it will be interesting to perform a more fine-grained analysis of the models’ local geometries and to develop theoretical guarantees for the effects of different pruning methods on the geometry of the network.

Finally, all of the future work mentioned so far will be helpful for developing an explainability-optimizing pruning method that does not significantly impact accuracy.

Additional Materials

Supplemental figures S1-S4 can be found online at thurj.org.

References

Bhatt, U., Weller, A., & Moura, J. M. F. (2020, May). *Evaluating and Aggregating Feature-based Model Explanations*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2005.00631> (arXiv:2005.00631 [cs]) doi: 10.48550/arXiv.2005.00631

Cai, J., Nguyen, K.-N., Shrestha, N., Good, A., Tu, R., Yu, X., . . . Serra, T. (2023, January). *Getting Away with More Network Pruning: From Sparsity to Geometry and Linear Regions*. arXiv. Retrieved 2024-11-08, from <http://arxiv.org/abs/2301.07966> (arXiv:2301.07966 [cs])

Chen, Z., Subhash, V., Havasi, M., Pan, W., & Doshi-Velez, F. (2024). *What makes a good explanation?: A harmonized view of properties of explanations*. Retrieved from <https://arxiv.org/abs/2211.05667>

Decker, T., Bhattarai, A. R., Gu, J., Tresp, V., & Buettner, F. (2024, June). *Provably Better Explanations with Optimized Aggregation of Feature Attributions*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2406.05090> (arXiv:2406.05090 [cs]) doi: 10.48550/arXiv.2406.05090

Dery, L., Kolawole, S., Kagy, J.-F., Smith, V., Neubig, G., & Talwalkar, A. (2024, February). *Everybody Prune Now: Structured Pruning of LLMs with only Forward Passes*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2402.05406> (arXiv:2402.05406 [cs]) doi: 10.48550/arXiv.2402.05406

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/1810.04805> (arXiv:1810.04805 [cs]) doi: 10.48550/arXiv.1810.04805

Elsayed, M., Farrahi, H., Dangel, F., & Mahmood, A. R. (2024). Revisiting scalable hessian diagonal approximations for applications in reinforcement learning. *ArXiv, abs/2406.03276*. Retrieved from <https://api.semanticscholar.org/CorpusID:270258083>

Enguehard, J. (2023, May). *Sequential Integrated Gradients: a simple but effective method for explaining language models*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2305.15853> (arXiv:2305.15853 [cs]) doi: 10.48550/arXiv.2305.15853

Frankle, J., & Carbin, M. (2019, March). *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/1803.03635> (arXiv:1803.03635 [cs]) doi: 10.48550/arXiv.1803.03635

Frantar, E., & Alistarh, D. (2023, March). *SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot*. arXiv. Retrieved 2024-12-07, from <http://arxiv.org/abs/2301.00774> (arXiv:2301.00774 [cs]) doi: 10.48550/arXiv.2301.00774

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015, October). *Learning both Weights and Connections for Efficient Neural Networks*. arXiv. Retrieved 2024-12-07,

from <http://arxiv.org/abs/1506.02626> (arXiv:1506.02626 [cs]) doi: 10.48550/arXiv.1506.02626

Hao, Y., Dong, L., Wei, F., & Xu, K. (2021, February). *Self-Attention Attribution: Interpreting Information Interactions Inside Transformer*. arXiv. Retrieved 2024-12-07, from <http://arxiv.org/abs/2004.11207> (arXiv:2004.11207 [cs]) doi: 10.48550/arXiv.2004.11207

Janizek, J. D., Sturmfels, P., & Lee, S.-I. (2020, June). *Explaining Explanations: Axiomatic Feature Interactions for Deep Networks*. arXiv. Retrieved 2024-12-07, from <http://arxiv.org/abs/2002.04138> (arXiv:2002.04138 [cs]) doi: 10.48550/arXiv.2002.04138

Khalifa, F. A., Abdelkader, H. M., & Elsaid, A. H. (2024). An analysis of ensemble pruning methods under the explanation of random forest. *Information Systems*, 120, 102310. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0306437923001461> doi: <https://doi.org/10.1016/j.is.2023.102310>

Kurtic, E., Campos, D., Nguyen, T., Frantar, E., Kurtz, M., Fineran, B., . . . Alistarh, D. (2022). *The Optimal BERT Surgeon: Scalable and Accurate Second-Order Pruning for Large Language Models*. arXiv. Retrieved 2024-11-08, from <https://arxiv.org/abs/2203.07259> (Version Number: 3) doi: 10.48550/ARXIV.2203.07259

Kwon, W., Kim, S., Mahoney, M. W., Hassoun, J., Keutzer, K., & Gholami, A. (2022). *A Fast Post-Training Pruning Framework for Transformers*. arXiv. Retrieved 2024-11-08, from <https://arxiv.org/abs/2204.09656> (Version Number: 2) doi: 10.48550/ARXIV.2204.09656

Li, J., Dong, Y., & Lei, Q. (2024, July). *Greedy Output Approximation: Towards Efficient Structured Pruning for LLMs Without Retraining*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2407.19126> (arXiv:2407.19126 [cs]) doi: 10.48550/arXiv.2407.19126

Li, Z., & Song, Z. (2024, April). Structured Pruning Strategy Based on Interpretable Machine Learning. In *2024 5th International Conference on Computer Engineering and Application (ICCEA)* (pp. 801–804). Hangzhou, China: IEEE. Retrieved 2024-12-03, from <https://ieeexplore.ieee.org/document/10603526> doi: 10.1109/ICCEA62105.2024.10603526

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019, July). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/1907.11692> (arXiv:1907.11692 [cs]) doi: 10.48550/arXiv.1907.11692

Lundberg, S., & Lee, S.-I. (2017, November). *A Unified Approach to Interpreting Model Predictions*. arXiv. Retrieved 2024-12-07, from <http://arxiv.org/abs/1705.07874> (arXiv:1705.07874 [cs]) doi: 10.48550/arXiv.1705.07874

Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024, January). *Towards Faithful Model Explanation in NLP: A Survey*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2209.11326> (arXiv:2209.11326 [cs]) doi: 10.48550/arXiv.2209.11326

Ma, X., Fang, G., & Wang, X. (2023, September). *LLM-Pruner: On the Structural Pruning of Large Language Models*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2305.11627> (arXiv:2305.11627 [cs]) doi: 10.48550/arXiv.2305.11627

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from <http://www.aclweb.org/anthology/P11-1015>

Meyer, R. A., & Avron, H. (2023). Hutchinson's estimator is bad at kronecker-trace-estimation. *ArXiv, abs/2309.04952*. Retrieved from <https://api.semanticscholar.org/CorpusID:261681808>

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022, October). SHAP-based explanation methods: A review for NLP interpretability. In N. Calzolari et al. (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 4593–4603). Gyeongju, Republic of Korea: International Committee on Computational Linguistics. Retrieved from <https://aclanthology.org/2022.coling-1.406>

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, March). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv. Retrieved 2024-12-08, from <http://arxiv.org/abs/1910.01108> (arXiv:1910.01108 [cs]) doi: 10.48550/arXiv.1910.01108

Sanyal, S., & Ren, X. (2021, August). *Discretized Integrated Gradients for Explaining Language Models*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2108.13654> (arXiv:2108.13654 [cs]) doi: 10.48550/arXiv.2108.13654

Sun, M., Liu, Z., Bair, A., & Kolter, J. Z. (2024, May). *A Simple and Effective Pruning Approach for Large Language Models*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2306.11695> (arXiv:2306.11695 [cs]) doi: 10.48550/arXiv.2306.11695

Sundararajan, M., Taly, A., & Yan, Q. (2017, June). *Axiomatic Attribution for Deep Networks*. arXiv. Retrieved 2024-12- 07, from <http://arxiv.org/abs/1703.01365> (arXiv:1703.01365 [cs]) doi: 10.48550/arXiv.1703.01365

Tran, C., Fioretto, F., Kim, J.-E., & Naidu, R. (2022, October). *Pruning has a disparate impact on model accuracy*. arXiv. Retrieved 2024-11-29, from <http://arxiv.org/abs/2205.13574> (arXiv:2205.13574 [cs]) doi: 10.48550/arXiv.2205.13574

Volkov, E. N., & Averkin, A. N. (2024, May). Local Explanations for Large Language Models: a Brief Review of Methods. In *2024 XXVII International Conference on Soft Computing and Measurements (SCM)* (pp. 189–192). Saint Petersburg, Russian Federation: IEEE. Retrieved 2024-12-03, from <https://ieeexplore.ieee.org/document/10554222/> doi: 10.1109/SCM62608.2024.10554222

Weber, D., Merkle, F., Schöttle, P., & Schlögl, S. (2023, February). *Less is More: The Influence of Pruning on the Explainability of CNNs*. arXiv. Retrieved 2024-12-08, from <http://arxiv.org/abs/2302.08878> (arXiv:2302.08878 [cs]) doi: 10.48550/arXiv.2302.08878

Yao, Z., Gholami, A., Shen, S., Mustafa, M., Keutzer, K., & Mahoney, M. W. (2021, April). *ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning*. arXiv. Retrieved 2024-12-09, from <http://arxiv.org/abs/2006.00719> (arXiv:2006.00719 [cs]) doi: 10.48550/arXiv.2006.00719

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019, November). *On the (In)fidelity and Sensitivity for Explanations*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/1901.09392> (arXiv:1901.09392 [cs]) doi: 10.48550/arXiv.1901.09392

Yu, L., & Xiang, W. (2023, June). *X-Pruner: eXplainable Pruning for Vision Transformers*. arXiv. Retrieved 2024-12-03, from <http://arxiv.org/abs/2303.04935> (arXiv:2303.04935 [cs]) doi: 10.48550/arXiv.2303.04935

Zhang, X., Zhao, J., & LeCun, Y. (2015, September). Character-level Convolutional Networks for Text Classification. *arXiv:1509.01626 [cs]*.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., . . . Du, M. (2023, November). *Explainability for Large Language Models: A Survey*. arXiv. Retrieved 2024-12-08, from <http://arxiv.org/abs/2309.01029> (arXiv:2309.01029 [cs]) doi: 10.48550/arXiv.2309.01029

Examining Period Poverty and Menstrual Equity in Nepal

Alissar Dalloul
Harvard College '27

This paper examines the issue of menstrual inequity in Nepal, highlighting the case of sixteen-year-old Anita Chand, who died from a snakebite while in menstrual exile. Despite being outlawed, the ancient practice of Chhaupadi — the Nepali tradition for menstrual exile — persists due to deep-rooted cultural beliefs that label menstruation as impure. By employing a socio-ecological model, biosocial lens, and interdisciplinary case studies, this paper analyzes Nepali traditions and social structures that promote menstrual stigma; the research demonstrates that period poverty, cultural stigmas, and voids in reproductive healthcare significantly contribute to poor (and even deadly) physical and mental health outcomes for menstruators. Moreover, in including historical and sociocultural contexts, this discussion addresses the overlaps between gender, class, and patriarchal systems that, together, further perpetuate menstrual inequity. In this way, employing a social-ecological model, a biosocial lens, and interdisciplinary studies, this discussion addresses the overlaps between gender, class, and patriarchal systems perpetuating menstrual equity. While using this biosocial lens to balance a respect for cultural mores with the need for change, this paper identifies the root social and historical problems contributing to menstrual equity — and provides comprehensive recommendations for tackling these challenges. Such recommendations focus on educational initiatives, policy reforms, and media engagement to dismantle harmful cultural norms and remediate menstrual inequity. The case of Anita Chand serves as a sobering example, underscoring the urgent need for sustained efforts that ensure no individual suffers due to a normal biological process.

Case Study: Nepali Teenage Girl Dies of Snake Bite in August 2023.

Anita Chand, a sixteen-year-old girl living in Nepal’s Baitadi district, died on August 9, 2023, from a snake bite she received while asleep. Anita’s death, however, goes beyond merely the label of a “snakebite;” rather, it reflects a complex reality rooted in ancient traditions, reproductive health education gaps, and inadequate menstrual hygiene measures (Reed, 2023).

Nepal’s challenges of limited access to menstrual products, poor menstrual hygiene practices, and cultural traditions rooted in menstrual stigmas pose life-threatening repercussions. In this case, Nepal’s cultural practice of *Chhaupadi*, or menstrual exile — a practice based on their age-old belief that menstruators are “unclean” and “untouchable” — played a significant role in this case’s tragic outcome (Reed, 2023). More specifically, Anita received this fatal snake bite while in menstrual exile — a *Chhaupadi* tradition that banishes menstruators to an unsanitary hut during their period. Despite the outlawing of *Chhaupadi* in 2005, its persistent practice raises questions about effective measures to eradicate harmful cultural norms (Gurung, 2023).

While the cause of Anita’s death is not explicitly attributed to “menstruation,” her death is inextricably linked to the stigma and poverty associated with it. This tragic incident encapsulates the need to view menstruation through a biosocial lens, acknowledging its broader impact on health and societal well-being.

Introduction Menstrual Equity as A Health Issue

Menstruation is a natural bodily process, which involves the monthly discharge of blood, mucosal tissue, and other materials from the uterus’s inner lining. Menstruators globally experience menstruation distinctively — with the age of menarche ranging from 9-16 years old and with menopause occurring between 35-55

years of age (Gurung, 2023). Period poverty, denoting the inaccessibility or unaffordability of essential menstrual products, emerges as a global health issue closely linked to socioeconomic challenges. The lack of access to menstrual products — notably labeled “poverty” — stems from a combination of factors, such as sociocultural norms and economic policies. Regrettably, menstruators — an inclusive term for all people who menstruate — face adverse health and social outcomes merely because they experience this normal bodily process.

The socioeconomic factors and consequences of period poverty are particularly evident in Nepal — a country where the 2017 Multidimensional Poverty Index (MPI) revealed that “28.6% of Nepalis were multidimensionally poor” (UNICEF, 2021; Multidimensional Poverty Peer Network). These poverty levels, coupled with menstrual product shortages and entrenched cultural stigmas that marginalize and isolate menstruators, exacerbate period poverty. Menstruators in Nepal, facing a lack of “adequate infrastructure to provide [health] education and healthcare,” therefore resort to unhygienic alternatives, such as reusing pads; these direly sought-out alternatives increase their risk for infections, harming both mental and physical health (Sharma, 2022).

Confronting menstrual equity barriers, particularly the stigmas emphasized in the initial case study, is crucial for mitigating the physiological impacts of period poverty, empowering menstruators, and advancing reform. Addressing period poverty necessitates a multifaceted approach that involves understanding and overcoming Nepal’s stigmas. These stigmas, rooted in misinformed beliefs, silence discussions and, accordingly, silence change in equity spheres.

This paper employs an interdisciplinary approach integrating historical analysis, sociocultural perspectives, and policy evaluation to investigate these issues. Drawing especially on a socioecological model and applying biosocial theories, including stigma theory, social suffering, and partial perspectives, this paper analyzes how

systemic factors contribute to menstrual stigma and period poverty. 1) The socio-ecological model dissects the issue across individual, interpersonal, community, and policy levels — revealing how cultural beliefs and structure serve as barriers to equity. 2) Case studies, such as the tragic death of Anita Chand, contextualize the real-world implications of menstrual stigma. 3) Comparative analyses of global health crises — like HIV/AIDS in China — provide insights into successful strategies for combatting social stigmas. In this way, by concurrently examining the effectiveness and gaps in current interventions, this paper aims to pinpoint the root causes of menstrual inequity and thereby guide targeted policy recommendations and community-level interventions.

Defining the “Biosocial” Categorization

The biosocial approach integrates biological and social dimensions to examine how cultural beliefs and systemic structures intersect with physical health processes. This framework emphasizes the necessity of examining both processes rather than relying solely on deterministic biological or social functions (Harris, 2018). In the context of menstrual health, this categorization highlights the interplay between menstruation as a biological function and the social structures — such as stigma, economic barriers, and healthcare access — that shape menstrual experiences. In this way, the biosocial lens enables a nuanced understanding of how period poverty and stigma are not only outcomes of individual or community ignorance — thereby helping to elucidate root causes and identify sustainable, culturally sensitive solutions.

Historical Contexts and Sociocultural Perspectives in Nepal

Stigma theory explores the social consequence of labeling and negative stereotyping’s social consequences — particularly by addressing the structural aspects of stigma, including its impact on individuals, families, social groups, and institutional discrimination. Scholars have also examined stigma theory within moral issues, where “stigmatized conditions threaten what is at stake for sufferers” (Yang, 2007). Applying this framework to Nepal reveals that menstruators are “sufferers” within a system that perpetuates inequity through harmful yet age-old constructs (Schomerus, 2021; Yang, 2007).

Nepali Hinduism and Chhaupadi: Menstrual “Untouchability”

The convergence of religious beliefs and sociocultural norms has given rise to Nepal’s practice of menstrual “untouchability” (UN Women, 2017; Crawford, 2014). In Nepali Hinduism, menstruating women are forbidden to “enter a temple or kitchen, share a bed with a husband, or touch a male relative,” as they are, during this period, considered “untouchable” (Crawford, 2014). Menstruators are also required to use separate utensils and are not allowed to see the sun, eat with others, cook food, or even worship. These beliefs have excluded menstruators from various aspects of daily life — whether social, religious, or familial (UN Women, 2017); a quintessential example is *Chhaupadi* or menstrual exile — which isolates menstruators to unsanitary huts during their period (Amatya, 2018).

The social construction of reality surrounding menstruation is rooted in historical and cultural factors; Nepal has constructed a biological reality that labels menstruation, and thus menstruators, as taboo and shameful. “Patriarchal societies, such as Nepal, objectify menstruating women’s bodies and portray them as impure, polluted,

dirty, untouchable, and harmful, fostering fear and detachment towards women and their bodies” (Gurung, 2023). This “objectification” enables restrictions on menstruators’ bodies, obligations, and decisions — ultimately culminating in period poverty and social isolation. Nevertheless, social construction theory provides a lens for intervention: Nepali society must seek to reshape, even dismantle these constructions of reality that perpetuate harmful norms for menstruators (Kleinman, 2010).

Contextualizing Menstrual Inequity in Nepal via a Socio-Ecological Model

Using the socio-ecological model as a conceptual framework to examine menstrual equity in Nepal reveals nuanced levels of community barriers to health across four levels: interpersonal, individual, community, and policy levels.

1) Individual level factors that exacerbate menstrual stigmas in Nepal include a menstruator’s “religion, caste, education level, socioeconomic status, self-silencing and internalized stigma, lack of knowledge and awareness of menstruation” (Gurung, 2023). Nepali beliefs saturated in misinformation include the following:

- a) menstruation causes illness in the family,
- b) contact with a menstruating woman can cause food to spoil, or
- c) if near a menstruator, someone could be attacked by a leopard (Gurung, 2023; Parajuli, 2017).

These misinformed beliefs result in Nepali menstruators self-isolating themselves, fearful of causing their loved ones harm.

2) Influence from immediate social relations —grandparents, parents, significant others, teachers, health care providers, etc. — also heightens stigma. A study conducted on Nepali married women that assessed the prevalence of menstrual restrictions showed that “nearly three out of every four [married] women (72.3 %) reported experiencing many menstrual restrictions, or two or more types of menstrual restrictions” (Gurung, 2023). Moreover, many women acquired their (misinformed) understanding of menstruation from their family members or peers, who were, too, influenced by generational stigmas: “Expectations of purifying... and secrecy around menstruation were imposed by adults” (Gurung, 2023). Additionally, a menstruator’s fear of being judged by their social group if they did not follow “untouchability” procedures, such as *Chhaupadi*, also perpetuates these stigmatized traditions. “Menstruation is [still] not culturally accepted as a natural biological function in Nepal [and persists] with a lack of familial and immediate circle societal support” (Gurung, 2023).

3) Within the community level — including but not limited to schools, hospitals, religious institutions, and government-driven organizations — voids in infrastructure and resources surrounding menstrual hygiene drives period poverty. Particularly, the lack of education on reproductive health in schools, coupled with a lack of awareness amongst teachers, providers, and religious leaders regarding menstrual stigma, hinders open conversations about menstrual equity in policy spheres (Thomson, 2019; Sharma, 2022). These institutions, otherwise, possess the potential to “enact regulations and policies that can reduce menstruation stigma in Nepal” (Gurung, 2023). However, once again, the essential step in remedying period poverty and menstrual inequity involves eliminating stigmas and thus encouraging discussions about menstruation.

4) Policies — ranging from educational to health — can affect meaningful change in these spheres. However, for these policies to be truly effective, they must be accompanied by meaningful efforts

to remedy harmful stigmas and practices. For example, *Chhaupadi* was banned in Nepal’s Supreme Court in 2005; and in 2018, Nepal’s government criminalized *Chhaupadi* (Amatya, 2018). Nevertheless, to this day, the tradition is still practiced due to entrenched, generational stigmas that impel individuals to defy legal barriers.

Haraway’s View of Feminist Epistemology

Recognizing that these socio-ecological levels encompass elements of diverse lived experiences, including socioeconomic, educational, and cultural backgrounds, is crucial for guiding effective interventions. This model can be linked to the theories of feminism and partial perspectives in that no single viewpoint can encapsulate a historical or social event; menstrual equity, for one, demands “multiple partial perspectives” (Haraway, 1988). More specifically, to understand menstrual equity through a feminist epistemological lens, we draw on Haraway’s reasoning — which challenges the notion of objectivity by asserting that perspectives are shaped by the specific cultural, historical, and social contexts of the observer. Haraway’s argument on epistemic privilege — which emphasizes that certain perspectives are privileged and may devalue other lived experiences — elucidates a root problem in menstrual equity. Policy-makers with “privilege” (in menstrual equity spheres, those with privilege are non-menstruators who typically identify as male) tend to overlook inequities such as period poverty, as these inequities — simply put — do not affect them (Haraway, 1988). Unfortunately, in Nepal, a patriarchal society, it is those with such privilege who dictate policy.

Nepal’s Current State of the Period: Menstrual Stigma’s Impacts on Society and Individual

Interrogating Hygiene and Cultural Judgments

As a preface, the framing of “poor hygiene” or “unhygienic” in menstrual practices often suggests implicit biases. For instance, labeling reusable materials like socks as unhygienic without acknowledging their proper absorbent use or cultural acceptability risks perpetuating stigma. As such, hygiene should be contextualized — recognizing that what constitutes “clean” or “safe” varies across settings and is influenced by resource availability, cultural norms, and personal preference. Rather than privileging disposable products as inherently superior, interventions should consider the environmental and economic sustainability of reusable options and recognize their potential to empower menstruators in resource-limited settings.

Still, at the same time, access to high-quality, hygienic menstrual products should be a fundamental right. While alternatives may be viable in different contexts, individuals should not be forced to rely on potentially unhygienic options— or at least not considered standard — due to financial constraints or lack of availability; unhygienic practices include wearing tampons for a prolonged time due to such economic constraints. In this way, ensuring equitable access to a range of menstrual products — including both disposable and reusable options — allows individuals to make informed choices based on their own needs, preferences, and circumstances. This approach acknowledges the diversity of menstrual experiences while advocating for policies and funding that guarantee access to products that meet the highest standards of health and dignity.

Chhaupadi Practices: Exile and its Consequences

A study examining *Chhaupadi* practices revealed that the majority of girls sampled were exiled during menstruation: “4% were exiled to traditional sheds, 82% to livestock sheds and 11% to

outside courtyards. Around 3% stayed inside the house but practiced some form of menstrual taboo. Only 30% of girls who stayed outdoors had toilet facilities” (Sharma, 2022; Amatya, 2018). Study participants reported psychological consequences, such as loneliness and insomnia. These issues are partly attributed to the fact that a significant percentage of exiled menstruators stayed in spaces without ventilation or even a mattress. “Nine of the menstruators in this study [similar to our initial case study] were bitten by a snake” (Sharma, 2022).

Now, bearing in mind Nepal’s historical and sociocultural contexts, consider this question again: why did sixteen-year-old Anita Chand die from a snakebite?

The cultural norm that menstruation is “impure” and “dirty” resulted in her banishment to a period hut, where she died from a poisonous snake bite (Gurung, 2023). As such, Anita’s death serves as a symbol for the other menstruators who have also died due to *Chaupaddi*’s horrific conditions — “from animal attacks and from smoke inhalation after lighting fires in windowless huts” (Reed, 2023). Anita and these menstruators essentially succumbed to the stigmas within their culture that ostracize “impure” menstruators (Gurung, 2023). Even so, those who survive menstrual exile and harmful traditions such as *Chhaupadi* must also confront a form of death — not biological but rather social in nature.

Menstrual Inequity Impacting Social Life: Education, Economy, and Social Death

Social isolation in itself, as experienced monthly by menstruators within Nepali customs of *Chhaupadi* and “untouchability,” negatively impacts mental and physical health. Recent studies link social deprivation with a “host of conditions,” including heart attacks, chronic diseases, mobility issues, high blood pressure, cancer, poor mental health, anxiety and depression, and a weakened immune system” (Umberson, 2010). Social exile also limits a menstruator’s ability to participate in the economy and receive education (Amatya, 2018). Such isolation can, therefore, be characterized as a social death, as menstruators experience exile from society and endure social and economic hardships — from missing job opportunities to losing time in school.

Social suffering refers to pain and distress caused by various social forces—including global and local economics, politics, social institutions, relationships, and regional cultures (Umberson, 2010). In menstrual equity, social suffering encompasses the pain and distress resulting from societal factors such as economic inequalities, cultural stigma, and limited access to menstrual products. This concept of suffering also emphasizes how institutions specifically designed for help and reform can, in fact, exacerbate social and health problems. Types of social suffering include structural violence, where suffering is structured by historical and economically driven processes related to factors like racism, sexism, and political violence (Umberson, 2010). In period poverty, structural violence is evident, particularly influenced by sex, poverty, and interconnected factors such as race and class. This theory, therefore, emphasizes the need for coordinated social and health policies to address structural and systemic problems within period poverty.

Within such social spheres, menstrual stigmas significantly impact education — contributing to school absenteeism in Nepal: “The major reasons for school absenteeism were discomfort, lack of continuous water supply, and shame or fear of staining [while on one’s period]” (Sharma, 2022). Statistics also show that there is a significant need for improved sex education, as only “47.5% girls learned

about menstruation in school” and “only 33.3% of the respondents used sanitary pads” (Parajuli, 2017). This disparity suggests that most menstruators learn about periods from their family or peers, who can especially act as vehicles of generational myths and stigmas. Moreover, these statistics suggest that a substantial portion of Nepal’s population lacks the means to afford or access essential menstrual products — harming their well-being and evidently hindering their school attendance (Yadav, 2018).

This theory, therefore, substantiates the argument that interventions for period poverty require a multifaceted approach: 1) addressing insufficient resources by providing tangible menstrual products, and 2) addressing menstrual stigmas and taboos, considering the social consequences of inadequate support systems, and ultimately working to dismantle societal structures that exacerbate the suffering associated with menstrual inequity.

Health Repercussions of Unhygienic Alternatives

Additional studies that have analyzed the practice of menstrual hygiene in Nepal found that “around 40% used sanitary pads during menstrual flow but most (65.1%) of them did not dispose them” (Sharma, 2022). This disparity sheds light on period poverty’s adverse impact on physical health — including increased risk for yeast infections, vulvar irritation, vaginal discomfort, and sometimes even life-threatening infections such as toxic shock syndrome (Farid, 2021). These consequences result from unhygienic menstrual practices such as wearing tampons and pads for a prolonged time, reusing sanitary pads, or resorting to household items as substitutes. Menstruators must resort to unhygienic alternatives due to the financial challenges that obstruct individuals from buying necessary products.

Impact of COVID-19 and Post-Disaster Responses on Menstrual Stigma and Period Poverty

Post-disaster responses can also be used to explore the confluence of menstruation, mental health, and period poverty; extreme conditions accentuate the grave impact of period poverty. For example, the COVID-19 pandemic significantly exacerbated issues of period poverty by dysregulating supply chains, further restricting access to menstrual products. The lockdown caused additional financial struggles, as the Nepal government placed prohibitory orders on farm production (a common familial job in Nepal).

A recent study which included 30 countries, one of them being Nepal found that during COVID, menstruators especially observed the following problems: “severe shortages of products (73%); inflated prices of pads and tampons (58%); reduced access to clean water to manage periods (51%); and increased stigma, shaming or harmful cultural practices (24%) as a result of increased social scrutiny at this time; [preexisting menstrual stigmas fueled the notion that proximity to a menstruator could lead to COVID and death]” (Plan International, 2020; Rohatgi, 2023). Moreover, despite the progress made to outlaw and criminalize *Chhaupadi* in Nepal, when COVID arose, “all focus was shifted to Covid” (Reed, 2023). People again “started staying in a shed, as “there were no programs and campaigns on *Chhaupadi* after Covid. People almost stopped talking about it” (Reed, 2023). The insufficient focus on menstruation during post-disaster responses contributes to delays in achieving menstrual equity; the aftermath of the 2015 Nepal earthquake provides another lens to examine menstrual stigma.

Despite the urgent delivery of immediate supplies post-earthquake, such as food, water, tents, and blankets, menstrual hygiene

products were overlooked. A study conducted among earthquake-affected menstruators in regions of Nepal found that “None of the respondents reported receiving menstrual adsorbents as relief materials in the first month following the earthquake” (Budhathoki, 2018). Menstrual products are, yet again, treated as luxury items rather than necessities due to misinformation and stigma.

Period Poverty, Menstrual Stigmas, and Mental Health

Menstrual inequity’s detriments to mental health are linked to both the lack of hygiene management methods and the cultural discrimination surrounding menstruation. Besides the stress concurrent with a lack of access to necessary menstrual products, Nepal’s “cultural construction of menstruation as disgusting, humiliating, and polluted can lead to women and girls’ [having] negative attitudes about their physical bodies” (Gurung, 2023). These constructions harm menstruators’ mental health, instilling low self-esteem and notions of inferiority. Moreover, the “conditioned self-silencing,” which arises from the shame surrounding menstruation, leads to many menstruators suffering in silence and, therefore, acts as a barrier to social change (Gurung, 2023).

Understanding the Socio-Ecological Model in Context with These Issues

In a systematic review conducted on menstrual health and hygiene in Nepal, prevalent issues included 1) mental health concerns, 2) menstrual hygiene practice concerns, and 3) reproductive issues. Examining menstrual health and hygiene through the socio-ecological model outlined several contributing factors: lack of proper water, sanitation, and hygiene (WASH) facilities in schools, lack of sex education in schools, lack of menstrual hygiene management (MHM) components in disaster relief efforts, and other influences such as cultural restrictions, school absenteeism, economic challenges, limited knowledge about menstruation (Thomson, 2019; Sharma, 2022).

Comparative Analyses: Thinking About Stigma Within Past Interventions

Addressing Stigma: Lessons from Comparative Analysis of Eating Disorders in Fiji

A study on “television, disordered eating, and young women in Fiji” illustrates how media influences societal norms (Becker, 2004). Pre-television introduction, Fijian social constructs favored a robust body and eating in secrecy was considered taboo. Yet, post-TV introduction, Fiji altered its ideal body type: young Fijian girls, “shaped by a desire for competitive social positioning during a period of rapid social transition,” sought to model the attributes on TV (Becker, 2004). Modeling these actresses marked “beginnings of weight and body shape preoccupation, purging behavior to control weight, and body disparagement” (Becker, 2004).

Becker’s paper, therefore, highlights the media’s influence on social constructions. In terms of period poverty, the media has shaped menstrual stigma in several ways — such as omitting menstrual products in TV shows and portraying menstruation, especially menarche, as a scary incident. Even so, the media, while capable of fostering harmful stigmas, also possesses the power to pacify such stigmas. Documentaries like *Period. End of Sentence* — which highlights women in India fighting against menstrual stigmas — and TV Shows, like “The Last of Us” — which reserve screen time to emphasize the treasure of finding tampons in an impoverished, post-apocalyptic world — help to normalize the discussion of menstruation.

Addressing Stigma: Lessons from Comparative Analysis of HIV/AIDS in China

The stigma surrounding HIV/AIDs in China draws parallels to menstrual stigma, as both stem from socially constructed norms that dictate an individual’s treatment as a “nonperson” (Jinhua, 2011). Discrimination in both cases intersects with other inequities based on gender, race, or class (Schomerus, 2021). The social death resulting from social isolation in the HIV/AIDs crisis — which makes it more difficult for patients to return to ordinary life — can also be likened to the social rejection menstruators face on their period. Menstrual stigma — fostering negative perceptions, such as “untouchability” — marginalizes menstruators

The stigma surrounding HIV/AIDs in China draws parallels to menstrual stigma, as both stem from socially constructed norms that dictate an individual’s treatment as a “nonperson” (Jinhua, 2011). Discrimination in both cases intersects with other inequities based on gender, race, or class (Schomerus, 2021). The social death resulting from social isolation in the HIV/AIDs crisis — which makes it more difficult for patients to return to ordinary life — can also be likened to the social rejection menstruators face on their period. Menstrual stigma — fostering negative perceptions, such as “untouchability” — marginalizes menstruators. Stigmatization, therefore, negatively impacts well-being, particularly mental health; it deters discussions about menstrual health problems, which can, therefore, lead to a menstruator not seeking care due to fear of discrimination.

Interventions

Existing Interventions & Efforts

In the 21st century, nonprofits such as PERIOD. and ThePadProject continue this effort, working to destigmatize conversations about menstruation, invoke legislative action, and increase access to menstrual products. Key interventions led by these organizations include: distributing free menstrual products in schools, conducting menstrual education programs, funding documentaries and research studies, implementing the “Menstrual Hygiene Day” campaign to raise awareness for menstruation, and propelling legal reforms to remove tampon taxes in America; (Period.org, 2023). Still, despite these significant strides toward equity, more must be done. For example, while eliminating certain U.S. states’ tampon tax is a positive step, menstrual products are still inaccessible to many. Affordability remains a crucial issue, and even with tax eliminations, these necessities still pose financial burdens. As such, existing interventions, while noteworthy, represent only the start of menstrual equity battles.

The documentary *Period. End of Sentence* (Sung, 2018), produced by ThePadProject, spreads awareness of period poverty by focusing on a rural Indian village. This film depicts how period poverty impacts menstruators’ attendance in school — showcasing a young girl who is unable to attend school due to stigma and lack of proper sanitary pads. The film includes interviews where both men and women describe periods as “disgusting,” compelling viewers to empathize with menstruators who feel publicly embarrassed while on their period (Sung, 2018). As viewers witness, “girls find[ing] it difficult to buy [pads] from the shop because there are a lot of men around,” this film emphasizes the urgent need to address patriarchal structures in regions like rural India and Nepal (Sung, 2018). Such awareness efforts — specifically, a film serving a global platform — shed light on the need to enhance menstruators’ lived experiences in various regions.

Recommendations: Thinking About Period Poverty with a Bio-social Lens

Therefore, understanding Nepal’s menstrual inequity through historical, sociocultural, and comparative lenses clarifies its root causes and informs targeted solutions. Future research should prioritize ethnographic studies that center on the voices of Nepali women and menstruators; capturing these perspectives will further illuminate the nuanced realities of menstrual equity and inform culturally sensitive interventions. Beyond the surface issue of limited access to menstrual products, a biosocial lens reveals how period poverty intersects with broader equity issues related to gender, race, culture, or socioeconomic status. By recognizing this, these efforts aim to expand menstrual product access while tackling the fundamental sociocultural and economic challenges — rooted especially in stigma — inextricable to menstrual inequity.

Mainly, education-centered strategies are vital to reducing menstrual stigma — empowering both menstruators and non-menstruators with accurate information. Implementing a standardized menstrual hygiene curriculum in schools would provide fundamental knowledge about biological aspects and hygiene practices; moreover, including non-menstruators in this dialogue is, too, essential for fostering an atmosphere of inclusivity and understanding — helping to dispel misconceptions and normalizing menstruation, especially within school environments. Accordingly, policies supporting regular training for educators is equally necessary — in that to effectively deliver this curriculum, teachers and healthcare professionals must be equipped with the tools and knowledge.

Community outreach programs, such as “door to door awareness programs” also offer a personalized approach to debunking myths and taboos surrounding menstruation — and therefore serve to eliminate the repercussions of damaging practices such as *Chhaupadi* (Gurung, 2023). Likewise, engaging and “educating local religious leaders, [so as to] incentivize menstrual education” could further foster greater acceptance and understanding (Gurung, 2023).

A study on period poverty in neighboring regions found that the majority of women and men received little to no education on “menstrual preparation, the menstrual cycle, [and] its causes before menarche” (Gurung, 2023). In this way, Menstrual education is necessary not only to help “women and girls make informed decisions for their body” but also to dispel deep-rooted myths that hinder equity (Gurung, 2023). Such educational strategies have been found to “reduce negative perceptions and attitudes of menstruation, menstruation hygiene products, feelings of shame and inferiority, low self-esteem, and the risk of infections and illnesses” (Gurung, 2023). Nevertheless, these educational reforms must be coupled with and enforced in policy. Policy recommendations should, therefore, aim to provide more mental health professionals, enforce menstrual education, and reform patriarchal caste systems — improving Nepali menstruators’ lived experience and weakening structures that contribute to their social suffering.

Public health initiatives can play a central role, particularly in addressing access-related barriers to menstrual products and health resources. Government-funded programs, or funded by organizations like PERIOD., to provide free or subsidized menstrual products in rural and underserved areas can ensure that essential products are consistently available. Such initiatives can significantly reduce financial barriers by establishing distribution points within schools, healthcare, and community centers.

Yet, the problem still lies in ensuring accountability of menstrual

health policies and efforts — in a culture especially influenced by menstrual stigma. As such, effective enforcement of anti-*Chhaupadi* laws is, too, essential; via community-based monitoring, including the appointment of health ambassadors, strengthened enforcement mechanisms, and awareness programs that highlight the dangers and legal consequences of *Chhaupadi* can further support gradual cultural shifts away from these harmful practices.

Moreover, advocacy efforts, as seen by organizations such as PERIOD., should be amplified in the media. With an increased voice, these nonprofits will generate more funds to distribute menstrual products, generate awareness campaigns, and incentivize legislative action throughout regions such as Nepal. Media, if strategically used, can also work to destigmatize menstruation by disseminating movies and images that normalize menstruation. The Fijian case study demonstrates that the media can alter social norms for the worse, yet by harnessing this potency to modify social norms for the better — as done in Period. End of Sentence — awareness can be deployed on a global scale.

Conclusion: Returning to the Case Study

The initial case study sheds light on Nepal’s deep-rooted social and cultural norms that lead to harmful practices, such as *Chhau-padi*. Multifaceted interventions are necessary to address period poverty’s biological and sociocultural dimensions — an approach combining policy reforms, education initiatives, and cultural aware-ness campaigns.

Although Anita’s death embodies just one case study, her death symbolizes the systemic issues surrounding menstrual inequity. The last reported death from *Chhaupadi*, before Anita, was in 2019, claiming the life of 21-year-old Parwati Budha Rawat. “Her death [had] prompted countrywide programs and campaigns to end the practice. Thousands of period huts were destroyed... people were getting information about menstruation and [the] law” (Reed, 2023). Yet, as other social and health problems arose, namely COVID, the conversation shifted — “people almost stopped talking about it” — a premise for Anita’s death two years later (Reed, 2023).

Anita’s death should, therefore, be regarded as a pressing symbol emphasizing the urgency, commitment, and persistence required to dismantle these harmful, deeply ingrained practices. Even in the 21st century, impenetrable factors persist, perpetuating stigmas that tragically cost lives. Sustained dedication — one that refuses to be eclipsed by other global issues — is required to ensure that no one’s well-being is compromised by their biological needs and that a normal bodily process hinders no one’s social existence.

References

Amatya P, Ghimire S, Callahan KE, Baral BK, Poudel KC (2018) Practice and lived experi-ence of menstrual exiles (*Chhaupadi*) among adolescent girls in far-western Nepal. PLOS ONE 13(12): e0208260. <https://doi.org/10.1371/journal.pone.0208260>.
Becker A. E. (2004). Television, disordered eating, and young women in Fiji: negotiating body image and identity during rapid social change. *Culture, medicine and psychiatry*, 28(4), 533–559. <https://doi.org/10.1007/s11013-004-1067-5>.
Budhathoki, S. S., Bhattachan, M., Castro-Sánchez, E., Sagtani, R. A., Rayamajhi, R. B., Rai, P., & Sharma, G. (2018). Menstrual hygiene management among women and adolescent girls in the aftermath of the earthquake in Nepal. *BMC women’s health*, 18(1), 33. <https://doi.org/10.1186/s12905-018-0527-y>.
Crawford, M., Menger, L. M., & Kaufman, M. R. (2014). ‘This is a natural process’: manag-ing menstrual stigma in Nepal. *Culture, health & sexuality*, 16(4), 426–439. <https://doi.org/10.1080/13691058.2014.887147>.

Farid, H. (June 1, 2021). Period equity: What it is and why it matters. *Harvard Health Publishing: Harvard Medical School*. <https://www.health.harvard.edu/blog/period-equity-what-is-it-why-does-it-matter-202106012473>.
Gurung, Ilmisha. (2023). “Menstruation stigma: A qualitative exploratory study of the lived experiences of Nepali women.” Electronic Theses and Dissertations. Paper 4105. <https://doi.org/10.18297/etd/4105>.
Haraway, D. (1988). Situated Knowledge: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3), 575–599. <https://doi.org/10.2307/3178066>.
Harris, K. M., & McDade, T. W. (2018). The Biosocial Approach to Human Development, Behavior, and Health Across the Life Course. *The Russell Sage Foundation journal of the social sciences : RSF*, 4(4), 2–26. <https://doi.org/10.7758/RSF.2018.4.4.01>
Jinhua, G. & Kleinman, A. (2011). Chapter Seven. Stigma: HIV/AIDS, Mental Illness, and China’s Nonpersons. In Deep China: The Moral Life of the Person (pp. 237-262). Berkeley: University of California Press. <https://doi.org/10.1525/9780520950511-009>.
Chapter Within The Book, Deep China.
Kleinman, A. (2010). Four social theories for global health. The Lancet, 375(9275), 1518-1519. [https://doi.org/10.1016/S0140-6736\(10\)60646-0](https://doi.org/10.1016/S0140-6736(10)60646-0). Article Within Jour-nal Multidimensional Poverty Peer Network. (no date). Nepal. https://mppn.org/paises_participantes/nepal/.
Parajuli, P., Paudel, N., & Shrestha, S. (2017). Knowledge and practices regarding menstrual hygiene among adolescent girls of rural Nepal. *Journal of Kathmandu Medical College*, 5(1), 23–27. <https://doi.org/10.3126/jkmc.v5i1.18262>.
Plan International. (2020). Periods in a Pandemic: Menstrual hygiene management in the time of COVID-19.
Reed, B. (2023, August 11). Teenage girl dies after being forced to stay in a ‘period hut’ in Nepal. The Guardian. <https://www.theguardian.com/global-development/2023/aug/11/teenage-girl-dies-after-being-forced-to-stay-in-a-period-hut-in-nepal>.
Rohatgi, A., & Dash, S. (2023). Period poverty and mental health of menstruators during COVID-19 pandemic: Lessons and implications for the future. *Frontiers in Global Women’s Health*, 4. <https://www.frontiersin.org/articles/10.3389/fgwh.2023.1128169>.
Schomerus G, Angermeyer MC (2021). Blind spots in stigma research? Broadening our perspective on mental illness stigma by exploring ‘what matters most’ in modern Western societies. *Epidemiology and Psychiatric Sciences* 30, e26, 1–6. <https://doi.org/10.1017/S2045796021000111>.
Sharma, A., McCall-Hosenfeld, J. S., & Cuffee, Y. (2022). Systematic review of menstrual health and hygiene in Nepal employing a social ecological model. *Reproductive health*, 19(1), 154. <https://doi.org/10.1186/s12978-022-01456-0>.
Sung, R. (Director). (2018). *Period. End of Sentence*. [Film].
Thomson, J., Amery, F., Channon, M., & Puri, M. (2019). What’s missing in MHM? Moving beyond hygiene in menstrual hygiene management. *Sexual and reproductive health matters*, 27(1), 1684231. <https://doi.org/10.1080/26410397.2019.1684231>.
Umberson, D., & Karas Montez, J. (2010). Social Relationships and Health: A Flashpoint for Health Policy. *Journal of Health and Social Behavior*, 51(1_suppl), S54-S66. <https://doi.org/10.1177/0022146510383501>
UNICEF. (2021). Multidimensional poverty criteria for allocation of equalization funds to subnational governments. In *A review of the use of multidimensional poverty measures essay*. Retrieved November 27, 2023, from <https://www.unicef.org/media/105966/file/Areviewoftheuseofmultidimensionalpovertymeasures.pdf>.
UN Women. (2017). *Abolishing Chhaupadi, breaking the stigma of menstruation in rural Nepal*. UN Women. <https://www.unwomen.org/en/news/stories/2017/4/feature-abolishing-chhaupadi-breaking-the-stigma-of-menstruation-in-rural-nepal>.
Yadav, R. N., Joshi, S., Poudel, R., & Pandeya, P. (2018). Knowledge, Attitude, and Practice on Menstrual Hygiene Management among School Adolescents. *Journal of Nepal Health Research Council*, 15(3), 212–216. <https://doi.org/10.3126/jnhrc.v15i3.18842>
Yang, L. H., Kleinman, A., Link, B. G., Phelan, J. C., Lee, S., & Good, B. (2007). Culture and stigma: adding moral experience to stigma theory. *Social science & medicine* (1982), 64(7), 1524–1535. <https://doi.org/10.1016/j.socscimed.2006.11.013>.

Ramanujan-Nagell Equation and Elliptic Curves

Lale Baylar and Karin Lund
Harvard College ’28 and Dartmouth College ’29

The Ramanujan-Nagell equation $x^2 + 7 = 2^n$ is a classic Diophantine equation with integer solutions. In this paper, we explore this equation using the theory of elliptic curves and computational techniques. We transform the Ramanujan-Nagell equation into elliptic curves of the forms $y^2 = x^3 - 7$, $y^2 = 2x^3 - 7$, and $y^2 = 4x^3 - 7$, and use the Nagell-Lutz theorem to analyze torsion points and solutions. Our primary research question is: *Can we classify all integer solutions to the Ramanujan-Nagell equation using elliptic curves and computational methods?* We provide an expository presentation on the groups of rational points on elliptic curves, offering insights into their structure and properties. Computational methods using SageMath are employed to verify and explore solutions. Our work reproduces and classifies all integer solutions (2, 1) and (2, -1) for $n = 3$, (2, 3) and (2, -3) for $n = 4$, (2, 5) and (2, -5) for $n = 5$, (4, 11) and (4, -11) for $n = 7$, and (32, 181) and (32, -181) for $n = 15$, which are all solutions of the Ramanujan-Nagell equation. This analysis demonstrates the effectiveness of elliptic curves and computational techniques in solving Diophantine equations.

Introduction

In number theory, the Ramanujan-Nagell equation, named after Indian mathematician Srinivasa Ramanujan and Norwegian mathemati-cian Trygve Nagell, is a specific Diophantine equation of the form $x^2 + 7 = 2^n$, where n and x are natural numbers. This equation has fascinated mathematicians for decades due to its elegant simplicity and the surpris-ing depth of its solutions. Its study connects classical number theory to the rich theory of elliptic curves and arithmetic algebraic geometry.

This particular equation was derived by Ramanujan, and Nagell later proved that the integer solutions for x only occur when $n = 3, 4, 5, 7$, and 15 . There are many methods utilized to understand this equation. For example, in “The Diophantine Equation $x^2 + 7 = 2^n$,” published in 1973 by the American Mathematical Society, the solution described uses a direct approach to the Diophantine equation. There are also other similar equations in the literature such as $x^2 + 11 = 3^n$, or more generally, $x^2 + p = k^n$ where k, n , and x are positive integers and p is a prime integer. In this paper, we employ the theory of elliptic curves and computational techniques, particularly using SageMath, to investigate the Ramanujan-Nagell equation. Our approach combines theoretical insights with computation tools to study and classify the solutions of this equation. The goal is to identify all integer solu-tions of the Ramanujan-Nagell equation by converting the equation into elliptic curves and analyzing their rational and integer points using the Mordell-Weil Theorem, the Nagell-Lutz Theorem, and the Mazur’s Theorem. We would like to remark that some of the SageMath computations, such as determining the generator(s) of the Mordell-Weil group and computing higher multiples of the generator(s), are important in our solution. We have used the Mordell-Weil Theorem, the Nagell-Lutz Theorem, and Mazur’s Theorem to determine the tor-sion subgroup of certain elliptic curves are trivial, and the SageMath computations we have provided also verifies our computation. We learned the material and were motivated by the insights from various materials, books, and lectures from (Silverman 2015), (Silverman 2009), (Lang 2012), (Cassels 1991), and (Knapp 2005).

Elliptic Curves and Group Law
Definitions and Theorems

An elliptic curve is a smooth, symmetrical curve defined by an equation like $y^2 = x^3 + Ax^2 + Bx + C$. These curves were central to Professor Andrew Wiles’ proof of Fermat’s Last Theorem, a famous problem that puzzled mathematicians for over 350 years (Wiles 1995).

Elliptic curves are an essential object of study in modern algebraic geometry and number theory. They exhibit fascinating properties that allow for a rich interplay between geometry and arithmetic. In particular, they form a group with the addition law, making them very important in various applications, including cryptography. In what follows, we review the definitions of elliptic curves, Weierstrass form, Bézout’s Theorem, and the Cayley-Bacharach Theorem, which are important for defining the group law on an elliptic curve.

Definition 1 (Elliptic Curve) *An elliptic curve is a smooth projective algebraic curve of genus one, defined over a field \mathbb{C} , with a specific point at infinity often denoted by O . In affine coordinates, an elliptic curve can be represented by a cubic equation of the form:*

$$y^2 = x^3 + Ax^2 + Bx + C$$

where $A, B, C \in \mathbb{C}$ and the discriminant $\Delta = -4A^3C + A^2B^2 - 4B^3 - 27C^2$ is non-zero, ensuring the curve is non-singular.

Elliptic curves are a type of projective curve and have degree $d = 3$. More generally, the form of projective plane curves is defined by their degree. A projective curve includes the coordinates in a projective plane that are defined by $P(X, Y, Z)$ as zeros of a homogeneous polynomial.

Definition 2 (Weierstrass Form) *An elliptic curve in Weierstrass form is given by the equation:*

$$y^2 = x^3 + Ax + B$$

where A and B are constants, typically over a field \mathbb{C} .

The Weierstrass equation captures the essence of elliptic curves, expressing the relationship between the x -coordinate and the y -coor-dinate through a cubic equation. The smoothness condition ensures that the curve has no cusps or singularities while projectivity allows for the inclusion of points at infinity, which plays a crucial role in the arithmetic of elliptic curves.

Bézout’s and Cayley-Bacharach Theorems

Theorem 1 (Bézout for $n = 2$) *Let $f(x_0, x_1, x_2)$ and $g(x_0, x_1, x_2)$ be two homogeneous polynomials in three variables over an algebraically closed field k , with degrees d_f and d_g , respectively. If these polynomi-als have only a finite number of common zeros in the projective plane P^2 , then the number of common zeros, counted with multiplicity, is equal to the product of the degrees:*

$$d_f \cdot d_g$$

In other words, if f and g are homogeneous polynomials in P^2 and

they intersect transversally, then the number of points in the intersection set $V(f) \cap V(g)$ is $d_1 \cdot d_2$.

Note that the algebraically closed field k in the statement above can be taken to be \mathbb{C} .

Theorem 2 (Cayley-Bacharach) *Let C_1 and C_2 be two plane curves of degrees d_1 and d_2 , respectively, in the projective plane P^2 over an algebraically closed field k . Assume that C_1 and C_2 intersect in exactly $d_1 \cdot d_2$ points (counting multiplicities). Let C be a curve of degree d passing through all but one of these intersection points. Then C passes through the remaining point as well.*

Note that the algebraically closed field k in the statement above can be taken to be \mathbb{C} .

The Bézout’s Theorem and the Cayley-Bacharach Theorem will be used below to define the group law on elliptic curves.

In this paper, our focus primarily lies on exploring the properties and behavior of elliptic curves defined over the field of rational numbers, denoted by \mathbb{Q} or with coefficients in \mathbb{Z} , and consider their rational or integer solutions only. We will mention important results below on rational points on these curves, particularly examining their integer solutions and their implications.

Elliptic Curve Group Laws

Let us now provide the precise definition of the group law on elliptic curves and explore some of its properties.

Given that (E, O) is an elliptic curve over k , the group law defines how to add points on E , resulting in another point on the curve. We define addition operation on $E(k)$ as follows the follows: $+: E(k) \times E(k) \rightarrow E(k)$

- **Identity Element:** If P is the identity element (denoted as O), then $P + Q = Q$ for any point Q , and vice versa.
- **Inverse Element:** The inverse of a point P is denoted as $-P$ and is defined as the reflection of P across the x -axis.
- **Point Doubling:** If $P = Q$, then $P + P$ is computed by drawing the tangent line to the curve at P and finding its intersection with the curve. The result is then reflected across the x -axis to obtain $P + P = 2P$.
- **Point Addition:** For distinct points P and Q , the line through P and Q intersects the curve at a third point R . The sum $P + Q$ is then defined as the reflection of R across the x -axis.

These operations define a group structure on the set of points of E , with the identity element O and associativity ensured by the geometric properties of elliptic curves, which we will explain below. This group structure on an elliptic curve is fundamental in various cryptographic protocols and number theory applications. In terms of coordinates, the above-defined operation can be given as follows:

- **Identity Element:** $(0 : 1 : 0)$, point at infinity. The point at infinity O serves as the identity element since $P + O = O + P = P$ for any point P on the curve E . Geometrically, adding O to any point P results in P itself, which is consistent with the identity element property.
- **Inverse Element:** Let P be the point with coordinates $(x : y : z)$. Then its inverse is $-P$ with coordinates $(x : -y : z)$. Geometrically, the inverse of P is the reflection of P about the x -axis. Thus, $P + (-P) = O$, ensuring the existence of inverses.
- **Closure:** For any two distinct points P and Q on E , the line passing through P and Q intersects E at a third point R . Geometrically, this is guaranteed by the tangent-secant law. Thus, the sum $P + Q = R$ is well-defined and lies on E , ensuring closure.

- **Commutativity:** Let P and Q be points on the curve. Then $P + Q = Q + P$. The commutativity property $P + Q = Q + P$ follows directly from the geometric construction of the sum point $P + Q$. Regardless of the order of addition, the resulting point remains the same due to the symmetry of the elliptic curve.
- **Associativity:** To prove associativity $(P + Q) + R = P + (Q + R)$ for any points P, Q , and R on an elliptic curve E , we use properties of the curve and Cayley-Bacharach’s theorem.

Consider E as an elliptic curve given by a Weierstrass equation, where points on E satisfy this cubic equation. A fundamental property of elliptic curves is that any line intersects E in exactly three points, counting multiplicities. For points P and Q on E , the line through P and Q intersects E at a third point, say R_1 . Reflecting R_1 across the x -axis gives $-(P + Q)$, and thus $P + Q$ is defined as the reflection of R_1 , denoted by R'_1 .

Now consider points P, Q , and R on E . To prove $(P + Q) + R = P + (Q + R)$, we analyze the points of intersection produced by lines connecting these points. The line through P and Q intersects E at P, Q , and $-(P + Q)$. Similarly, the line through Q and R intersects E at Q, R , and $-(Q + R)$.

According to Bézout’s theorem, the two lines will intersect E at a total of nine points, counting multiplicities. Recall Cayley-Bacharach theorem in the special case states that if two cubic curves intersect in exactly nine points, any cubic curve passing through eight of these points also passes through the ninth. Here, the cubic curves are the elliptic curve E itself and the union of lines L_1 through P and Q , and L_2 through Q and R .

The nine points of intersection are $P, Q, R, -(P + Q), -(Q + R)$, and additional points that align with the intersection properties of cubic curves. Consider the point $(P + Q) + R$, which lies on E and aligns with the points derived from P, Q , and $-(P + Q)$. Similarly, the point $P + (Q + R)$ lies on E and aligns with Q, R , and $-(Q + R)$. By Cayley-Bacharach, since we can choose eight points on E (out of these nine intersection points) that satisfy the conditions, the ninth point must also lie on the curve.

Hence, both $(P + Q) + R$ and $P + (Q + R)$ must lie on E through the Cayley Bacharach theorem’s implication. Therefore, $(P + Q) + R$ and $P + (Q + R)$ must be the same point, proving associativity. Using Cayley-Bacharach’s theorem that we have included below and applying it to the intersection properties of cubic curves, we establish that the points resulting from the associative addition on the elliptic curve E must coincide, confirming $(P + Q) + R = P + (Q + R)$.

This completes the proof of the group law on an elliptic curve.

Mordell’s Theorem, Mazur’s Theorem, Nagell-Lutz Theorem

The Mordell-Weil theorem states that the rational points on an elliptic curve can be built from a finite set of points using addition. The Mordell-Weil theorem provides insights into the distribution of rational points on abelian varieties, including elliptic curves, over number fields. It states that the group of K -rational points on an abelian variety A defined over a number field K is finitely generated.

Specifically, for an elliptic curve E defined over K , the group $E(K)$ of K rational points, which is known as the Mordell-Weil group, can be expressed as the direct sum of a finite torsion subgroup and a free abelian group of finite rank. In mathematical terms:

$$E(K) \cong \mathbb{Z}^r \oplus T$$

where r is the rank of the elliptic curve and T is a finite torsion subgroup.

Theorem 3 (Mordell-Weil) *Let E be an elliptic curve defined over a*

number field K . Then, the group of rational points $E(K)$ is finitely generated and can be expressed as the direct sum of a torsion subgroup $E(K)_{tors}$ and a free abelian group \mathbb{Z}^r , where r is the rank of $E(K)$.

Mazur’s theorem provides a classification of the possible torsion subgroups of elliptic curves over the rationals. This theorem gives a precise list of the possible structures that the torsion subgroup can take for any elliptic curve defined over \mathbb{Q} .

Theorem 4 (Mazur’s Theorem) *If E is an elliptic curve defined over the rational numbers \mathbb{Q} , then torsion subgroup $E(\mathbb{Q})_{tors}$ is isomorphic to one of the following 15 groups:*

$$\mathbb{Z}/n\mathbb{Z} \text{ or } \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2m\mathbb{Z}$$

where n and m are positive integers satisfying $1 \leq n \leq 10$ or $n = 12$, and $1 \leq m \leq 4$.

We will use Mazur’s Theorem and Nagell-Lutz Theorem to determine the torsion subgroups of a certain family of elliptic curves below.

The Nagell-Lutz Theorem is a fundamental result in the theory of elliptic curves and plays a significant role in the study of rational points on these curves. It provides a way to find the integer points of finite order (torsion points) on an elliptic curve. The torsion points are points on an elliptic curve with finite order, meaning they return to identity after a finite number of self additions. The following Nagell-Lutz helps identify these points by checking if the y -coordinate divides the discriminant of the elliptic curve.

Theorem 5 (Nagell-Lutz Theorem) *If E is an elliptic curve given by the equation $y^2 = x^3 + Ax + B$, where A and B are integers, and if $P = (x, y)$ is a rational point of finite order on E , then either:*

1. *$y = 0$, in which case the point $P = (x, y)$ has order two, or*
2. *y is an integer that divides $D = 4A^3 + 27B^2$, which immediately implies that y^2 divides D , which is the discriminant of the elliptic curve E .*

It follows from the Nagell-Lutz Theorem that, if a rational point $P = (x, y)$ with $y \neq 0$ lies on an elliptic curve, either it corresponds to an integral solution of the equation defining the curve or it satisfies a special divisibility condition related to the coefficients of the curve equation.

Rational Points on Elliptic Curves

Let E be an elliptic curve defined over a field k , given by the Weierstrass equation

$$E : y^2 = x^3 + Ax + B,$$

where $A, B \in k$ and the discriminant $\Delta = -16(4A^3 + 27B^2) \neq 0$. The set of rational points on E over k , denoted $E(k)$, consists of solutions $(x, y) \in k \times k$ that satisfy the Weierstrass equation. Our field k will be \mathbb{C}, \mathbb{R} , or \mathbb{Q} , but the formulas below hold for general field k .

Addition Formula

Given two rational points $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ on E , the formula for the sum $P + Q$ is given by the following:

$$x_3 = \lambda^2 - x_1 - x_2,$$

$$y_3 = \lambda(x_1 - x_3) - y_1,$$

where $\lambda = (y_2 - y_1)/(x_2 - x_1)$ if $P \neq Q$, and $\lambda = (3x_1^2 + A)/2y_1$ if $P = Q$.

Doubling Formula

Using the above, the doubling formula for a rational point $P = (x_1, y_1)$ on E is given by:

$$x_3 = \lambda^2 - 2x_1,$$

$$y_3 = \lambda(x_1 - x_3) - y_1,$$

where $\lambda = (3x_1^2 + A)/2y_1$.

In what follows, we will apply the above formulas to analyze the groups of rational points of some examples.

Example 1 Let us consider the elliptic curve E_1 defined by the equation $Y^2Z = X^3 - 4XZ^2 + 4Z^3$. To check for nonsingularity of E_1 , we analyze the partial derivatives:

$$\partial F/\partial X = 3X^2 - 4Z^2, \partial F/\partial Y = 2YZ, \partial F/\partial Z = Y^2 + 8XZ + 12Z^2$$

Setting each partial derivative equal to zero and solving the resulting system of equations, we find the following solutions:

$$\partial F/\partial X = 0 \rightarrow 3X^2 - 4Z^2 = 0 \rightarrow X^2 = 4Z^2/3$$

$$\partial F/\partial Y = 0 \rightarrow Y = 0 \text{ or } Z = 0$$

$$\partial F/\partial Z = 0 \rightarrow Y^2 + 8XZ + 12Z^2 = 0.$$

The first equation implies that either $X = \pm 2Z/\sqrt{3}$ or $X = 0$. The second equation indicates that either $Y = 0$ or $Z = 0$. The third equation yields a quadratic equation in X, Y , and Z , which can be solved for Y^2 to obtain solutions in terms of X and Z . By further analysis, we find that the only common solutions satisfying all three equations are when $X = 0, Y = 0$, and $Z = 0$, which can not be on our curve because it can not be on a projective space. Thus, there are no nontrivial common solutions. Therefore, E_1 is nonsingular.

Next, we will consider the torsion rational points on elliptic curve E_1 . To compute the torsion points, we first transform the equation into its Weierstrass form: $y^2 = x^3 - 4x + 4$, and then apply the Nagell-Lutz theorem to find the torsion points. Recall that Nagell-Lutz Theorem states that if (x, y) is a rational point on an elliptic curve E given by $y^2 = x^3 + ax + b$, where $a, b \in \mathbb{Z}$, then either $y = 0$ or y^2 is a perfect square that divides D , and x and y are integers with $|x| \leq \max(1, |2b|)$.

In our case, the Weierstrass form of E_1 is $y^2 = x^3 - 4x + 4$. Here, $a = -4$ and $b = 4$. We have either $y = 0$ or y^2 divides $D = 4^2 \cdot 11$. It follows that we have one of the following possibilities for y : $y = 0, \pm 1, \pm 2, \pm 4$. If $y = 0$, then $x^3 - 4x + 4 = 0$, which has no integer solutions. If $y = \pm 1$, then $x^3 - 4x + 4 = 1$, which has solution $(x, y) = (1, \pm 1)$. If $y = \pm 2$, then $x^3 - 4x + 4 = 4$, which has three solutions $(x, y) = (0, \pm 2), (x, y) = (-2, \pm 2), (x, y) = (2, \pm 2)$. If $y = \pm 4$, then $x^3 - 4x + 4 = 16$, which has no integer solutions.

Thus, it follows that the torsion subgroup $E_1(\mathbb{Q})_{tors}$ is trivial, meaning it contains only the identity element. We have also provided SAGE computation verifying the triviality of the torsion subgroup $E_1(\mathbb{Q})_{tors}$ and computed its rank (**S1**).

Example 2 Let C be the following non-singular cubic curve with the form $y^2 = x^3 + ax^2 + bx + c$:

$$C: y^2 = f(x) = x^3 + 1.$$

Let us take points $P = (x_1, y_1) = (-1, 0)$ and $Q = (x_2, y_2) = (2, 3)$ on curve C . Using the formulas from (1), we have that:

$$\lambda = (y_2 - y_1)/(x_2 - x_1) = (3 - 0)/(2 - (-1)) = 1$$

$$x_3 = \lambda_2 - x_1 - x_2 = 12 - (-1) - 2 = 0$$

$$y_3 = \lambda(x_1 - x_3) - y_1 = 1((-1) - 0) - 0 = -1$$

$$P + Q = (x_3, y_3) = (0, -1)$$

Using point $R = (x_2, y_2) = (2, 3)$ on our previous curve C and the formula

$$\lambda = dy/dx = f'(x)/2y$$

we obtain:

$$\lambda = 3x^2/2y = 3(2)^2/2(3) = 2.$$

Using the fact that $\lambda = 2$ and the equations from (2), we have:

$$x_4 = \lambda^2 - 2x_2 = 2^2 - 2(2) = 0$$

$$y_4 = \lambda \cdot (x_2 - x_4) - y_2 = 2(2 - 0) - 3 = 1$$

To continue with point doubling with curve C : $y^2 = x^3 + 1$, we

can conclude whether (2, 3) is a point of finite order. To do so, we'll use the methods of point doubling and point addition:

We left off on the point (x₄, y₄) = (0, 1) in our previous example. Adding (0, 1) and (2, 3), we have:

$$\begin{aligned}\lambda &= (3 - 1)/(2 - 0) = 1 \\ x_5 &= \lambda^2 - 0 - 2 = -1 \\ y_5 &= 1 \cdot (0 - (-1)) - 1 = 0\end{aligned}$$

Now we have the point (x₅, y₅) = (-1, 0). Once again, we repeat the process above:

$$\begin{aligned}\lambda &= (3 - 0)/(2 - (-1)) = 1 \\ x_6 &= \lambda^2 - (-1) - 2 = 0 \\ y_6 &= 1(-1 - 0) - 0 = -1\end{aligned}$$

We can also see this from our previous example of point addition above. Now, we have point (x₆, y₆) = (0, -1).

$$\begin{aligned}\lambda &= (3 - (-1))/(2 - 0) = 2 \\ x_7 &= \lambda^2 - 0 - 2 = 2 \\ y_7 &= 2(0 - 2) - (-1) = -3\end{aligned}$$

Now, we have the point (x₇, y₇) = (2, -3). The example concludes here because, when attempting to compute our next point, the denominator of λ will be zero. Thus, we stop calculating points here, as we've arrived at the origin point.

We have also provided SAGE computation verifying the torsion subgroup E₁(\mathbb{Q})_{tors} and computed its rank (S2).

Example 3 For this example, let our curve be C : y² = x³ + 8. Using the following discriminant formula:

$$D = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2,$$

we have that the discriminant of this curve is -1728 since a = 0, b = 0, and c = 8. Thus, the possibilities for the y-coordinate are:

$$\{0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 8, \pm 12, \pm 24\}$$

Then, we can calculate the x-coordinates for each of these y-values:

$$\begin{aligned}0 &= x^3 + 8 \rightarrow x = -2 \\ 1 &= x^3 + 8 \rightarrow x = -7^{1/3} \\ 4 &= x^3 + 8 \rightarrow x = -4^{1/3} \\ 9 &= x^3 + 8 \rightarrow x = 1 \\ 16 &= x^3 + 8 \rightarrow x = 2 \\ 36 &= x^3 + 8 \rightarrow x = 28^{1/3} \\ 64 &= x^3 + 8 \rightarrow x = 56^{1/3} \\ 144 &= x^3 + 8 \rightarrow x = 136^{1/3} \\ 576 &= x^3 + 8 \rightarrow x = 568^{1/3}\end{aligned}$$

(Note that the sign of y does not alter the x-coordinates)

Note that the point (-2, 0) is a torsion point, and in fact it generates the torsion subgroup of the elliptic curve C: y² = x³ + 8.

The x-coordinates are integers when y = 3 and y = 4. Thus, our possible points are P₁ = (1, 3) and P₂ = (2, 4). Next, calculating 2P₁ and 2P₂, we compute that the x-coordinates of 2P₁ and 2P₂ are not integers. Thus, they are not points of finite order, meaning they are of infinite order. In fact, any of these two points generates an infinite part of the Mordell-Weil group of the elliptic curve C: y² = x³ + 8.

We have also provided SAGE computation verifying the torsion subgroup E_C(\mathbb{Q})_{tors} $\cong \mathbb{Z}_2$ and computed its rank (S3).

Solving Ramanujan-Nagell Equation Using Elliptic Curves and SAGE Computations

In this section, we consider the Ramanujan-Nagell equation y² + 7 = 2ⁿ (*)

and provide its solution using the theory of elliptic curves:

To solve this equation, let us first rewrite (*) as follows to identify its solutions among the integer points of one of three elliptic curves.

We will consider the cases n = 3k, n = 3k + 1, and n = 3k + 2 separately.

For n = 3k:

$$y^2 + 7 = (2^k)^3 \rightarrow y^2 = x^3 - 7 \quad \textbf{Case (1)}$$

For n = 3k + 1:

$$y^2 + 7 = 2(2^k)^3 \rightarrow y^2 = 2x^3 - 7 \quad \textbf{Case (2)}$$

For n = 3k + 2:

$$y^2 + 7 = 4(2^k)^3 \rightarrow y^2 = 4x^3 - 7 \quad \textbf{Case (3)}$$

where in each case above, we set x = 2^k.

Recall that the set of rational solutions on these elliptic curves forms a finitely generated abelian group of finite rank. Furthermore, this group is the direct sum of the subgroups of points of finite order (the torsion subgroup) and the subgroup of points of infinite order, according to the Mordell-Weil Theorem:

$$E(\mathbb{Q}) = E_{\text{finite}} \oplus E_{\text{infinite}}.$$

Remark: While we will initially apply the Nagell-Lutz Theorem to identify potential torsion points with integer coordinates on the elliptic curves under consideration, it is crucial to emphasize the role of Mazur's Theorem in concluding the structure of the torsion subgroup. The Nagell-Lutz Theorem provides necessary conditions for a point to be torsion: namely, that the coordinates are integers and that the y-coordinate (when non-zero) divides the discriminant of the curve. This allows us to generate a finite list of candidate torsion points, such as P = (2, 1) and Q = (2, -1) and 2P = (32, 181) and 2Q = (32, -181) on the curve E : y² = x³ - 7 (see below).

However, determining whether these points are indeed of finite order requires further analysis. Here, Mazur's Theorem plays a key role. It classifies all possible torsion subgroups of elliptic curves over the rational numbers \mathbb{Q} and states that the torsion subgroup E(\mathbb{Q})_{tors} must be isomorphic to one of the following fifteen groups: $\mathbb{Z}/n\mathbb{Z}$ for 1 ≤ n ≤ 10 or n = 12, or $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2m\mathbb{Z}$ for 1 ≤ m ≤ 4.

By computing several multiples of P using SageMath (from 2P up to 12P and 2Q up to 12Q), and observing that none of these yield the identity point O or repeat earlier points, we conclude that P and Q do not lie in any of the allowable finite torsion subgroups classified by Mazur's Theorem. Thus, P and Q must be points of infinite order.

We will first find the points of finite order using the Nagell-Lutz Theorem, considering each case:

Case (1): No Torsion → No Points of Finite Order

We use the Nagell-Lutz Theorem to find finite order points on the elliptic curve y² = x³ - 7. In fact, using the Nagell-Lutz Theorem, Mazur's Theorem, and some SageMath computations, we show the torsion subgroup of this elliptic curve is trivial. First, we check whether the curve has points of order 2:

$$x^3 - 7 = 0 \rightarrow x = 7^{1/3} \notin \mathbb{Z}$$

Since the equation above does not have an integer solution, it follows that the elliptic curve does not have a point of order 2. Next, we compute the discriminant of the elliptic curve y² = x³ - 7 using the formula below. We then apply the Nagell-Lutz Theorem to find all torsion points of this elliptic curve.

$$D = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2$$

where a = 0, b = 0, and c = -7:

$$D = 0 + 0 + 0 + 0 - 27(-7)^2 = -1323 = -3^3 \cdot 7^2.$$

It follows that the possible candidates for y are:

$$\{\pm 1, \pm 3, \pm 7, \pm 21\}$$

We then calculate the x-coordinates for these y-values:

$$\begin{aligned}1 &= x^3 - 7 \rightarrow x = 8^{1/3} \notin \mathbb{Z} \\ 9 &= x^3 - 7 \rightarrow x = 16^{1/3} \notin \mathbb{Z} \\ 49 &= x^3 - 7 \rightarrow x = 56^{1/3} \notin \mathbb{Z}\end{aligned}$$

$$441 = x^3 - 7 \rightarrow x = 448^{1/3} \notin \mathbb{Z}$$

We get integer x-coordinates when y = ±1. Thus, our possible points are P₁ = (2, 1) and P₂ = (2, -1). To ensure that these candidates are solutions to the original equation (*), we must check that the x-coordinate is a power of 2, which it is (2 = 2¹). We use SageMath commands in S4 to check whether P₁ and P₂ are points of finite or infinite order, determine the torsion subgroup, the rank of the given elliptic curve, and the generators of the Mordell-Weil group.

Using the SageMath computation (S4), we see that nP_i are all different points for any 1 ≤ n ≤ 12, and we have -P₁ = P₂. From Mazur's Theorem, it follows that the torsion subgroup of this elliptic curve consists of only the identity element. This has also been verified by the SageMath commands (S4). Since there are no torsion points on the elliptic curve y² = x³ - 7, P₁ and P₂ must be points of infinite order. Furthermore, the SageMath commands in S4 also verify that P₁ (and P₂) are points of infinite order, and they are each generators of the Mordell-Weil group, which is isomorphic to \mathbb{Z} .

Since the elliptic curve in Case (1) does not have any torsion points, we can only use the points of infinite order, the generators P₁ and P₂ of the Mordell-Weil group, to potentially identify more integer solutions on this elliptic curve. Let us take points P = P₁ = (2, 1) (or equally Q = P₂ = (2, -1), which is the same as the point -P). Knowing that P and Q are of infinite order and since they both are generators of the Mordell-Weil group, we can use them to generate more solutions. Using SageMath, we compute 2P, 2Q, 3P, 3Q, 4P, 4Q, ... , 12P, 12Q.

We find additional solutions: 2P = (32, 181) and 2Q = (32, -181). Since nP and nQ (|n| > 2) involve fractions, we will show that Case (1) does not have any additional integer solutions.

Theorem 6 Let E/ \mathbb{Q} be the elliptic curve defined by y² = x³ - 7, and let P = (2, 1) denote a generator of the Mordell-Weil group E(\mathbb{Q}) $\cong \mathbb{Z}$. Then the integral points on E satisfy the following:

- The complete set of integral points is E(\mathbb{Z}) = {O, (2, ±1), (32, ±181)}.
- For all integers n with |n| ≥ 3 the multiple nP does not have integral coordinates; that is, nP $\notin \mathbb{Z}^2$ for all |n| ≥ 3.

Proof. (I) By Siegel's theorem, the set E(\mathbb{Z}) is finite. Moreover, Baker's method using linear forms in logarithms yields effective bounds for the size of integral points. In particular, Baker (1968) showed that if P generates E(\mathbb{Q}), then there exists an explicit constant C such that any integral point (x, y) satisfies

$$\max(|x|, |y|) < \exp(C),$$

where C depends on the canonical height of P and the minimal Weierstrass model. This result was further refined by Baker-Coates and later by Bugeaud, Mignotte, and Siksek.

For the curve E : y² = x³ - 7, computations (using Magma or Sage) show that C < 10¹⁰, allowing one to exhaustively search for integral points by checking nP for small values of |n|. Since P = (2, 1) generates E(\mathbb{Q}), every rational point on E is of the form nP for some n $\in \mathbb{Z}$. A direct computation yields:

$$\begin{aligned}P &= (2, 1), \\ -1 \cdot P &= (2, -1), \\ 2P &= (32, 181), \\ -2P &= (32, -181).\end{aligned}$$

All other multiples nP with |n| < 100 yields non-integral coordinates, and by the aforementioned bound, no further integral points exist. Hence, E(\mathbb{Z}) = {O, (2, ±1), (32, ±181)}.

(II) Suppose |n| ≥ 3. We present two independent arguments.

Height argument: The canonical Néron-Tate height satisfies ĥ(nP) = n²ĥ(P), where ĥ(P) ≈ 0.76009. From Silverman's inequality (see Advanced Topics in the Arithmetic of Elliptic Curves, Thm. 1.1), we have:

$$\log|x(nP)| \geq 2\hat{h}(nP) - C_E = 2n^2 \hat{h}(P) - C_E,$$

where C_E is an explicit constant depending on E. Since ĥ(P) > 0, this lower bound grows quadratically in n, showing that |x(nP)| becomes arbitrarily large as |n| increases. Thus, integrality of x(nP) is ruled out for large n, unless exceptional cancellation occurs.

Division polynomial argument: Let x(nP) = Φ_n(2)/Ψ_n²(2), where Φ_n and Ψ_n are the classical division polynomials. The denominator Ψ_n²(2) is a square of an integer, and its prime factors are governed by a recurrence:

$$\Psi_{m+2}(\Psi_{m-1})^2 \equiv \Psi_{m-2}(\Psi_{m+1})^2 \pmod{p}$$

valid for any prime p. In particular, for all |n| ≥ 3, one finds that Ψ_n(2) has nontrivial prime divisors. Moreover, Φ_n(2) and Ψ_n²(2) are coprime in \mathbb{Z} , since their zero sets correspond to distinct algebraic loci.

Hence, the denominator of x(nP) cannot cancel, so x(nP) is not an integer. This can be checked explicitly by computing valuations at bad primes. Thus, nP $\notin \mathbb{Z}^2$ for all |n| ≥ 3.

n	ĥ(nP)
1	0.76009
2	3.04036
3	6.84084
4	12.16150
5	19.00234
6	27.36338

The table above confirms the quadratic growth of ĥ(nP) = n²ĥ(P) and supports the claim that integral points cannot occur for |n| ≥ 3.

We next use the integer points obtained to find the solutions of the Ramanujan-Nagell equation:

For P = (2, 1) and Q = (2, -1), we can set x = 2 and y = 1 or y = -1 in Case (1)'s equation for y² + 7 = (2^k)³ with n = 3k :

$$(\pm 1)^2 + 7 = (2^k)^3 \rightarrow k = 1 \rightarrow n = 3$$

Similarly, for 2P = (32, 181) and 2Q = (32, -181):

$$(\pm 181)^2 + 7 = (2^k)^3 \rightarrow k = 5 \rightarrow n = 15$$

The integer solutions to the equation y² + 7 = 2ⁿ in this case are:

$$(2, 1), (2, -1), (32, 181), (32, -181)$$

The points (2, 1) and (2, -1) correspond to n = 3, and the points (32, 181) and (32, -181) correspond to n = 15.

Case (2): No Torsion Points

We use the Nagell-Lutz Theorem to find finite order points on the elliptic curve y² = 2x³ - 7. Similarly, using the Nagell-Lutz Theorem, Mazur's Theorem, and some SageMath computations, we show the torsion subgroup of this elliptic curve is also trivial. First, we check whether the curve has points of order 2:

$$2x^3 - 7 = 0 \rightarrow x = (7/2)^{1/3} \notin \mathbb{Z}.$$

Thus, there is no point of order 2. Next, we compute the discriminant of y² = 2x³ - 7 and determine the torsion points:

$$4y^2 = 8x^3 - 28 \rightarrow (2y)^2 = (2x)^3 - 28.$$

Letting y' = 2y and x' = 2x, we have:

$$(y')^2 = (x')^3 - 28.$$

The discriminant is:

$$D = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2$$

where a = 0, b = 0, and c = -28:

$$D = -27(-28)^2 = -21168 = -2^4 \cdot 3^3 \cdot 7^2.$$

The possible candidates for y' are:
 $\{\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 7, \pm 14, \pm 21, \pm 42\}$
We calculate x-coordinates for these y-values:
 $1 = (x')^3 - 28 \rightarrow x' = 29^{1/3},$
 $4 = (x')^3 - 28 \rightarrow x' = 32^{1/3},$
 $9 = (x')^3 - 28 \rightarrow x' = 37^{1/3},$
 $16 = (x')^3 - 28 \rightarrow x' = 44^{1/3},$
 $36 = (x')^3 - 28 \rightarrow x' = 4,$
 $49 = (x')^3 - 28 \rightarrow x' = 77^{1/3},$
 $196 = (x')^3 - 28 \rightarrow x' = 224^{1/3},$
 $441 = (x')^3 - 28 \rightarrow x' = 469^{1/3},$
 $1764 = (x')^3 - 28 \rightarrow x' = 1792^{1/3}.$

We get integer coordinates when $y' = \pm 6$. Thus, the possible points are $P_1 = (4, 6)$ and $P_2 = (4, -6)$. We must ensure that these candidates satisfy the original equation (*) with x-coordinate a power of 2, which it is ($4 = 2^2$). Again, we use SageMath commands in **S5** to check whether P_1 and P_2 are points of finite or infinite order, determine the torsion subgroup, the rank of the given elliptic curve, and the generators of the Mordell-Weil group.
Using the SageMath computation (**S5**), we see that nP_1 are all different points for any $1 \leq n \leq 15$, and we have $-P_1 = P_2$. From Mazur’s Theorem, it follows that the torsion subgroup of this elliptic curve consists of only the identity element. This has also been verified by the SageMath commands (**S5**). Since there are no torsion points on the elliptic curve $(y')^2 = (x')^3 - 28$, P_1 and P_2 must be points of infinite order. Furthermore, the SageMath commands in **S5** also verify that P_1 (and P_2) are points of infinite order, and they are each generators of the Mordell-Weil group, which is isomorphic to \mathbb{Z} .
Since the elliptic curve in Case (2) does not have any torsion points, we can only use the points of infinite order, the generators P_1 and P_2 of the Mordell-Weil group, to potentially identify more integer solutions on this elliptic curve. Let us take points $P = P_1 = (4, 6)$ (or equally $Q = P_2 = (4, -6)$, which is the same as the point $-P$). Knowing that P and Q are of infinite order and since they both are generators of the Mordell-Weil group, we can use them to generate more solutions. Using SageMath, we compute $2P, 2Q, 3P, 3Q, 4P, 4Q, \dots, 12P, 12Q$.
We find additional solutions: $2P = (8, 22), 2Q = (8, -22), 3P = (37, 225),$ and $3Q = (37, -225)$. Since nP and nQ ($|n| > 3$) involve fractions, we can similarly show that Case (2) does not have any additional integer solutions. We have that the following theorem holds:

Theorem 7 *Let E/\mathbb{Q} be the elliptic curve defined by $y^2 = x^3 - 28$, and let $P = (4, 6)$ denote a generator of the Mordell-Weil group $E(\mathbb{Q}) \cong \mathbb{Z}$. Then the integral points on E satisfy the following:*

- The complete set of integral points is $E(\mathbb{Z}) = \{O, (4, \pm 6), (8, \pm 22), (37, \pm 225)\}$.*
- For all integers n with $|n| \geq 4$, the multiple nP does not have integral coordinates; that is, $nP \notin \mathbb{Z}^2$ for all $|n| \geq 4$.*

Proof. The proof is similar to the proof of Theorem 6 above.
We verify the results:
For $P = (4, 6)$ and $Q = (4, -6)$, we can set $x' = 4$ and $y' = 6$ or $y = -6$ in Case (2)’s equation for $y^2 + 7 = 2(2^k)^3$ with $n = 3k + 1$.
Because $x' = 2x$ and $y' = 2y$, we have that $x = 2$ and $y = 3$ or $y = -3$:
 $(\pm 3)^2 + 7 = 2(2^k)^3 \rightarrow k = 1 \rightarrow n = 4$
For $2P = (8, -22)$ and $2Q = (8, -22)$, we can set $x' = 8$ and $y' = 22$ or $y = -22$.
Because $x' = 2x$ and $y' = 2y$, we have that $x = 4$ and $y = 11$ or $y = -11$:
 $(\pm 11)^2 + 7 = 2(2^k)^3 \rightarrow k = 2 \rightarrow n = 7$

For $3P = (37, 225)$ and $3Q = (37, -225)$, we can set $x' = 37$ and $y' = 225$ or $y = -225$.
Because $x' = 2x$ and $y' = 2y$, we have that $x = 18.5$ and $y = 112.5$ or $y = -112.5$.
Since y is not an integer, we cannot find a power of 2 for which the equation $y^2 + 7 = 2(2^k)^3$ is solvable. Therefore, $(37, 225)$ and $(37, -225)$ are not solutions.
The integer solutions to the equation $y^2 + 7 = 2^n$ in this case are:
 $(2, 3), (2, -3), (4, 11), (4, -11)$.
The points $(2, 3)$ and $(2, -3)$ correspond to $n = 4$, and the points $(4, 11)$ and $(4, -11)$ correspond to $n = 7$.

Case (3): No Torsion
Similarly, we use the Nagell-Lutz Theorem to find finite order points on the elliptic curve. Again, using the Nagell-Lutz Theorem, Mazur’s Theorem, and some SageMath computations, we determine the torsion subgroup of this elliptic curve to be trivial. We first check if the elliptic the curve $y^2 = 4x^3 - 7$ for points of order 2:
 $4x^3 - 7 = 0 \rightarrow x = (7/4)^{1/3} \notin \mathbb{Z}$.
Thus, there is no point of order 2.
Letting $y' = 4y$ and $x' = 4x$, we have:
 $(y')^2 = (x')^3 - 112$
We find the discriminant of $(y')^2 = (x')^3 - 112$:
 $D = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2$
where $a = 0, b = 0$, and $c = -112$:
 $D = -27(-112)^2 = -338688 = -2^8 \cdot 3^3 \cdot 7^2$.
The possible candidates for torsion $(y')^2$ are:
 $\{1, 4, 9, 16, 36, 49, 64, 144, 196, 256, 441, 576, 784, 1764, 12544, 28224, 112896\}$.

We can then calculate the x' -coordinates for each of these y' -values:
 $1 = (x') - 112 \rightarrow x' = 113^{1/3},$
 $4 = (x') - 112 \rightarrow x' = 116^{1/3},$
 $9 = (x') - 112 \rightarrow x' = 121^{1/3},$
 $16 = (x') - 112 \rightarrow x' = 128^{1/3},$
 $36 = (x') - 112 \rightarrow x' = 148^{1/3},$
 $49 = (x') - 112 \rightarrow x' = 161^{1/3},$
 $64 = (x') - 112 \rightarrow x' = 176^{1/3},$
 $144 = (x') - 112 \rightarrow x' = 256^{1/3},$
 $196 = (x') - 112 \rightarrow x' = 308^{1/3},$
 $256 = (x') - 112 \rightarrow x' = 368^{1/3},$
 $441 = (x') - 112 \rightarrow x' = 553^{1/3},$
 $576 = (x') - 112 \rightarrow x' = 688^{1/3},$
 $784 = (x') - 112 \rightarrow x' = 4314^{1/3},$
 $1764 = (x') - 112 \rightarrow x' = 1876^{1/3},$
 $7056 = (x') - 112 \rightarrow x' = 8167^{1/3},$
 $12544 = (x') - 112 \rightarrow x' = 12656^{1/3},$
 $28224 = (x') - 112 \rightarrow x' = 28336^{1/3},$
 $112896 = (x') - 112 \rightarrow x' = 113008^{1/3},$

We get integer coordinates when $y' = \pm 20$. Thus, the possible points are $P_1 = (8, 20)$ and $P_2 = (8, -20)$. We must ensure these points satisfy the original equation (*) with x-coordinate a power of 2. Again, we use SageMath commands in **S6** to check whether P_1 and P_2 are points of finite or infinite order, determine the torsion subgroup, the rank of the given elliptic curve, and the generators of the Mordell-Weil group.
Using the SageMath computation (**S6**), we see that nP_1 are all different points for any $1 \leq n \leq 15$, and we have $-P_1 = P_2$. From Mazur’s Theorem, it follows that the torsion subgroup of this elliptic curve consists of only the identity element. This has also been verified by the SageMath commands (**S6**). Since there are no torsion points on the

elliptic curve $(y')^2 = (x')^3 - 112$, P_1 and P_2 must be points of infinite order. Furthermore, the SageMath commands above also verify that P_1 (and P_2) are points of infinite order, and they are each generators of the Mordell-Weil group, which is isomorphic to \mathbb{Z} .
Since the elliptic curve in Case (3) does not have any torsion points, we can only use the points of infinite order, the generators P_1 and P_2 of the Mordell-Weil group, to potentially identify more integer solutions on this elliptic curve. Let us take points $P = P_1 = (8, 20)$ (or equally $Q = P_2 = (8, -20)$, which is the same as the point $-P$). Knowing that P and Q are of infinite order and since they both are generators of the Mordell-Weil group, we can use them to generate more solutions. Using SageMath, we compute $2P, 2Q, 3P, 3Q, 4P, 4Q, \dots, 12P, 12Q$.
Since nP and nQ ($|n| > 1$) involve fractions, we have that Case (3) does not have any additional integer solutions.

Theorem 8 *Let E/\mathbb{Q} be the elliptic curve defined by $y^2 = x^3 - 112$, and let $P = (8, 20)$ denote a generator of the Mordell-Weil group $E(\mathbb{Q}) \cong \mathbb{Z}$. Then the integral points on E satisfy the following:*

- The complete set of integral points is $E(\mathbb{Z}) = \{O, (8, \pm 20)\}$.*
- For all integers n with $|n| \geq 2$, the multiple nP does not have integral coordinates; that is, $nP \notin \mathbb{Z}^2$ for all $|n| \geq 2$.*

Proof. The proof is similar to the proof of Theorem 6 above.
We verify the results:
For $P = (8, 20)$ and $Q = (8, -20)$, we can set $x' = 8$ and $y' = 20$ or $y = -6$ in Case (3)’s equation for $y^2 + 7 = 4(2^k)^3$ with $n = 3k + 2$.
Since $x' = 4x$ and $y' = 4y$ for Case 3, we have $x = 2$ and $y = 5$ or $y = -5$:
 $(\pm 5)^2 + 7 = 4(2^k)^3 \rightarrow k = 1 \rightarrow n = 5$
The integer solutions to the equation $y^2 + 7 = 2^n$ in this case are:
 $(2, 5), (2, -5)$
The points $(2, 5)$ and $(2, -5)$ correspond to $n = 5$.

Conclusion
In this study, we have explored the Ramanujan-Nagell equation $y^2 + 7 = 2^n$ using the theory of elliptic curves and SageMath computations. We analyzed three distinct cases by converting the equation into elliptic curves of the forms $y^2 = x^3 - 7, y^2 = 2x^3 - 7$, and $y^2 = 4x^3 - 7$.
For each case, we employed the Nagell-Lutz Theorem to investigate possible torsion points, confirming that none of the elliptic curves have torsion points of order 2. By analyzing the torsion subgroups and computing solutions of infinite order using SageMath, we identified all possible integer solutions to the equation. The comprehensive results are summarized as follows:

- For the elliptic curve $y^2 = x^3 - 7$:
 - Solutions include $(2, 1)$ and $(2, -1)$ and $(32, 181)$ and $(32, -181)$, corresponding to the $n = 3$ and $n = 15$ cases, respectively, of the Ramanujan-Nagell equation.
- For the elliptic curve $y^2 = 2x^3 - 7$:
 - Solutions include $(4, 6)$ and $(4, -6)$ and $(8, 22)$ and $(8, -22)$, corresponding to the $n = 4$ and $n = 7$ cases, respectively, of the Ramanujan-Nagell equation.
- For the elliptic curve $y^2 = 4x^3 - 7$:
 - Solutions include $(8, 20)$ and $(8, -20)$ corresponding to the $n = 5$ case of the Ramanujan-Nagell equation.

Combining results from all three cases, all the integer solutions to the equation $y^2 + 7 = 2^n$ are:
 $(2, 1), (2, -1), (2, 3), (2, -3), (2, 5), (2, -5), (4, 11), (4, -11), (32, 181), (32, -181),$

where $(2, 1)$ and $(2, -1)$ correspond to $n = 3, (2, 3)$ and $(2, -3)$ correspond to $n = 4, (2, 5)$ and $(2, -5)$ correspond to $n = 5, (4, 11)$ and $(4, -11)$ correspond to $n = 7$, and $(32, 181)$ and $(32, -181)$ correspond to $n = 15$. The integer solutions to $y^2 + 7 = 2^n$ are:

x	y	n
2	±1	3
2	±3	4
2	±5	5
4	±11	7
32	±181	15

Our results confirm all the solutions and demonstrate the effectiveness of combining theory and computation.
Appendix
Supplementary materials are available online at thurj.org.

References
Anthony W. Knapp, Elliptic Curves, *Mathematical Surveys and Monographs*, 2005, Volume 40.
Cassels, J. W. S. (1991). Lectures on elliptic curves. *London Mathematical Society Student Texts*, vol. 24. Cambridge University Press, Cambridge. George Green Library, QA565 CAS.
D. Zagier, Large Integral Points on Elliptic Curves, *Mathematics of Computation*, Volume 48, Number 177, January 1987, Pages 425–436.
Ebbinghaus, H.-D., Hermes, H., Hirzebruch, F., Koecher, M., Mainzer, K., Neukirch, J., Prestel, A., and Remmert, R. (1991). Numbers. *Graduate Texts in Mathematics*, vol. 123. Springer-Verlag.
Eisenbud, D., Green, M., and Harris, J. (1996). Cayley-Bacharach theorems and conjectures. *Bulletin of the American Mathematical Society*, 33(3), 295–324.
Ireland, K., and Rosen, M. (1990). A classical introduction to modern number theory. Second ed. *Graduate Texts in Mathematics*, vol. 84. Springer Verlag, New York.
Lemmermayer, F. (2018). Elliptic curves, historical remarks. Available at <http://www.fen.bilkent.edu.tr/~franz/ta/ta01.pdf>, last visited Sep 2018.
Lynn, B. (n.d.-a). Elliptic curves - the weierstrass form. <https://crypto.stanford.edu/pbc/notes/elliptic/weier.html>
Mead, D. G. (1973). The Equation of Ramanujan-Nagell and [y2]. *Proceedings of the American Mathematical Society*, 41(2), 333–341. <https://doi.org/10.2307/2039090>
P. Ingram, Multiples of Integral Points on Elliptic Curves, *Journal of Number Theory*, Volume 129, Issue 1, January 2009, Pages 182–208.
SageMath documentation: https://doc.sagemath.org/html/en/reference/arithmetic_curves/sage/schemes/elliptic_curves/ell_point.html
SageMath documentation: https://doc.sagemath.org/html/en/reference/arithmetic_curves/sage/schemes/elliptic_curves/ell_torsion.html
Serge Lang, Elliptic Curves, Springer, 2012, *Graduate Texts in Mathematics*, Volume 231.
Silverman, J. H., and Tate, J. T. (2015). Rational points on elliptic curves. Second ed. *Undergraduate Texts in Mathematics*. Springer, Cham. George Green Library, QA565 SIL.
Silverman, J. H. (1994). *Advanced topics in the arithmetic of elliptic curves* (Vol. 151). Springer.
Silverman, J. H. (1999). Computing rational points on rank 1 elliptic curves via L-series and canonical heights. *Mathematics of Computation*, 68(226), 835–858.
Silverman, J. H. (2009). *The arithmetic of elliptic curves* (2nd ed., Vol. 106). Springer.
Sutherland, A. (2021, February 17). 18.783 S2021 Lecture 1: Introduction to elliptic curves: Elliptic curves: Mathematics. MIT OpenCourseWare. https://ocw.mit.edu/courses/18-783-elliptic-curves-spring-2021/resources/mit18_783s21_slides/
Washington, L. C. (2008). Elliptic curves. Second ed. *Discrete Mathematics and its Applications* (Boca Raton, FL). Chapman Hall/CRC, Boca Raton, FL. George Green Library, QA565 WAS and QA567.2.E44 WAS.
Wiles, A. (1995). Modular Elliptic Curves and Fermat’s Last Theorem. *Annals of Mathematics*, 141(3), 443–551. <https://doi.org/10.2307/2118559>

Nullius in Verba: Artists, Corpses and Empiricism in Post-Enlightenment England and Beyond

Taylor Larson
Harvard College '25

In this paper, I investigate the intersection of artistic training, anatomical science, and ethical inquiry in post-Enlightenment England, focusing on the Royal Academy of Arts and its use of anatomical casts made from human remains. Through the work of figures like William Hunter and Joshua Reynolds, I explore how the Academy embraced empirical observation to reshape artistic pedagogy and challenge the boundaries between art and science. Central to my analysis is the Anatomical Crucifixion, a cast created from the corpse of an executed criminal, which exemplifies the period's shifting attitudes toward the human body—from sacred entity to object of study. I then extend this historical examination to contemporary practices, particularly the controversial exhibition *Body Worlds*. Drawing connections between past and present, I argue that the ethical tensions surrounding the display of human remains—issues of consent, commodification, and spectacle—persist today. By placing these practices in conversation, I reflect on how institutions produce and legitimize knowledge through the display of the dead, and what these displays reveal about our ongoing struggle to balance scientific inquiry with respect for human dignity.

Introduction

In May 2024, the Royal Academy of Art in London reinstated their cast collection after a short renovation period; while many had been placed in storage temporarily during the construction period, some had been out of the public eye for many years before. The majority of the collection, numbering about 140 in total, is an impressive assemblage of plaster casts made from marble statues, ranging from ancient Greek and Roman pieces to works by Michelangelo, all purchased with the intent of being used by students to learn both anatomy and classical beauty ideals. It was required that young students first become adept at creating drawings of these sculpture casts before they would be allowed into the Life Drawing room, where they would then draw, paint and sculpt based on live human models. Although students had to first become successful at drawing based on plaster sculpture casts before being allowed to reference live human models, the bodies of deceased models were not off the table for use in their work. Among the casts reinstated at the Academy, three differentiate themselves by the nature of their reference; rather than taking the shape of famous pieces of classical sculpture, they were created using the remains of executed criminals (**Figure 1**). Each was made in a collaboration between artists and scientists, working together to obtain a cadaver and utilize it in the process of crafting an anatomical cast, with the intention that it would become a tool of learning for students of the Royal Academy.

Crossovers between the world of fine arts and medicine were not new. In particular, collaborations between anatomists and artists for the sake of creating anatomical studies were commonplace in Europe during and after the Enlightenment era. Although many anatomists were themselves skilled artists, they often called upon professional artists/engravers to aid in the creation of published content in order to most adequately convey their discoveries to a wider audience (Petherbridge, 1977). This commonplace collaboration points toward the overarching fixation on detailed accuracy and realism in the realm of anatomical illustrations,



Figure 1. The three plaster casts created using human remains, seen hanging in the Royal Academy. Photo credit: Goppion Photo Archive.

which is fitting to be an object of focus in a field so dedicated to precision with the goal of empirical, objective truth. The establishment of The Royal Society in 1660 was inspired by the “new science”—the scientific method that has been shaping our world since the seventeenth century. The Society's first charter, written in 1662, established the motto “Nullius in verba,” or “Take

nobody's word for it.” This motto was indicative of the Society's dedication to peer-reviewing and experimentation for the purpose of constantly challenging established ideas, as well as the degree to which they valued hands-on experience. Initially composed of men from various backgrounds and professions, the formation of the Society was demonstrative of a more widespread interest in experimentation and exploration in fields that had previously been defined by knowledge passed down through generations of practitioners. This moment marked a departure from centuries of reliance on ancient texts and religious dogma, giving rise to a more evidence-based approach.

Starting in the mid-to-late eighteenth century, observers increasingly noted that, despite England's accomplishments in science and literature, it lacked comparable distinction in the visual arts—no English painters had achieved the renown of the nation's celebrated poets or natural philosophers (Kemp, 1992). William Hunter, who would become the Academy's first Professor of Anatomy, would say: “When we have already gone so far beyond the ancients in science, in every thing besides, are we never to excel them in works of imagination?” (cited in Kemp, 1975). This line is very much indicative of the idea behind the foundation of the Royal Academy of Arts in 1768; while England had long since emphasized furthering themselves in fields like science, the arts had been neglected. The Academy's first President, Joshua Reynolds, would allude to this in the dedication of his *Discourses*:

“By your illustrious predecessors were established marts for manufacturers, and colleges for science; but for the arts of elegance, those arts by which manufactures are embellished, and science is refined, to found an Academy was reserved for your Majesty” (Reynolds, 1797).

Both Reynolds and Hunter intended for the Academy to go forth educating students on the ‘scientific’ principles of art—it was by this method they sought to differentiate themselves from other academies across the continent. The term “scientific principles of art,” as used by Reynolds and Hunter, referred to the emphasis on empirical observation and systematic study of anatomy as foundational to creating accurate representations of the human body. This contrasted with earlier art training that prioritized idealized forms over observational accuracy.

Teaching Anatomy in Art

The detailed study of anatomy as a part of an artist's curriculum was likewise not a new phenomenon; *écorché*, the French word for “flayed,” is often used to describe the practice of visually recreating the figure of a human or animal with its skin removed to reveal the complex muscle and skeletal structure. Since the 15th century, western artists had been incorporating such studies into their portfolio with the intent of bettering their ability to realistically depict bodies in all sorts of poses, their future work guided by a complex understanding of how the body moved and operated. It became a standard component of classical artistic training, with the idea that one must fully understand every detail of how the body works in order to successfully render it realistically—a cornerstone of classical French art education that was adopted into the curricula of academies across Europe.

Many paintings and drawings have been created of anatomy lessons taking place in both medical and art academies, sometimes

in close quarters gathered around a table or within a large dissection theater. There are certain trademarks of these works that all appear very similarly across the different countries in which they were created: a congregation of smartly dressed men, writing utensils in one hand and paper in the other as the lecturer pointed to and explained details about what they were meant to draw. However, although on a surface level these lessons all appeared to be quite similar, there was some variety in the lecturers' intended outcome for the students in attendance. Art historian Andrew Graciano, in his paper “Anatomy in the Drawing Room,” wrote about how formal art academies in Europe had “long privileged historical subjects,” with beginner-level students often being required to base drawings off casts and classical sculptures before moving on to drawing based on a live model (Graciano, 2019). The imitation of the idealized anatomy in such sculptures would become internalized by art students over time, giving them a mental “catalog of perfection” with which to “‘correct’ the inevitably flawed living model” (Reynolds, 1797). The Academy's first President, Joshua Reynolds, would write about this “principal defect,” wherein students would “change the form according to their vague and uncertain ideas of beauty, and make a drawing rather of what they think the figure ought to be, than of what it appears” (Reynolds, 1797). As they were drawing, students would in real time alter the figure before them in order for it to conform to classical art standards of beauty and the ideal form as they had been taught by hours of close study.

This long-standing component of artistic education was one facet where President Joshua Reynolds and William Hunter, Professor of Anatomy, sought to differentiate themselves from other institutions of art education. They intended for the Royal Academy to operate grounded in what they saw as the “scientific principles of art.” Hunter's lessons were specifically crafted around the idea that students should start by drawing exactly what they see to develop a deep understanding of how the body operated and what it looked like doing so, with the belief that “to make solid proficiency in the study of any Art, it is observed that it is of infinite service to be grounded in its Elements, its scientific and demonstrable principles” (Hunter, cited Kemp, 1975). He would carry out dissections for the students of the academy to be audience to, as well as give detailed lectures with the assistance of a live model, cadaver cast, and skeletal remains all at once. In each lecture, Hunter would choose a single muscle group to discuss in great detail, utilizing his various models to give examples of each all the way from flesh to bone. “Hunter's lecture about one aspect of the whole body, albeit in depth, considering the interrelation of skin, muscles and bones, provided his audience with a focused accuracy informed by science” (Graciano, 2019). Hunter intended for his students to develop a deep understanding of the body and its functions so that they could recreate it faithfully rather than be influenced by the style consistencies of antiquity. Academy President Joshua Reynolds would later go on to clarify that merely imitating nature was not the highest form of art, and that it was up to artists to consider the abstract and perfect the imperfections of nature, thus making her more beautiful. He described the base understanding of nature as the mechanical components of art, which should serve as the foundation for an artist before moving on to the poetic components, using artistic expertise to create something more beautiful than how it existed originally. In an article comparing and contrasting the beliefs of

both Reynolds and Hunter, Oxford Art Historian Martin Kemp described the anatomist's approach to art:

“Hunter, by contrast, had come to espouse an uncompromising empiricism, and took his place on the extreme wing of characteristically British tradition. He was utterly committed to observational science, founded upon minute scrutiny, systemic description in words and images, and inductive analysis” (Kemp, 1992).

To aid his teachings, William Hunter would go on to create several écorché style casts with the help of flayed corpses that he was able to acquire due to his history as a surgeon and connections within the field of anatomy. Since the establishment of the Murder Act in 1752, surgeons had been the primary recipients of corpses from the gallows—the bodies of convicted murderers, condemned to the post-mortem punishment of dissection, which was most often performed for educational purposes. Its long title being “An Act for better preventing the horrid Crime of Murder,” the act was created in response to a widespread panic particularly centralized in London, which was the result of several homicides being heavily discussed and sensationalized by the press (Tarlow, 2018). Prior, those convicted of murder had already been fated to receive capital punishment; yet, it was determined by lawmakers that the extremity of this retribution was not severe enough to be a significant deterrent. Thus, they created the Murder Act with the intent “that some further terror and peculiar mark of infamy be added to the punishment of death” (Pickering, 1765). The law went beyond solely a death sentence for the convicted and added on the awareness that after death, one's body would not be subject to any sort of formal burial. Their postmortem location was either public dissection and anatomization or the gibbet, where the corpse would be hung in chains and left to decompose in front of the eyes of the living with whom they had once been a member of. Throughout the law's lifespan, 1166 individuals would be sentenced to death, and of them, 80% would be given to practitioners of medicine for dissection and anatomization (King, 2000). This act proved to be insufficient in meeting the demands for corpses available to carry out dissections upon as the fields of medicine and surgery rapidly advanced, with increased populations of students for whom teachers of anatomy would need to acquire corpses to educate them. To make up for this gap between supply and demand, individuals known as “resurrectionists” began unearthing newly buried corpses to sell them directly to medical academies. At times, the bodies would be stolen from the graves directly by teachers and students. The issue would culminate in 1828 when it was discovered that two men, William Burke and William Hare, had carried out at least 16 murders of victims whose bodies they would immediately sell to the University of Edinburgh. The Anatomy Act of 1832 would eventually give surgeons and medical students access to any and all unclaimed bodies (“An Act for regulating Schools of Anatomy,” 1832). The eighteenth to nineteenth century was a period notable for high levels of usage of human cadavers as tools.

On January 15th of 1770, the Royal Academy Council minutes record that “The President was desired to make an application to the Master of the Surgeons Company for a body to be [dissected] in the Royal Academy by Hunter,” and the March minutes of that year record intent to reimburse Hunter for the corpse (accessed

via Stephen, 2019). It is unclear which of his anatomical casts this particular cadaver would be utilized for—only two of the casts made using flayed criminal corpses still exist in the collections of the Academy today. One of these is an untitled écorché figure created in 1771, which is less identifiable with the living in that its face has been deconstructed down to skeleton and muscle (Figure 2). Although both of the other anatomical casts were also created with a flayed body, they have not been painted over and remain a single color, which serves to help disguise the degree to which the body has been deconstructed for the piece. In this standing figure, the bright red color of the musculature contrasted with the lighter hues used for the visible joints and sinews clinging to them make blatant to the viewer that what they are seeing is a bodily interior, an anatomical model which retains an obvious connection to the realm of science and is thus easy to view with a more clinical and impersonal perspective. The other figure created



Figure 2. Unidentified maker (production supervised by William Hunter), Écorché figure, cast probably 1771. Plaster cast. 1715 mm x 610 mm x 475 mm, Weight: 65 kg. © Photo: Royal Academy of Arts, London.

under the supervision of Hunter and made with the assistance of a real corpse is the *Smugglerius* cast (Figure 3), modeled after the ancient Roman statue *The Dying Gaul*. The subject has his face angled downwards, hiding his expression: because in the original *Dying Gaul* statue the man's face is lifted and slightly visible, it is possible that the creators of the *Smugglerius* made an intentional effort to hide the face as it would have been difficult to manipulate the cadaver's face so as to imitate the reference.



Figure 3. William Pink (production supervised by Agostino Carlini, R.A.), *Smugglerius*. 1834, (original cast ca. 1775). Plaster cast. 755 mm. Photograph © The Royal Academy of Arts, London.

The third cast possessed by the Academy is that of a flayed male corpse hanging from a cross. Named the *Anatomical Crucifixion*, it combines the visible face and singular coloration of the other two pieces, and it is these factors that make it the most recognizable as a once-living body (Figure 4). One of the most disturbing features of the cast is the expression, particularly the eyes. Upon first glance, it appears that the figure has had their head lolled to the side, eyes closed and mouth slightly open, appearing to suggest that either the crucified victim had already succumbed to their injuries and died, or was no longer capable of mustering the energy to keep their head raised and eyes open. However, the shape of the eyes present the most tell-tale evidence of the violence this body experienced before ending up in this cast state: they are exceedingly rounded, bulging out from the face as a result of the hanging of the cadaver model James Legg. The face and its evidence of the death experienced by its owner are the most visceral reminders of the fact that this sculpture cast was molded from the body of a real man, sentenced to death for the crime of murder.

The Anatomical Crucifixion

The name *Anatomical Crucifixion* is almost snide sounding, seeming to say that as opposed to all other crucifixion depictions, this one is the most anatomically correct rendition. Indeed, its creation was inspired by a desire among several artists to prove their theory that, although a scene recreated by many great artists, classical depictions of the crucifixion of Christ did not appear “natural” (“Obituary,” 1846). Three members of the Royal Academy—sculptor Thomas Banks, portrait painter Richard Cosway, and the current Academy President Benjamin West—approached prominent surgeon Joseph Constantine Carpue to request his assistance in procuring a body with which they could imitate the crucifixion position and create a model with which to prove whether they had been correct in their

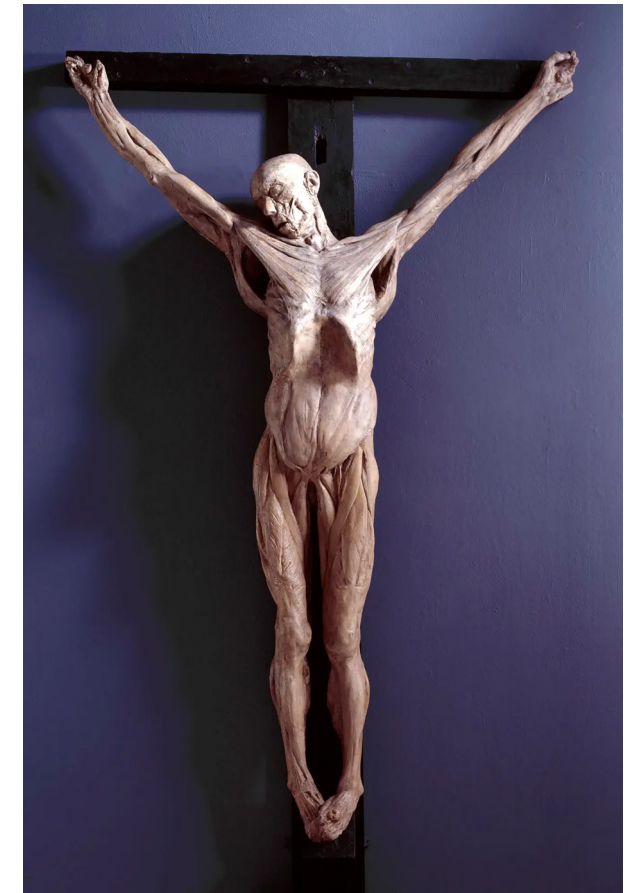


Figure 4. Thomas Banks, R.A., *Anatomical Crucifixion* (James Legg), 1801. Plaster cast, early 19th century. 2315 mm x 1410 mm x 340 mm, Weight: 74 kg. Photograph © The Royal Academy of Arts, London.

theory. William Hunter had passed away in 1783, predating the creation of this cast made in 1801. Thus, the Academy artists had seen Carpue as their point of contact between the art world and science, and therefore their key to accessing a cadaver for use in their anatomical study. All three artists involved in the creation of this gruesome piece would have been students of the academy before the death of Hunter, and so would have taken lessons of anatomy under him, which may have been the inspiration for this particular project. What we can understand as certain inspiration for this undertaking was an alleged story about Michelangelo, wherein he was said to have tied a model to a crucifix before stabbing him to death in order to obtain an accurate concept of the impact of crucifixion on the human body. As written by European Art Historian Meredith Gamer, this story appears to have been well-known in the early nineteenth century as it was recounted by the Academy artists and appeared in a *Morning Chronicle* publication (Gamer, 2014). In his final addition to his *Discourses*, Joshua Reynolds, who would have been President within the three artists' time at the Academy, placed great emphasis on the work of Michelangelo, whom he referred to as the “founder and father of modern art” (Reynolds, 1797). He discussed how Michaelangelo wielded a profound understanding of not only the mechanical, scientific nature of his subjects, but that he also possessed the ability to poeticize the subject matter, aestheticizing it to the point of achieving greatness.

“I will not say that Michaelangelo was eminently poetical, only because he was greatly mechanical; but I am sure that mechanic excellence invigorated and emboldened his mind to carry paintings into the regions of poetry...” (Reynolds, 1797).

At the end of this speech, Reynolds finished off with saying the name Michaelangelo one last time, having decided that he desired the name to be his final words spoken in the Academy chair. It is worth considering whether this speech had such a profound effect on the Academists so as to inspire them to want to reattempt his supposed anatomical experiment.

After the artists approached Joseph Carpue, he got to work sourcing a corpse for their use in the creation of the casts. Because only the bodies of executed murderers were available for use, surgeons would keep track of legal proceedings and maintain contact with colleagues in order to be aware of any upcoming executions, soon-available bodies that they could lay claim to. Shortly after submitting the application for a corpse subject, an incident took place at the Chelsea Hospital in October 1801, which led to a cadaver becoming available: that of an elderly man named James Legg. The Chelsea Hospital ran on a military-like system wherein ex-military inhabitants would become Captains of their own company composed of other patients; Legg was one of these Captains and had recently fought with a fellow pensioner before shooting him in the chest (“Dying Behaviour, &c. of James Legg,” 1801). Before a charter addendum to the Murder Act created in 1822, it was required that the place to which an executed corpse was delivered be within 400 yards of the execution site. Because of this, a building was erected near the location at which Legg was set to be executed, and stocked with the wood for a crucifix as well as other materials necessary for the cast creation. In order for it to appear the most “accurate,” the process of transporting and posing the cadaver was done quickly so that rigor mortis would not have time to completely set in before they had the chance to place the corpse on a cross and allow it to fall into place. The Academicians and Carpue attended Legg’s execution, where he was to meet his end alongside a man who had murdered his own wife, Richard Stark. The “event” was attended by hundreds of onlookers, and, “after hanging the usual time [the men] were cut down and delivered to the surgeons for dissections” (“Dying Behaviour, &c. of James Legg,” 1801). Legg was taken to the designated space in which the casts were to be created. One plaster cast was created with his entire body intact, and another—the one that is still visible in the Royal Academy today—was completed after the corpse had been flayed. This is similar to the other two casts that had been created when Hunter was still working at the Academy: all three of the plaster casts still exist only in their final state of exposed muscle.

Implications

What does it mean to disturb the ephemerality of a corpse? Historically, the human corpse held sacred value, often afforded burial rituals to mark its social and religious significance. The transition toward treating corpses as empirical tools reflected the broader Enlightenment shift toward scientific rationalism. Yet, this utilitarian approach to human remains raises profound ethical dilemmas: at what point does the pursuit of knowledge override the need for dignity?

The shift in perception was not merely due to scientific advancement but was also influenced by cultural and legal changes, including increasing secularization and changes in penal practices. The use of executed criminals’ bodies, seen as a posthumous extension of punishment, marks the convergence of public justice, education, and art. During the period in which James Legg had been executed, the act of denying a corpse burial was seen as an extension of capital punishment. Part of that punishment was the knowledge of one’s corpse being exposed to the eyes of many, whether in a dissection theater or hung in chains; there would in both instances, however, be a point at which the corpse decayed beyond being recognizable, no longer visually identifiable with the living person it had once been. If we are to regard the public display of one’s corpse as a punishment, one could surmise that to capture in visual media the appearance of one’s corpse would seem to extend that punishment indefinitely. Indeed, after its creation, the cast would be displayed in varying locations and viewed by many who flocked to it with curiosity and eagerness to examine the figure. However, among those viewing it were individuals who were learning from it intending to improve their art, which served the purpose designated by its creators. The creators of this piece would not likely have considered it as something meant to prolong the punishment of its subject, but rather as a valuable tool of education. It is not simply a macabre reminder of the price to be paid for committing a capital crime, but instead becomes an image of restitution “as it takes the body of a murderer—one who committed the ultimate crime against society—and puts it into the service of that society” (Gamer, 2014). After both James Legg’s flayed corpse and the cast had been displayed for a short period of time, Thomas Banks would write to the Royal Academy that he desired they take the *Anatomical Crucifixion* into their collection so that it could be utilized as a useful tool for the Academy students and professor of Anatomy (Banks, accessed via Gamer, 2014). The Academy would accept his proposal and by 1802 the cast was on display.

The *Anatomical Crucifixion* provokes intrigue, particularly in the fact that it was not created by or under the request of an anatomist, unlike the prior casts overseen by William Hunter. The piece was created not only to better grasp human anatomy, but with the intent of specifically interrogating the appearance of a crucified body: an image that has been recreated countless times by artists of all skill levels. This means that there existed an exhaustive catalogue of works to reference if one sought to create their own crucifixion scene; the creation of an anatomical tool for the purpose of more accurately depicting a crucified body is representative of a widespread belief at the time when rapid scientific advancement had brought into question the act of relying upon pre-existing knowledge and emphasized personal investigation and experience. At the time of its creation, two of the artists involved were in the process of creating pieces involving the crucifixion scene. Benjamin West was in the midst of designing the monumental west window of Saint George’s chapel at Windsor Castle, while Richard Cosway was working on a series of detailed drawings depicting scenes from throughout the life of Christ (Gamer, 2014). Their choice to create an anatomical model of the *Crucifixion* brought both religion and art into conversation with principles of the scientific revolution, as it regarded direct observation and evidence over received knowledge. Their collective treatment of the human

corpse as a tool of empirical study was indicative of the shift from corpses being viewed as particularly sacred objects with intrinsic value, which had been happening over the course of the last century. These cadaver casts would go on to become staples in the sketchbooks of the many artists who passed through the Royal Academy, immortalizing them in countless images.

We live in a time when museum collections are facing growing scrutiny over what is appropriate to display, even as visual documentation and the sharing of information have expanded dramatically. Alongside the questions of how and why the remains were acquired, the major ethical dilemma regarding the display of human remains involves complex discussions of the corpse as property, discussions that become even more complicated as topics like consent and religion are introduced. One main justification for displaying remains—that it serves a greater purpose for the betterment of humanity through education—has reappeared in contemporary public discussions and social critique whenever the question of using human corpses as tools arises. The British Museum website has a page dedicated to its collection of human remains, and on the subject of display states: “Surveys show that most visitors are comfortable with and expect to see human remains as an element of our Museum displays” (Fletcher, 2014). The controversial travelling exhibition *Body Worlds*, which was investigated in 2004 after claims arose that their plastinated body displays were created using the remains of executed Chinese prisoners, has been repeatedly criticized in the media, and remains a subject of ethical debate. Like the British Museum, the *Body Worlds* website has a page detailing the results of a survey conducted at the end of their exhibitions across several cities, stating that 87% of the visitors stated that they knew more about the human body after their tour (Fonseca, 2013).

The ethical tensions present in historical anatomical casts are paralleled by modern controversies, particularly surrounding the aforementioned *Body Worlds* exhibition. Created by Gunther von Hagens, *Body Worlds* features plastinated human corpses posed in dynamic, sometimes theatrical positions and is framed as both educational and artistic. Despite claims of public health benefit and anatomical education, the exhibition has drawn criticism for sensationalism and commodification of the human body (Goulding et al., 2013). Visitors often experience conflicting reactions—fascination, discomfort, even reverence—when confronted with these carefully staged corpses. The exhibition’s success, reportedly drawing over 29 million attendees globally, underscores the ongoing allure of death when mediated through the lens of science and spectacle. Critics argue that *Body Worlds* transforms human remains into aesthetic commodities, stripping bodies of personal identity while highlighting the body as both machine and object. As Christina Goulding and colleagues have observed, this dehumanizing treatment allows viewers to dissociate the exhibits from their once-living origins, mirroring historical instances in which anatomical casts served educational aims while erasing individual identities (Goulding et al., 2013).

Goulding et al. outline five interpretive frameworks through which audiences read the *Body Worlds* displays: as spectacle, as mortality salience, as commodity, as machine, and as dehumanized body. These readings, deeply dependent on audience subjectivity, underscore the exhibition’s paradox. The plastinated corpses are both objectified and aestheticized; they are

emotionally distanced yet viscerally intimate. Exhibits such as a pregnant woman with her fetus exposed evoke both educational intrigue and ethical discomfort, challenging viewers to reconcile the bodies’ prior humanity with their current objecthood. The intentional staging—chess-playing skeletons, yogic poses, flayed ballerinas—intensifies the spectacle, and thus risks trivializing the deceased even as it claims to promote anatomical education.

Moreover, the commercial success of the exhibition complicates its claims to public good. While von Hagens positions himself as a democratizing force reclaiming dissection from the medical elite, others see *Body Worlds* as exploitative—a blend of macabre entertainment and for-profit science. The boundary between museum and sideshow becomes blurred, reflecting not only modern society’s complex relationship with death but also a consumerist approach to mortality, where corpses become both spectacle and brand.

This modern manifestation of anatomical display reinforces the argument that the ethical dilemmas explored in historical contexts remain deeply relevant today. Whether in the halls of the Royal Academy or in global touring exhibitions, the presentation of human remains continues to straddle the uneasy line between scientific inquiry, artistic expression, and public consumption. After being retrieved from storage only last year, the Royal Academy casts have been placed back on display in the Academy halls. In a video made announcing the reinstallment, Head of Fine Art Processes Mark Hampson acknowledges that “in the contemporary art school, the casts don’t have an obvious teaching purpose or function.” He elaborates that they are no longer needed as the “kind of Google Image search they were for the Georgian students.” Despite this, as with the previously mentioned displays around the world which feature depictions of or actual human remains, they continue to be placed before audiences for the purpose of education or inspiration. The ethical dilemmas surrounding the display of plaster casts made from human bodies and the use of plastinated corpses in exhibitions like *Body Worlds* reveal enduring tensions between scientific inquiry, artistic representation, and public spectacle. Both practices raise fundamental questions about consent, the treatment of human remains, and the role of institutions in curating images of death for educational or aesthetic purposes.

This complication is an ongoing ethical paradox: the human corpse, once removed from its social identity, is often treated as an anonymous specimen rather than a former person. The tension between scientific utility and ethical responsibility persists, with modern debates mirroring historical anxieties over bodily autonomy and institutional authority. Whether displayed in an 18th-century art academy or a contemporary traveling exhibition, the use of preserved human bodies forces us to confront fundamental questions about dignity, consent, and the limits of educational display. While our understanding of ethics has evolved, the underlying dilemma remains unchanged: at what point does the pursuit of knowledge come at the cost of human respect? Visual media is created and shared now more than ever before, and in a society where most people have access to image search engines, it’s increasingly difficult to justify the use of real corpses to create empirical tools for the purpose of allowing individuals to see something with their own eyes. Rapid advancement necessitates constant interrogations of ethical and moral standards.

References

Banks, T. (1801). *Anatomical Crucifixion (James Legg)*. Plaster cast. Royal Academy of Arts, London.

Banks, T. (1802). *Letter to the President and Council of the Royal Academy*, July 22.

Bryant, J. (2005). The Royal Academy’s ‘Violent Democrat’: Thomas Banks. *The British Art Journal*, 6(3), 51–58.

Gamer, M. (2014). Criminal and Martyr: The Case of James Legg’s Anatomical Crucifixion. In S. M. Promey (Ed.), *Sensational Religion: Sensory Cultures in Material Practice*. Yale University Press.

Gamer, M. (2019). The Smugglerius, Re-Viewed. *The Sculpture Journal*, 28(3), 331–344. <https://doi.org/10.3828/sj.2019.28.3.5>

Goulding, Christina, et al. “Reading the Body at von Hagen’s ‘body Worlds.” *Annals of Tourism Research*, vol. 40, 2013, pp. 306–30, <https://doi.org/10.1016/j.annals.2012.08.008>.

Graciano, A. (2019). Anatomy in the Drawing Room at Felix Meritis Maatschappij in Amsterdam: Between Skin and Bones, Theory and Practice. In *Visualizing the Body in Art, Anatomy, and Medicine Since 1800*. Routledge.

Hunter, W. (1975). *Dr. William Hunter at the Royal Academy of Arts*. M. Kemp (Ed.). University of Glasgow Press.

Iaccarino, M. (2001). Science and Ethics. *EMBO Reports*, 2(9), 747–750. <https://doi.org/10.1093/embo-reports/kve191>

Kemp, M. (1992). True to Their Natures: Sir Joshua Reynolds and Dr William Hunter at the Royal Academy of Arts. *Notes and Records of the Royal Society of London*, 46(1), 77–88. <https://doi.org/10.1098/rsnr.1992.0004>

King, P. (2000). *Crime, Justice, and Discretion in England, 1740-1820*. Oxford University Press.

McCormack, H. (2019). Joseph Banks and William Hunter: Where the Royal Society Meets the Royal Academy. *Journal for Maritime Research*, 21(1–2), 119–142. <https://doi.org/10.1080/21533369.2020.1763634>

Obituary: Joseph Constantine Carpue, F.R.S. (1846). *The Lancet*, 47(1171), 166–168.

Petherbridge, D., & Jordanova, L. J. (1997). *The Quick and the Dead: Artists and Anatomy*. University of California Press.

Pickering, D. (1765). *The Statutes at Large: from the 23d to 26th Year of King George II*. Joseph Bentham, Printer to the University, for Charles Bathurst.

Pink, W. (c. 1834). *Smugglerius*, after Agostino Carlini, RA. Plaster cast of 1776 original. Royal Academy of Arts, London.

Reynolds, J. (1767–1792). *Sir Joshua Reynolds Papers*. Series: II. Manuscripts, 1774 and Undated; Series: I. Letters, 1767–1791 (Box 1, no. 009564559). Aeon.

Reynolds, J. (1797). *The Works of Sir Joshua Reynolds, Knt. Late President of the Royal Academy*. T. Cadell Jun. and W. Davies.

Reynolds, J. (1768–1792). *The Works of Sir Joshua Reynolds, Knt. Late President of the Royal Academy* (no. 990078007620203941). Aeon.

Stephens, R. (2019). The Minute Books of the Royal Academy Under Sir Joshua Reynolds, 1768–92. *The Volume of the Walpole Society*, 81, 1–454. JSTOR, <https://www.jstor.org/stable/26906782>

Stimson, D. (1948). *Scientists and Amateurs, a History of the Royal Society*. H. Schuman.

Stephan, C. N., & Fisk, W. (2021). The Dubious Practice of Sensationalizing Anatomical Dissection (and Death) in the Humanities Literature. *Journal of Bioethical Inquiry*, 18(2), 221–228. <https://doi.org/10.1007/s11673-021-10095-2>

Tarlow, S., & Battell Lowman, E. (2018). *Harnessing the Power of the Criminal Corpse*. Palgrave Macmillan.

Unidentified maker (1801). *Dying Behaviour, &c. of James Legg, for the Murder of William Lamb; and Richard Stark, for the Murder of His Wife*. Broadside. Woodcut. Bodleian Library, University of Oxford.

Unidentified maker (attributed to W. Hunter) (1771). *Écorché figure*. Plaster cast. Royal Academy of Arts, London.

Features



Are We Entering the Dark Ages of Science?

Leah Lourenco '26

Introduction

For nearly 140 years, the National Institutes of Health (NIH) has supported the development of critical medical advances in the United States (NIH, 2014). In a statement made on February 7th, 2025, following the inauguration of US President Donald Trump, the NIH reiterated its mission of “[seeking] fundamental knowledge about the nature and behavior of living systems’ in order to enhance health, lengthen life, and reduce illness and disability.” During the same announcement, the NIH informed the public that it would more than halve the amount of grant funding allocated to indirect costs, which are used to fund facilities and administration costs (NIH, 2025). This major change comes as the Trump administration seeks to reduce government spending more broadly. The Trump administration has claimed that this cut of indirect costs by the NIH will save over \$4 billion annually; however, the

research community holds concerns about the future of biomedical research under this new funding framework (Seminera, 2025).

What is the NIH?

The NIH has been central to the United States’ role as a leader in biomedical innovation and advancement since its founding in 1887, in a rudimentary laboratory within the Marine Hospital Service. It was only in 1930 that the organization was formally renamed to the National Institute of Health; over the subsequent decades, 27 institutes, including the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, and the National Institute of Mental Health, were formed to create the NIH that we know today (NIH, 2015a; NIH, n.d.).

Grants from the NIH are a significant source of funding for biomedical research across the country. In fiscal year (FY) 2024, the NIH budget was more than \$47 billion,

83% of which supports extramural research conducted in research bodies outside of the NIH, while 11% goes to intramural research (Sekar, 2024; NHGRI, 2015). Grants can be awarded to a wide variety of organizations, whether they be internal or external to the United States, for-profit or non-profit, or public or private. This can include other federal institutions, universities, non-profits, hospitals, and even individuals (NIH, 2024a).

Countless lives have been altered by technology developed by NIH-funded projects (NIH, 2015b). When the Vanderbilt Comprehensive Care Clinic opened in 1994, it was little more than a palliative care facility for patients suffering from human immunodeficiency virus (HIV) with little hope for survival. As the years have passed, medical innovation has made it possible to live life with an HIV diagnosis, without significant risk of the disease progressing into AIDS. Steve Raffanti, co-founder of the clinic, remarks on the dramatic change in the patient base of the clinic, “Our mortality rate in ’96 dropped 93% and has stayed down ever since. You would walk through the hallways of the clinic where all the exam rooms were and some would run out and hug you and say, ‘Oh, Dr. Raffanti, you can’t believe how well I feel!’” (NIH, 2016a). Funding from the NIH was essential in bringing about this dramatic reversal of HIV mortality. In FY 1995, the NIH provided \$1.3 billion in funding to HIV/AIDS research, while by FY 2000, funding for the same research had reached \$2.1 billion (Kates & Summers, 2004).

Although many NIH-funded projects address the deadliest, most prevalent conditions that people face—including HIV/AIDS, cancer, and cardiovascular disease—the NIH also provides support to patients with rare diseases. For example, twins Alexis and Noah Beery faced a rare neurological disease, called dopa-responsive dystonia, early in life. It was only through NIH-supported genetic sequencing that the cause of the disease was accurately discovered. This discovery provided clinicians with the ability to properly treat the Beery twins, ultimately giving them a healthy life (Knox, 2011; NIH, 2016b).

NIH Funding in Early 2025

The February 7th statement by the NIH shook the research community by announcing that a 15% indirect cost rate would be applied to any active or new grants after February 10, 2025. This stands in stark contrast to the previous indirect cost rates, which have averaged approximately 28% over the past decade, and in some cases, even more.

When considering indirect costs, there are two distinct components to consider: facilities and administration (NIH, 2025). Facility costs account for depreciation and interest on the debt of research facilities, equipment, and capital improvements, as well as utilities and maintenance (NIH, 2024b; Clark & Klumpp, 2025; NIH, 2025). Administrative costs are a catch-all for everything not facilities-related, including the salaries of grant administrators and staff who have to manage the administrative, financial, regulatory, and safety activities necessary to meet federal regulations, as well as accounting and legal fees.

In response to the statement by the NIH, Dr. John Shaw, Vice Provost for Research at Harvard University, filed a court declaration on February 10th, which argued that “a sudden and unexpected reduction in the indirect cost rate would be disastrous [to Harvard’s research endeavors].”

“A sudden and unexpected reduction in the indirect cost rate would be disastrous [to Harvard’s research endeavors]”

Shaw cited the many impacts these cuts would have on Harvard and research at large. He projected that Harvard would need to make immediate staff cuts, ultimately slowing the progression of current research projects. Halted projects could become obsolete or require further repetition, creating more work and necessitating more funding. Shaw noted the potential impacts on the Boston area: Harvard University’s research has a significant impact on the local economy, employing over 18,700 citizens and creating new ventures in the private sector. Finally, Shaw recognized the gap in technological advancement and economic growth that could occur between the United States and competitor nations if research lags (Shaw, 2025). Shaw’s concerns about the NIH announcement echoed throughout the Harvard research community. In one article by *The Harvard Crimson*, nine researchers across several schools within the University expressed concerns that this policy change would be the end of some important research groups, such as the Brugge Laboratory in the Department of Cell Biology at Harvard Medical School and the Fortune Laboratory studying tuberculosis at the Harvard T.H. Chan School of Public Health (Patel & Yoon, 2025).

For the time being, Judge Angel Kelley of the United States District Court for the District of Massachusetts has blocked the policy of decreased indirect cost support from taking effect nationwide. This ruling was made in response to two lawsuits claiming that the new NIH policy violated federal law (Stein, 2025). If the temporary policy block is lifted, there will be a significant impact on the wider research community. Dr. Donald Ingber, Founding Director of the Wyss Institute at Harvard University, released an article on February 13, 2025, entitled “Bringing the American Economic Flywheel to a Screeching Halt,” where he addresses how these research funding cuts could impact the United States’ reign as a major economic leader.

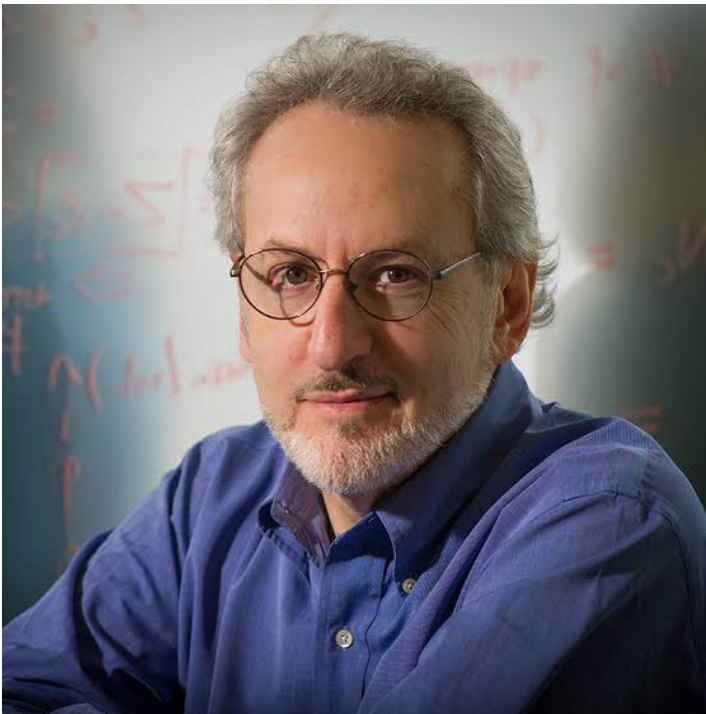


Figure 1. Donald Ingber (Courtesy of the Global Virus Network)

In early March, THURJ interviewed Ingber, who expanded on his motivations for writing the article as well as the implications of the NIH policy change for the United States and Harvard as research bodies. Once the announcement had been made, many researchers commented on the detriment such a change would bring to medical advancement; however, Ingber set out to discuss a different perspective. In his statement, Ingber thought it was important to emphasize the disastrous economic consequences of such a funding change, given that economic growth and success are a major focus of the current federal administration. “I think all they care about is economic competitiveness, international

competitiveness—the economy, and I thought we needed to translate the impact of this decision on that sort of outcome,” Ingber told THURJ. Technologies derived from NIH-funded research have been a key driver of economic growth due to their ability to circulate funds and generate jobs. If NIH-funded research declines, biomedical advancements will likely slow, attracting fewer of the world’s top researchers to the American system.

Moving Forward

In its February 7th statement, the NIH argued that the funding cuts were being implemented “to ensure that as many funds as possible go towards direct scientific research costs rather than administrative overhead,” implying that the \$4 billion the administration expects to save will be funneled into direct scientific research costs (NIH, 2025; Marquez & Bush, 2025). In response, Ingber said, “I think the chance that they're going to shift the money from indirect costs into direct costs—it's slim at best.” Ingber argues that it would likely take years to make such an adjustment, and it remains true that the haste with which the funding changes were administered is a real threat to ongoing research projects.

Unexpectedly, the Trump administration has since decided to freeze \$2.2 billion in federal funding to Harvard, following Harvard President Alan Garber’s refusal to adhere to the Trump administration’s excessive demands regarding academic freedom and discipline on campus. This announcement came on April 14, 2025 and has had an immediate impact on current research activities across the University, including those of Ingber, who received an order to stop two projects related to his organ-on-a-chip technology (Rai & Sundar, 2025).

Prior to these most recent cuts, Ingber acknowledged, “I think if you ask what it feels like for faculty and students—we're confused.” Until further decisions can be reached, it is critical to remain informed about the ongoing dialogue surrounding funding and the role that research plays in broader society. As Ingber writes, “Change can only come about when many voices are heard.”

“Ingber acknowledged, ‘I think if you ask what it feels like for faculty and students—we're confused’”

Ultimately, scientific progress will inevitably be made. Whether the private sector assumes a larger role in innovation or the government reconsiders its current path, researchers will continue to push to make the world better, day by day. Ingber assures us: “I think science is something that will always be done; even in the dark ages, there were people doing interesting stuff behind the scenes, you know?”

References

2.3.2 Eligibility. (2024a, April). National Institutes of Health (NIH). Retrieved March 30, 2025, from https://grants.nih.gov/grants/policy/nihgps/html5/section_2/2.3.2_eligibility.htm.

7.9 Allowability of Costs/Activities. (2024b, April). National Institutes of Health (NIH). Retrieved March 30, 2025, from https://grants.nih.gov/grants/policy/nihgps/html5/section_7/7.9_allowability_of_costs_activities.htm.

Baker, S. Nature Index 2024 Research Leaders: Standout performers make their mark in health sciences | News | Nature Index. (2024, June 18). Retrieved March 30, 2025, from <https://www.nature.com/nature-index/news/nature-index-research-leaders-standout-performers-health-sciences>.

Bringing the American Economic Flywheel to a Screeching Halt. (2025, February 13). Wyss Institute. <https://wyss.harvard.edu/news/a-letter-from-our-founding-director-bringing-the-american-economic-flywheel-to-a-screeching-halt/>.

Chronology of Events. (2015a, February 11). National Institutes of Health (NIH). <https://www.nih.gov/about-nih/what-we-do/nih-almanac/chronology-events>.

Clark, D. & Klumpp, L. Understanding the NIH’s New Indirect Cost Rate Policy: What Nonprofit and Higher Education CFOs Need to Know | BDO. (2025, February 25). Retrieved March 30, 2025, from <https://www.bdo.com/insights/industries/nonprofit-education/understanding-the-nih-s-new-indirect-cost-rate-policy-what-nonprofit-and-higher-education-cfos-need>.

Frequently Asked Questions About NHGRI Research. (2015, October 23). Retrieved March 30, 2025, from <https://www.genome.gov/12011002/about-nhgri-research-fact-sheet>.

History. (2014, October 31). National Institutes of Health (NIH). <https://www.nih.gov/about-nih/who-we-are/history>.

“I never knew how it was to feel free.” (2016b, July 8). National Institutes of Health (NIH). <https://www.nih.gov/about-nih/who-we-are/i-never-knew-how-it-was-feel-free>

“In the first twelve months, we lost over 300 people.” (2016a, July 8). National Institutes of Health (NIH). <https://www.nih.gov/about-nih/who-we-are/first-twelve-months-we-lost-over-300-people>.

Institutes at NIH. (n.d.). National Institutes of Health (NIH). Retrieved March 30, 2025, from <https://www.nih.gov/institutes-nih>.

Kates, J., & Summers, T. (2004, March). Trends in U.S. Government Funding for HIV/AIDS: Fiscal Years 1981 to 2004. Progressive Health Partners & Kaiser Family Foundation.

Knox, R. (2011, June 16). Genome Maps Solve Medical Mystery For Calif. Twins. NPR. <https://www.npr.org/sections/health-shots/2011/06/18/137204964/genome-maps-solve-medical-mystery-for-calif-twins>

Marquez, A. & Bush, E. NIH announces it’s slashing funding for indirect research costs. (2025, February 8). NBC News. <https://www.nbcnews.com/politics/politics-news/nih-announces-slashing-funding-indirect-research-costs-rcna191337>.

NOT-OD-25-068: Supplemental Guidance to the 2024 NIH Grants Policy Statement: Indirect

Cost Rates. (2025, February 7). National Institutes of Health (NIH). Retrieved March 30, 2025, from <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-25-068.html>.

Patel, D. & Yoon, G. Harvard Researchers Brace for Impact As NIH Threatens To Limit Support For Indirect Costs | News | The Harvard Crimson. (2025, February 14). <https://www.thecrimson.com/article/2025/2/14/researchers-brace-impacts-funding/>.

Rai, A. & Sundar, S. Stop-Work Orders Roll In for Harvard Researchers After \$2.2 Billion Pause in Federal Funds | News | The Harvard Crimson. (2025, April 16). <https://www.thecrimson.com/article/2025/4/16/stop-work-orders/>.

Seminera, M. Trump cuts to NIH research funding halt Duke projects | AP News. (2025, March 8). Retrieved March 30, 2025, from <https://apnews.com/article/trump-cuts-research-funding-nih-duke-7f24b33bbad54490583520536ab40e0c>.

Sekar, Kavya. National Institutes of Health (NIH) Funding: FY1996-FY2025. (2024, June 25). [Legislation]. Retrieved March 30, 2025, from <https://www.congress.gov/crs-product/R43341>.

Shaw, J. H. (2025, February 10). DECLARATION OF JOHN H. SHAW.

Stein, R. NIH funding policy temporarily blocked by judge: Shots—Health News: NPR. (2025, February 11). Retrieved March 30, 2025, from <https://www.npr.org/sections/shots-health-news/2025/02/08/g-s1-47383/nih-announces-new-funding-policy-that-rattles-medical-researchers>.

Voices of the NIH Community. (2015b, March 27). National Institutes of Health (NIH). <https://www.nih.gov/about-nih/who-we-are/voices-nih-community>.

FEATURES

The Neuroscience of Time Perception and the Ethics of Altering It

Kolbjørn Skarpnes & Rita Elmkvist Nilsen / NTNU & Kavli Institute for Systems Neuroscience.

David Kim '27

Introduction

Throughout history, human societies have been governed by the passage of time. From coordinating agricultural cycles in ancient civilizations to the split-second transmission of information over the Internet, time has served as the foundation of our routines and social order. Time, however, is not as objective as it seems on the face of a clock: for many people, five minutes on a thrilling roller coaster passes by much more quickly than five minutes waiting in an unmoving line.

Strangely, our sense of time is elastic: unlike for primary senses, including vision or hearing, the human body houses no single organ responsible for sensing time and its passage. Instead, our brains construct our sense of time through complex neural circuits, which makes understanding variations in

time perception highly enigmatic. Researchers have begun uncovering how various brain regions interface to track time, as well as the specific neurotransmitters and conditions that might alter this unique sense. Yet, despite these advances, ethical questions arise: how might manipulating someone's sense of time affect society, from medicine to criminal justice? Thus, even as time perception is increasingly understood from a neuroscientific lens, its bioethical implications must be thoroughly explored.

Neural Clocks and Human Time Perception

In recent years, our understanding of how the human mind tracks time has become more nuanced and complex; studies suggest that multiple brain regions—including the prefrontal cortex, cingulate cortex, parietal lobe, supplementary motor area, and basal ganglia—shape our perception of time. However, no single “control center” has been identified, and the

specific contributions of each region remain widely unclear (Fontes et al., 2016).

Regardless, some initial clues about the roles of the regions have emerged. The supplementary motor area is active when we count seconds or tap rhythms, suggesting that it helps us produce timed actions. On the other hand, the prefrontal cortex aids in attention and memory of time, which might manifest when one estimates how long an event lasts and compares it to their expectations. Additionally, the basal ganglia—particularly its connections with dopaminergic neurons in the midbrain—form a core timing circuit. Within this circuit, evidence from neuroimaging indicates that the dorsal striatum and the nigrostriatal dopamine pathway may function as a sort of neural pacemaker. For instance, patients with Parkinson's disease (PD) lose the dopaminergic neurons that compose a critical part of this pathway, and have been shown to have trouble keeping up with time perception and time-related tasks (Fung et al., 2021).

To this end, PD research increasingly demonstrates that dopamine may play a significant role in modulating the speed of our internal clock. High dopamine levels may make our neural clock faster (and thus perceive time as shorter), whereas lower levels have the opposite effect, making time seem longer (Apaydin et al., 2018). This dopaminergic-dependent mechanism may help explain the popular saying, “time flies when we're having fun,” as pleasurable experiences stimulate dopamine release and thus shorter perceived time.

“This dopaminergic-dependent mechanism may help explain the popular saying, ‘time flies when we're having fun,’ as pleasurable experiences stimulate dopamine release and thus shorter perceived time”

Other neurotransmitters have also been popularly linked to processing time: for instance, hallucinogens like lysergic acid diethylamide (LSD) famously slow down how humans perceive time by affecting serotonin-involved pathways (Shebloski & Broadway, 2016). Similarly, GABA—the brain's primary

inhibitory neurotransmitter—has been implicated to predict time perception (Terhune et al., 2014). These findings indicate that identifying drugs modulating neurotransmitter activity could functionally speed up or slow down our internal time clocks, controlling the intricate and elusive system that investigators are continually closer to mapping.

Current Frontiers and Future Directions

Although our understanding of time perception has greatly progressed in the past few decades, many questions about its scientific implications linger. For short timescales—that is, at the level of milliseconds—our brains often use *dedicated* sensory circuits, such as auditory circuits, for perception. However, for longer durations leading up to minutes, our brains instead might rely on more general neuronal populations in the cortex and basal ganglia (Paton & Buonomano, 2019).

Recently, the concept of a “neural population clock” has gained traction, which holds that a network of neurons is responsible for time as opposed to a single neuron (Penhune & Zatorre, 2019). Researchers have found that a neuronal group in the premotor cortex displayed activity that reliably followed the passage of time, suggesting that this network may be able to keep track of time changes (Merchant et al., 2014). In parallel, neuroscientists have also been working to integrate their studies of time perception with other fields, such as reward learning in behavioral studies. For example, investigators recently showed that reward expectation circuits within the brain and reward-dependent learning might have bidirectional effects. Furthermore, studies investigating time perception in clinical disorders (like depression, anxiety, and schizophrenia) have suggested that temporal aspects of mood and emotion from these disorders are key characteristics in the clinical experience (Amadeo et al., 2022).

Despite its increasing feasibility due to biotechnological advancements, the prospect of altering individual time perception raises profound philosophical considerations. The brain's construction of time contributes significantly to our ideas of identity and continuity, which are essential in forming core memories and other crucial aspects of personhood. Scientific advances that help us correct timing “problems”, such as those caused by PD or depression, hold promise for improving quality of life. Further, because consciousness is often understood as a

continuous stream of experiences, these tools may offer deeper insights into how the brain weaves these moments and perceptions together into a metaphysical stream.

Perhaps the most intriguing of these philosophical considerations is the concept of intentionally manipulating or engineering our perception of time. If we can somehow grasp and comprehend the neural code for time, could we use this knowledge with good intent and purpose? To date, no such technology exists (such as reliable and invasive brain stimulation), but current research has been building on the possibilities, particularly as we learn more about neurotransmitters and neural pathways involved over time. The idea of altering time perception is increasingly within the grasp of neuroscience, necessitating conversations about how it should—or shouldn’t—be used.

Ethical Considerations on Manipulating Time Perception

Moving toward a future in which manipulating time perception seems a distinct possibility, the ethics behind such manipulation remain contentious and unclear.

With clinical disorders, these temporal interventions may be straightforwardly considered beneficial, as they may reduce suffering. For instance, an individual undergoing chronic pain or a painful medical procedure might benefit from a sped-up internal clock (wherein the time of suffering is effectively reduced through dissociation); additionally, a patient in terminal care might want to compress their sense of time such that it passes more quickly and peacefully (Garcia-Romeu et al., 2016). In other cases, such as depression, a drug that speeds up time could make days feel as if they progress at a more “normal” pace and potentially reduce the suffering associated with slow days feeling interminable. Throughout cases such as these, however, patient autonomy and consent must be upheld: despite their potential as treatment for conditions requiring palliative care, the non-consensual use of these therapeutics would negate their benefit.

However, manipulating time perception might not always be the most ethically appropriate or even psychologically beneficial. For instance, patients with post-traumatic stress disorder may be further harmed by introducing temporal alterations, as these therapies may worsen their dissociation or prevent emotional processing, which requires a length of time unique to

each individual (Biggs et al., 2019). Additionally, in disorders with psychosis or mania, time perception may already be unstable, and interventions might lead to unintended, unknown consequences. Due to the current lack of understanding of how temporal processing works and might be modified, their implementation must be carefully considered to mitigate unintended clinical consequences, such as memory alteration (Kolber, 2011).

However, biotechnological advancements may go beyond treating illness and be further used in daily life. For instance, some people may wish to make time pass by more quickly while studying, where they retain the same amount of information, yet over a period that feels subjectively shorter. Alternatively, in a moment of exhilaration, such as a celebration, one might instead want to extend their sense of time to feel their joy for a longer duration. These cases, while improving everyday life, call for questions about safety and fairness: would these tools be accessible to all, or would they instead create new inequalities and exacerbate existing ones? An additional concern is that our sense of time is inherently unique to our identities—purposefully altering how we sense time might change our fundamental sense of self, especially if such tools are commodified and incorporated into widespread life (Johnson, 2025). From an existentialist or phenomenological perspective, the human experience of time forms a critical part of our being in the world, such that modifying it beyond what is considered natural may blur the boundaries between our truly authentic and our engineered experiences. If time-altering tools are commercialized and made prevalent, our perception of meaningful moments might be reduced simply to products that could be bought, sold, and modified at will.

“An additional concern is that our sense of time is inherently unique to our identities—purposefully altering how we sense time might change our fundamental sense of self”

Beyond daily life, significant discord on the ethics of using time-altering substances lies within the justice

system. Could manipulating time perceptions be used to carry out prison sentences? Some ethicists, like Rebecca Roache and her colleagues, have considered the concept of artificially extending prison sentences to carry out realistically shorter sentences (Roache, 2014). For example, a criminal could opt to serve a shorter sentence but take a drug that makes it seem to them as if 50 years passed by, enabling them to reintegrate into society while still being sufficiently punished. This practice would magnify the psychological punishment of imprisonment, without physically keeping individuals in prison as long. However, this idea challenges long-held beliefs surrounding humane punishment and individual rights, notions which would be challenged by the use of a temporal drug that might cause unintended, debilitating, counteracting consequences, detracting from the essence of justice.

Conclusion

While time perception manipulation seems like a distant reality, it sits at a unique intersection between the scientific possibilities and the ethics of our minds. Depending on its usage, it could benefit clinical populations that suffer from pain but also carry a significant risk of abuse, from exacerbating inequities in daily life to destabilizing ethical criminal justice. The groundwork of research on temporal perception continues to build daily, necessitating difficult questions as we push our understanding of the brain’s clock to further limits. As we venture deeper into the science of time perception, we must tread carefully, balancing innovation with ethics and recognizing that the manner in which we experience time is inseparable from our humanity.

References

Amadeo, M. B., Esposito, D., Escelsior, A., Campus, C., Inuggi, A., Pereira Da Silva, B., Serafini, G., Amore, M., & Gori, M. (2022). Time in schizophrenia: A link between psychopathology, Psychophysics and Technology. *Translational Psychiatry*, 12(1). <https://doi.org/10.1038/s41398-022-02101-x>

Apaydin, N., Üstün, S., Kale, E. H., Çelikag, I., Özgüven, H. D., Baskak, B., & Çiçek, M. (2018). Neural mechanisms underlying time perception and reward anticipation. *Frontiers in Human Neuroscience*, 12. <https://doi.org/10.3389/fnhum.2018.00115>

Biggs, Q. M., Ursano, R. J., Wang, J., Krantz, D. S., Carr, R. B., Wynn, G. H., Adams, D. P., Dacuyan, N. M., & Fullerton, C. S. (2019). Daily variation in post traumatic stress symptoms in individuals with and without probable post traumatic stress disorder. *BMC Psychiatry*, 19(1). <https://doi.org/10.1186/s12888-019-2041-7>

Fontes, R., Ribeiro, J., Gupta, D. S., Machado, D., Lopes-Júnior, F., Magalhães, F., Bastos, V. H., Rocha, K., Marinho, V., Lima, G., Velasques, B., Ribeiro, P., Orsini, M., Pessoa, B., Araujo Leite, M. A., & Teixeira, S. (2016). Time Perception Mechanisms at central nervous system. *Neurology International*, 8(1). <https://doi.org/10.4081/ni.2016.5939>

Fung, B. J., Sutlief, E., & Hussain Shuler, M. G. (2021). Dopamine and the interdependency of time perception and reward. *Neuroscience & Biobehavioral Reviews*, 125, 380–391. <https://doi.org/10.1016/j.neubiorev.2021.02.030>

Garcia-Romeu, A., Kersgaard, B., & Addy, P. H. (2016). Clinical applications of hallucinogens: A review. *Experimental and Clinical Psychopharmacology*, 24(4), 229–268. <https://doi.org/10.1037/pha0000084>

Johnson, M. (2025, February 7). *Why the neuroscience of identity is bound to time and memory*. Neuroscience Of. <https://www.neuroscienceof.com/human-nature-blog/identity-neuroscience-time-memory-self>

Kolber, A. (2011). Give memory-altering drugs a chance. *Nature*, 476(7360), 275–276. <https://doi.org/10.1038/476275a>

Merchant, H., Bartolo, R., Pérez, O., Méndez, J. C., Mendoza, G., Gámez, J., Yc, K., & Prado, L. (2014). Neurophysiology of timing in the hundreds of milliseconds: Multiple layers of neuronal clocks in the medial premotor areas. *Advances in Experimental Medicine and Biology*, 143–154. https://doi.org/10.1007/978-1-4939-1782-2_8

Paton, J. J., & Buonomano, D. V. (2018a). The neural basis of timing: Distributed mechanisms for diverse functions. *Neuron*, 98(4), 687–705. <https://doi.org/10.1016/j.neuron.2018.03.045>

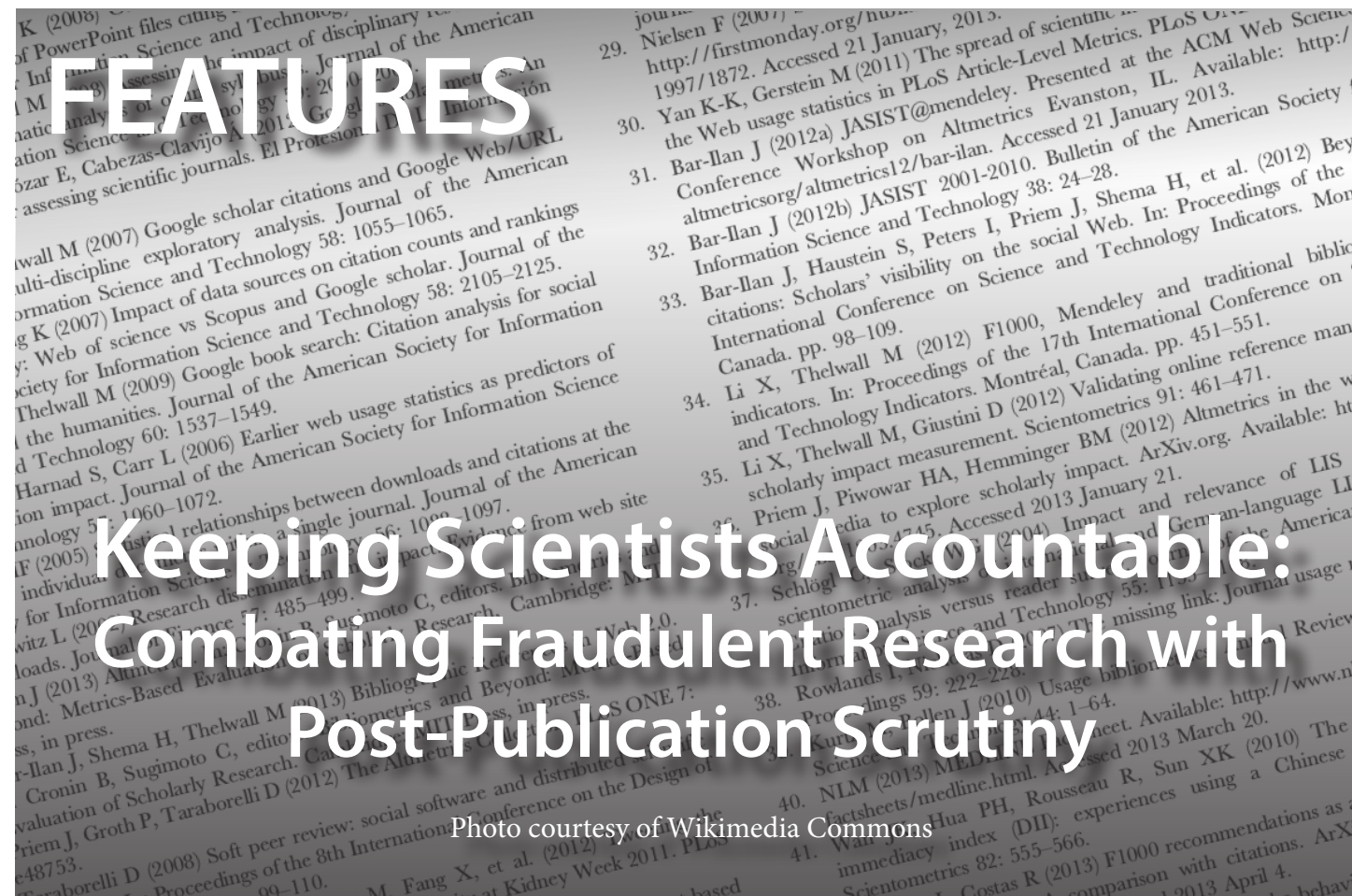
Paton, J. J., & Buonomano, D. V. (2018b). The neural basis of timing: Distributed mechanisms for diverse functions. *Neuron*, 98(4), 687–705. <https://doi.org/10.1016/j.neuron.2018.03.045>

Penhune, V. B., & Zatorre, R. J. (2019). Rhythm and time in the premotor cortex. *PLOS Biology*, 17(6). <https://doi.org/10.1371/journal.pbio.3000293>

Roache, R. (2014, March 25). *The future of punishment: A clarification: Practical ethics. The future of punishment: a clarification*. <https://blog.practicaethics.ox.ac.uk/2014/03/the-future-of-punishment-a-clarification/>

Shebloski, K. L., & Broadway, J. M. (2016a). Commentary: Effects of psilocybin on time perception and temporal control of behavior in humans. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00736>

Terhune, D. B., Russo, S., Near, J., Stagg, C. J., & Cohen Kadosh, R. (2014). GABA predicts time perception. *The Journal of Neuroscience*, 34(12), 4364–4370. <https://doi.org/10.1523/jneurosci.3972-13.2014>



Keeping Scientists Accountable: Combating Fraudulent Research with Post-Publication Scrutiny

Photo courtesy of Wikimedia Commons

Antonino J.P. Libarnes '28

Introduction

In recent years, scientific integrity has become increasingly difficult to maintain. Multiple scandals have shaken public confidence in the scientific process, including within the Harvard community. In 2023, research integrity watchdog Data Colada accused Harvard Business School professor Francesca Gino of fabricating data in several behavioral research papers, leading to her being put on administrative leave and initiating a \$25 million lawsuit against Harvard (Hamid & Yuan, 2023; Simonsohn et al., 2023). Additionally, in 2024, Sholto David of For *Better Science* identified widespread scientific misconduct among the leadership of the Dana-Farber Cancer Institute, a cancer treatment and research center affiliated with Harvard Medical School (Mueller, 2024). While extreme, these cases are symptoms of deeper, systemic issues in how research is produced, reviewed, and published.

Scientific publishing operates under enormous pressure. Over 3 million articles in science and engineering were

published in 2022, and a 2021 analysis identified over 75,000 unique journals (National Science Board, National Science Foundation, 2023; Singh et al., 2021). This is a jarring difference from the 1.3 million published articles in 23,750 journals in 2006. (Björk et al., 2009). A major factor driving the volume of both research articles and journals is the long-standing “publish or perish” mentality present in science; research output is often a major factor in faculty hiring and promotion decisions, as well as graduate program admissions and post-doctoral research appointments. In a 2010 poll, over two-thirds of researchers claimed that research metrics were used in making hiring, promotion, or tenure appointments (Abbott et al., 2010). As a result, researchers may prioritize generating a high *quantity* of publications rather than focusing on their *quality* (Rawat & Meena, 2014).

The pressure to increase output and the associated rise in publication volume risks pushing fraudulent work through publication channels. The sheer volume of research articles being published is too much for the peer review system to handle. When submitted to a standard journal, a manuscript is initially reviewed by the journal's editor, after which it is passed on to a team of two to three peer reviewers for additional feedback (Publons, 2018). If all papers were published via this process, at least 6.6 million peer reviews had to be conducted

for the 3.3 million papers published in 2022. Recruiting peer reviewers to fulfill this demand is becoming increasingly difficult, especially considering they are usually full-time researchers themselves (Dance, 2023). Consequently, not all submitted manuscripts receive proper scrutiny, making it easier for misconduct to pass through the filter of peer review. Additionally, because some manuscripts within the scientific literature are contaminated by fraudulent data, the practices built upon them can lead to incorrect conclusions (“10.Q. Scientific Session,” 2024). It is critical to identify such articles as they can harm ongoing and future research.

Pipelines for Low-Quality Research

The proliferation of low-quality research is streamlined by various organizations partaking in dubious publishing practices. The high demand for publication has led to the rise of organizations that unethically ease the publication process for profit. Many researchers, to earn promotions and tenure, utilize these methods to quickly increase their citation and publication metrics.

One way for researchers to easily publish papers is through the use of paper mills, which are companies that sell manuscripts using fake data to researchers (Christopher, 2021). These organizations churn out fraudulent and even plagiarized manuscripts for a fee, allowing scientists to pay their way out of doing actual research. Articles produced by paper mills often cite each other, fabricate image data, and utilize nonsensical but convincing figures to seem legitimate. A 2023 analysis found that almost 2% of papers published in 2022 showed signs of being produced by a paper mill, a large difference from an estimated <0.1% of papers published in 2000 (Van Noorden, 2023). In 2023, the former publisher Hindawi retracted over 8,000 articles that were produced by paper mills (Kincaid, 2023).

Some publishers and journals deliberately provide avenues for low-quality research. “Predatory journals” accept almost any paper for a high publication fee, often priced in the thousands of dollars (Beall, 2012). This model facilitates misconduct, allowing researchers to publish several papers with little review. Meanwhile, this model harms honest researchers. Jeffery Beall, the librarian who coined the term “predatory journal,” states that “when a researcher’s work is published alongside articles that are plagiarized [...] it becomes tainted by association” (Beall, 2012). Cabell’s Predatory Reports, an updated database of predatory journals, listed over 15,000 predatory journals in 2021, up from 12,000 in 2019 (*The Source / Mountain to Climb*, 2021).

Post-Publication Peer Review

In response to the limitations of traditional peer review and the rise of predatory journals, post-publication peer review (PPPR) has emerged as a way to evaluate published literature. PPPR allows for ongoing, transparent scrutiny and commentary of research articles, in contrast to the closed-door process of traditional peer review (Hunter, 2012). This process can act as a second layer of quality control in case traditional peer review, or the lack thereof, fails to prevent fraudulent works from being published.

One of the most prominent platforms for PPPR is PubPeer, a site where researchers can publicly comment on others’ published research (Townsend, 2013). Any indexed article is available to be commented on, and researchers can comment anonymously if they choose to do so. The platform is often used for identifying flaws in papers and instances of manipulation or fraudulent data. PubPeer has received over 300,000 comments since its founding; 57,000 comments were written in 2024 alone, a drastic increase from the little over 2,100 written in 2013 (Einstein Foundation, 2024). Some publishers and journals, such as *PLOS*, provide their own comment sections for the same purpose (Wakeling et al., 2020).

The benefits of PPPR are significant. These commenting systems can help to identify research issues at any point after publication. In February 2025, a Nobel Prize laureate’s article published in 2017 was retracted after PubPeer comments highlighted possible data manipulation (Travis, 2025). Real-time commenting promotes quick responses and corrections, as opposed to the long and obscured process of peer review. Moreover, PubPeer’s option for anonymity has protected commenters from retaliation, evidenced by a Michigan court case where a researcher implicated in PubPeer comments attempted to sue the commenters (Servick, 2015). An appeals court ruled that PubPeer did not have to identify the anonymous commenters, successfully preventing repercussions toward them (McCook, 2016). PPPR encourages researchers to partake in ethical practices, since unethical behavior is more readily exposed.

However, PPPR platforms also hold controversies. While PubPeer is anonymous, other comment platforms and journals may not be. As such, colleagues may fear retaliation if they comment on the articles of those they work with or those who hold positions over them (Daungsupawong & Wiwanitkit, 2024). The change in scientific discourse fostered by PPPR has concerned some, as it increasingly drifts away from overall discussion towards scrutiny of minute details (Blatt, 2015). Michael Blatt, editor-in-chief of the journal *Plant Physiology*, states that PPPR comments

often “do no more than flag perceived faults and query the associated content.” Additionally, some researchers have made accusations of defamation or cyberstalking from commenters who have criticized their work. The Michigan court case that tested PubPeer’s anonymity protections involved a researcher claiming that comments on his articles were defamatory, preventing him from receiving a position at the University of Mississippi (Koziol, 2016). The former Department of Medicine chair and physician-in-chief emeritus at the Brigham and Women’s Hospital, Joseph Loscalzo, claims to have received malicious emails and comments via PubPeer (Joelving, 2023).

Despite these challenges, PPPR remains a powerful tool to promote self-correction in science. As flawed work becomes more commonplace with the rising volume of research produced, post-publication scrutiny remains important as a safeguard for research integrity.

Other Means of Scrutiny

Beyond PPPR, a body of independent watchdogs, databases, and individuals helps to keep research accountable. By operating outside of the traditional publishing structure, these sources offer a more agile way to uncover and publicize questionable research practices.

One of the most well-known examples of this is the website *Retraction Watch*. This website, updated daily, disseminates information about retracted or questionable articles and publishes journalistic investigations on scientific integrity (Balyakina, 2022). These articles offer visibility and transparency into the research process, which is often very opaque. This organization also offers an online database of retracted articles, allowing other researchers to monitor scientific integrity at scale (“*Retraction Watch* Database User Guide,” 2018). Other websites and blogs offer similar oversight of unethical practices. The website *For Better Science* is another independent scientific integrity watchdog. *For Better Science* identified image manipulation in several papers published by leading researchers at the Dana-Farber Cancer Institute, leading to high-profile retractions and resignations (David, 2024). Another website, *Data Colada*, was critical in identifying the data manipulation present in several behavioral science articles published by Professor Gino of Harvard Business School (Simonsohn et al., 2023).

Alongside institutional efforts, individuals have also helped in maintaining oversight. Scientist Elizabeth Bik has analyzed many articles to identify image manipulation, receiving the Einstein Foundation Award for promoting quality in research (Einstein Foundation, 2024; Vidal & Raoult, 2025). Image manipulation may occur in key data

How an image sleuth uncovered possible tampering

Vanderbilt University neuroscientist Matthew Schrag found apparently falsified images in papers by University of Minnesota, Twin Cities, neuroscientist Sylvain Lesné, including a 2006 paper in *Nature* co-authored with Karen Ashe and others. It linked an amyloid-beta (Aβ) protein, Aβ*56, to Alzheimer’s dementia.

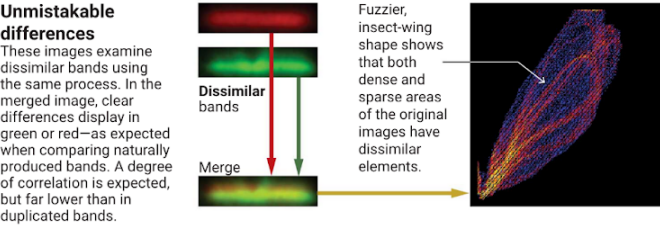
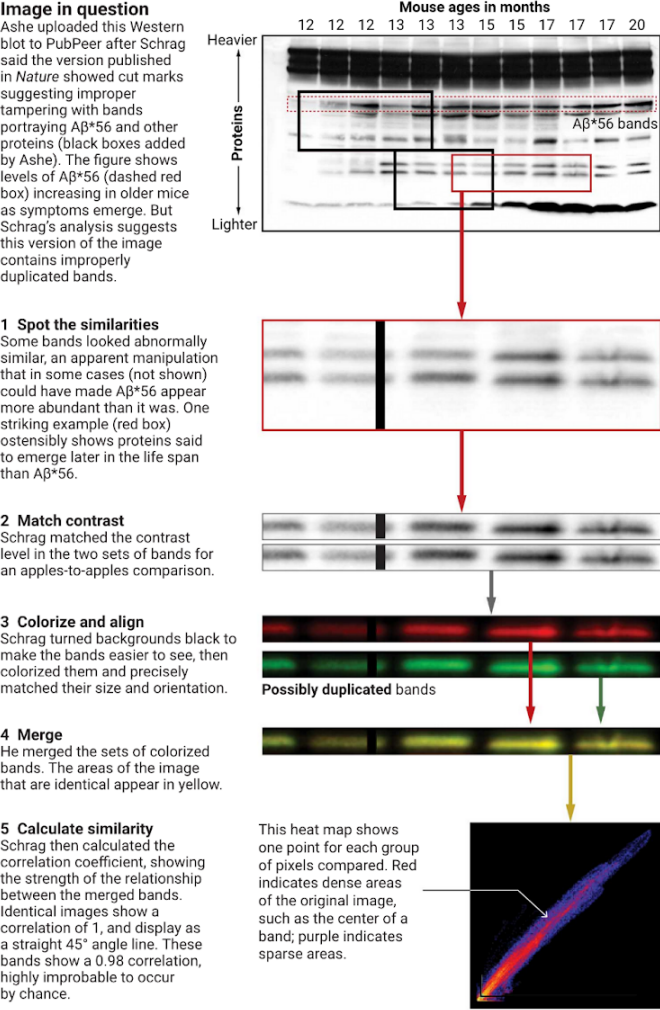


Figure 1. Bands on a key Western blot appeared to be duplicated. Graphic from Bickel/*Science*. Data from S. Lesne et al. (from Piller, 2022)

sources, such as Western blots, in an attempt to skew the results and conclusions of an experiment. The retraction of a highly-cited Alzheimer’s disease research article highlights this issue (Lesné et al., 2006). This article extended upon the prevailing theory that clumps of proteins known as amyloid-beta are the cause of Alzheimer’s disease: fraudulent Western blot images suggested that a certain protein subtype was a major contributor (Piller, 2022). These manipulations were identified by individual watchdogs analyzing the images, and the article’s retraction made several news headlines.

Conclusion

The rise of post-publication analysis via PPPR and independent watchdogs signals a shift to a more transparent and self-correcting culture of science. Unlike traditional peer review, which is often opaque and limited to a small group of reviewers, these platforms encourage open and anonymous critique, normalizing the questioning of published work. This openness also encourages researchers to be more rigorous and make their data accessible, knowing that their work could be subject to public scrutiny.

By raising the standards for scientific rigor and transparency, post-publication review counteracts the risks of fraudulent science created by the “publish or perish” culture. PPPR ensures that quality, not just productivity, drives scientific advancement. In doing so, it safeguards scientific integrity and supports a culture where well-substantiated research is highly valued.

References

10.Q. Scientific session: Evolving landscape of scientific publishing practices: implications for public health. (2024). *The European Journal of Public Health*, 34(Suppl 3), ckae144.676. <https://doi.org/10.1093/eurpub/ckae144.676>

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Metrics: Do metrics matter? *Nature*, 465(7300), 860–862. <https://doi.org/10.1038/465860a>

Balyakina, E. A. (2022). Retraction Watch: A tool for informing academia about ethical violations in publications. *Science Editor and Publisher*, 6(2), Article 2. <https://doi.org/10.24069/SEP-21-12>

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, 489(7415), 179–179. <https://doi.org/10.1038/489179a>

Björk, B.-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: Yearly volume and open access availability. *Information Research*, 14(1), 391.

Blatt, M. R. (2015). Vigilante Science. *Plant Physiology*, 169(2), 907–909. <https://doi.org/10.1104/pp.15.01443>

Christopher, J. (2021). The raw truth about paper mills. *FEBS Letters*, 595(13), 1751–1757. <https://doi.org/10.1002/1873-3468.14143>

Dance, A. (2023). Stop the peer-review treadmill. I want to get off. *Nature*, 614(7948), 581–583. <https://doi.org/10.1038/d41586-023-00403-8>

Daungsupawong, H., & Wiwanitkit, V. (2024). Evaluating the pros and cons of anonymous commenting on PubPeer. *Formosan Journal of Surgery*, 57(5), 224. <https://doi.org/10.1097/FS9.0000000000000142>

David, S. (2024, January 2). Dana-Farberications at Harvard University. *For Better Science*. <https://forbetterscience.com/2024/01/02/dana-farberications-at-harvard-university/>

Einstein Foundation. (2024). *Elisabeth Bik – Einstein Foundation Award*. <https://award.einsteinfoundation.de/award-winners-finalists/recipients-2024/elisabeth-bik>

Einstein Foundation. (2024). *PubPeer – Einstein Foundation Award*. <https://award.einsteinfoundation.de/award-winners-finalists/recipients-2024/pubpeer>

Hamid, R., & Yuan, C. (2023, August 3). Embattled by Data Fraud Allegations, Business School Professor Francesca Gino Files Defamation Suit Against Harvard. *The Harvard Crimson*. <https://www.thecrimson.com/article/2023/8/3/hbs-prof-lawsuit-data-fraud-defamation/>

Hunter, J. (2012). Post-Publication Peer Review: Opening Up Scientific Conversation. *Frontiers in Computational Neuroscience*, 6. <https://doi.org/10.3389/fncom.2012.00063>

Joelving, F. (2023, December 4). Cyberstalking pits Harvard professor against PubPeer. *Retraction Watch*. <https://retractionwatch.com/2023/12/04/cyberstalking-pits-harvard-professor-against-pubpeer/>

Kincaid, E. (2023, December 19). Hindawi reveals process for retracting more than 8,000 paper mill articles. *Retraction Watch*. <https://retractionwatch.com/2023/12/19/hindawi-reveals-process-for-retracting-more-than-8000-paper-mill-articles/>

Koziol, M. (2016, August 12). Meet the researcher with 13 retractions who’s trying to sue PubPeer commenters: Fazlul Sarkar. *Retraction Watch*. <https://retractionwatch.com/2016/08/12/meet-the-researcher-who-tried-to-take-on-pubpeer-commenters-fazlul-sarkar/>

Lesné, S., Koh, M. T., Kotilinek, L., Kaye, R., Glabe, C. G., Yang, A., Gallagher, M., & Ashe, K. H. (2006). RETRACTED ARTICLE: A specific amyloid-B protein assembly in the brain impairs memory. *Nature*, 440(7082), 352–357. <https://doi.org/10.1038/nature04533>

McCook, A. (2016, December 7). PubPeer wins appeal of court ruling to unmask commenters. *Retraction Watch*. <https://retractionwatch.com/2016/12/07/pubpeer-wins-appeal-court-ruling-unmask-commenters/>

Mueller, B. (2024, January 22). Top Cancer Center Seeks to Retract or Correct Dozens of Studies. *The New York Times*. <https://www.nytimes.com/2024/01/22/health/dana-farber-cancer-studies-retractions.html>

National Science Board, National Science Foundation. (2023). *Publications Output: U.S. Trends and International Comparisons. Science and Engineering Indicators 2024*. National Center for Science and Engineering Statistics. <https://ncses.nsf.gov/pubs/nsb202333/>

Piller, C. (2022). *Potential fabrication in research images threatens key theory of Alzheimer’s disease* [Dataset]. <https://doi.org/10.1126/science.ade0209>

Publons. (2018). *Publons’ Global State Of Peer Review 2018* (0 ed.). Publons. <https://doi.org/10.14322/publons.GSPR2018>

Rawat, S., & Meena, S. (2014). Publish or perish: Where are we heading? *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, 19(2), 87–89.

Retraction Watch Database User Guide. (2018, October 23). *Retraction Watch*. <https://retractionwatch.com/retraction-watch-database-user-guide/>

Servick, K. (2015). Michigan judge asks PubPeer to turn over anonymous user information. *Science*. <https://doi.org/10.1126/science.aab0354>

Simonsohn, U., Nelson, L., & Simmons, J. (2023, June 17). [109] *Data Falsificada (Part 1): “Clusterfake.”* Data Colada. <http://datacolada.org/109>

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, 126(6), 5113–5142. <https://doi.org/10.1007/s11192-021-03948-5>

The Source / Mountain to climb. (2021, September 1). The Source. <https://blog.cabells.com/2021/09/01/mountain-to-climb/>

Townsend, F. (2013). Post-publication Peer Review: PubPeer. *Editors’ Bulletin*, 9(3), 45–46. <https://doi.org/10.1080/17521742.2013.865333>

Travis, K. (2025, February 14). ICYMI: Second paper by Nobel laureate Thomas Südhof retracted. *Retraction Watch*. <https://retractionwatch.com/2025/02/14/icymi-second-retraction-nobel-thomas-sudhof/>

Van Noorden, R. (2023). How big is science’s fake-paper problem? *Nature*, 623(7987), 466–467. <https://doi.org/10.1038/d41586-023-03464-x>

Vidal, C., & Raoult, D. (2025). *Report on the allegations of Mrs. BIK*. AMU - Aix Marseille Université. <https://hal.science/hal-04926449>

Wakeling, S., Willett, P., Creaser, C., Fry, J., Pinfield, S., Spezi, V., Bonne, M., Founti, C., & Medina Perea, I. (2020). ‘No comment’? A study of commenting on PLOS articles. *Journal of Information*

FEATURES

Rethinking Alzheimer's: A New Autoimmune Perspective

Photo courtesy of Dmitri Volochek, Adobe Stock.

Hiteyjit Singh Gujral '27

Visiting Undergraduate Student

Introduction

Alzheimer's Disease (AD) is a widespread neurodegenerative disease that damages collections of neurons throughout the brain. As a subset of dementia, AD often manifests as symptoms that affect the functioning of the brain, such as memory loss, cognitive decline, and the impairment of several other functions, hampering an individual's ability to perform daily activities (Alzheimer's Association, 2025). As of 2024, dementia cases, of which between 60% and 70% are caused by AD, have risen to a total of 6.9 million within the United States and 55 million across the world (Alzheimer's Association, 2024). With around 60% of the global burden of AD falling disproportionately on developing regions with limited access to symptom-alleviating medications, AD's impact has necessitated heavy

attention and investment to develop our understanding of this disease (World Health Organization, 2025). For decades, researchers have searched for answers, debating the underlying mechanisms of Alzheimer's. Now, after years of uncertainty, a compelling new theory is gaining traction: the autoimmune hypothesis.

The "Current" Mechanism

AD's distinction from other forms of dementia lies in its identity as a protein-misfolding disease, in which two protein oligomers are primarily implicated: Amyloid-Beta ($A\beta$) and tau. In healthy brains, $A\beta$ —part of the larger amyloid precursor protein—contributes to synaptic plasticity and transmission, and tau aids in stabilizing the neuron's cytoskeleton (Hefter et al., 2020). However, in AD, tau proteins become insoluble within the cell, forming 'tangles' and disrupting the cell's internal transportation mechanism. Simultaneously, $A\beta$ begins to clump together to form 'plaques' surrounding the outside of the neuron (Bloom, 2014). The misfolding of these proteins is part of the

popular "amyloid cascade" hypothesis. This theory frames $A\beta$ misfolding and accumulation as the initiating event in AD's pathogenesis, where tau pathology, inflammation, and neuronal death occur as a cascade of downstream effects (Karran et al., 2011). While this theory has long enjoyed popularity, limited efficacy of theory-derived therapeutics, along with emerging data on immune activity, have led scientists to consider alternative explanations for AD's etiology (Kepp et al., 2023).

The New Perspective

Recent scientific investigations and alternative interpretations of existing data suggest that AD is not simply a disorder of protein accumulation, but also an autoimmune condition, where the brain's defense mechanisms react against healthy tissue. Although $A\beta$ accumulation remains central to our understanding of AD, it is now being presented as a potential trigger, target, and even mediator of immune processes, rather than as the initiating event of AD (Chen & Holtzman, 2022).

The immune system is a critical component in detecting and removing disease from our bodies. Notably, however, the brain has its own independent immune system, operating relatively separately from the rest of the body. While similarities do exist between brain immunity and peripheral immunity, specialized structures such as the blood-brain barrier (BBB) ensure that brain function remains physically protected from pathogens circulating in the bloodstream (Hubbard & Binder, 2016). Such an "immune privilege" allows the brain to be largely protected from peripheral dangers. Since peripheral and central immune components are isolated from each other, the usual cross-interaction occurs through buffer mediums (eg. cerebrospinal fluids) and is limited to communication rather than a sharing of resources to modulate responses (Mapunda et al., 2022). It is a trade-off to keep the brain protected.

One particular view of AD as an autoimmune disease, called the AD² hypothesis, presents a model of the disease as a disorder with the brain's short-term, rapid immune response system (innate immunity) (Weaver, 2022). The study frames $A\beta$ as part of the brain's innate response system,

and suggests its potential function as an active signalling, immunoregulatory, and antimicrobial protein. However, these properties of $A\beta$ further appear to drive pathogenic neuroinflammation by increasing microglial activity and nonspecific programmed neuronal death (apoptosis). Additionally, while $A\beta$ could be released for regular immunomodulation in response to damage-associated molecular pattern-stimulating events, it may further promote an antimicrobial response without the presence of such a pathogen; critically, observed physiological similarities between neurons and bacteria appear to drive an occasional mistargeting of neurons through $A\beta$'s antimicrobial properties, and eventual necrotic cell death. Further in vitro studies have also shown this autoimmune-induced necrosis to release products that drive a further increase in extracellular $A\beta$ accumulation (Meier-Stephenson et al., 2022). Enhanced microglial activity also appears

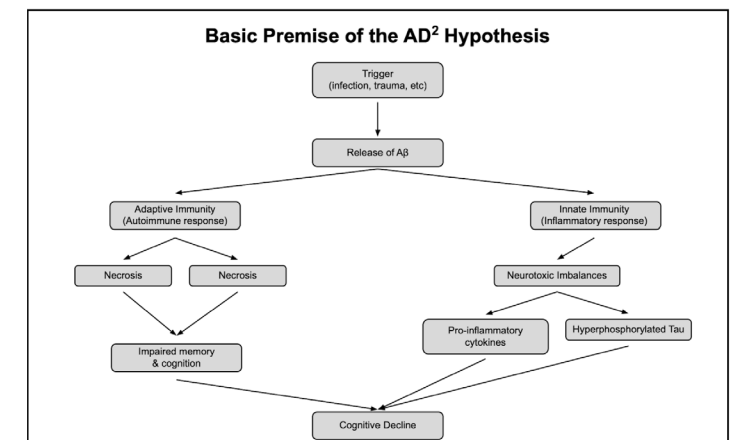


Figure 1. Basics of the AD² model as inspired by Weaver et al. (2022).

to accentuate hyperphosphorylated Tau (p-Tau). With maladaptive neuroinflammatory responses as a result of $A\beta$ accumulation, a vicious cycle forms where an increase in p-Tau causes further pro-inflammatory microglial activation, triggering a cycle of further Tau hyperphosphorylation (Jorfi et al., 2023).

Apart from the AD² hypothesis, some scientists also suggest that additional aspects of the brain and peripheral immune systems are at play in AD. Astrocytes, originally meant for neuroprotection through functions such as clearing excess neurotransmitters, appear to be neurotoxic when overactive, reducing their ability to cleave and clear excess $A\beta$ that has accumulated (Wei & Morrison, 2019; Kim et al., 2018). Additionally, some investigations find that $A\beta$ has neurotoxic effects on astrocytes through a disruption of their ionic balance, inducing an increase in intracellular ions and eventual oxidative stress-induced neuronal death (Abramov & Duchon, 2005).

“Data suggest AD is not simply a disorder of protein accumulation, but also an autoimmune condition”

Interestingly, peripheral immune system structures have also been implicated in AD. Recent studies suggest that T-cells—specialized white blood cells central to the adaptive immune response—can ‘infiltrate’ brain tissue under certain

“Reconceptualization of AD as an autoimmune disorder marks an exciting frontier in neurodegenerative research”

conditions (Jorfi et al., 2023). In post-mortem analyses of AD rat models, certain T-cells were found adhering to blood vessels in the brain, contributing to memory impairment and an increase in neuronal cell death (Jorfi et al., 2023). Similar findings in both human subjects and transgenic animal models reveal correlations between disease progression and the infiltration of cytotoxic T-cells into the brain parenchyma (Jorfi et al., 2023). Yet, the story isn't entirely one-sided. In some mouse models, specific T-cell subtypes have been linked to reduced cognitive impairment and enhanced immunoregulatory responses against AD (Jorfi et al., 2023). These seemingly contradictory roles suggest that T-cell involvement in AD may depend heavily on the context and conditions in which it is involved, as well as the T-cell subtypes involved.

Neuroinflammation in AD may also modulate the production of certain metabolic byproducts, potentially affecting disease symptoms. Tryptophan, an amino acid obtained through the diet, usually supports the brain’s immunity through metabolic pathways that produce microglial-activating metabolites (Meier-Stephenson et al., 2022). However, as noted by Meier-Stephenson et al. (2022), AD patients appear to have elevated levels of downstream neurotoxic tryptophan metabolites (e.g., quinolinic acid). Certain compounds within tryptophan's catabolic cascade, such as IDO-1, have also been connected to inflammatory signalling molecules (cytokines), with AD-associated inflammatory environments influencing IDO-1 and production of neurotoxic tryptophan metabolites. However, some tryptophan-derived metabolites, like those from the serotonergic pathway, can be protective, suggesting that tryptophan and its derivatives play a nuanced and varied role in AD progression. The effects of the immune environment on AD appears to extend beyond metabolism; for instance, secretion of the immune-related marker SPP1 triggers microglia to engulf synapses. According to research by De

Schepper et al. (2023), this early synaptic pruning by microglia may critically contribute to neurodegenerative processes, highlighting potential therapeutic targets for preventing synapse loss. Ultimately, the immune environment likely determines which properties are exhibited, where it may be both neurotoxic and therapeutic.

Epidemiological evidence also appears to support the autoimmune hypothesis. Individuals with existing autoimmune conditions appear approximately 70 percent more likely to develop AD (Ramey et al., 2025), which is in line with immune-related triggers as described above. Although this finding by no means establishes a causal relationship between autoimmune disease and AD, it is further data implicating the immune system as playing some role in AD.

Discussion

If AD is indeed an autoimmune disorder, the therapeutic landscape for its treatment could change dramatically. AD immunotherapy could shift treatment toward precisely modulating neuronal immune responses rather than merely targeting protein deposits. This opens the door for therapies designed to target known inflammatory pathways implicated in neurodegeneration, such as cytokine-driven inflammation, which could be a more effective way to slow down AD progression. Given the involvement of microglia and astrocytes in AD, therapies aimed at their selective modulation may also offer significant benefits by reducing harmful inflammation and cerebrospinal fluid (CSF) dysfunction. Meanwhile, therapeutic strategies based on metabolism in AD are also gaining ground, especially those centered around tryptophan with its dual neurotoxic or neuroprotective role. Specific metabolites, such as 5-hydroxytryptamine, have shown exciting promise by not only limiting Aβ aggregation but also modulating inflammatory cytokines like IL-1β (Meier-Stephenson et al., 2022; Savonije & Weaver, 2023). Biomarker identification, such as tracking SPP1 secretions and subtle neuroinflammation, could further enhance early detection of and personalized treatment strategies for AD. Lastly, targeted cognitive therapies, building upon immune-pathway insights, might complement biological interventions to optimize patient outcomes.

However, one needs to approach these interesting revelations with caution—there are current limitations with this autoimmune paradigm. The research that grounds it is relatively recent, and further lacks substantial experimental data, particularly in humans. The brain's immune system is difficult to study non-invasively, and translating findings

from animal models to human pathology presents significant challenges. Although correlational evidence and interpretation of pre-existing experimental data indeed have begun to surface, more time and research will be needed to determine the exact role the autoimmune pathway plays with respect to AD, including whether it is a sole trigger of the disease or a result of a prior initial protein cascade. Nevertheless, pinpointing these correlations has a significant impact on shifting the neuroscientific world's perspective from solely an amyloid-cascade-based hypothesis, and instead prompts us to consider a greater mechanism at play. As described, these changes in understanding of AD development and progression open the possibility of better targeted therapies and earlier biomarkers to detect AD pathologies.

Conclusion

The reconceptualization of AD as an autoimmune disorder marks an exciting frontier in neurodegenerative research. Rather than viewing Aβ as simply a rogue protein, these theories hold it as a key element in the complex neuronal immune response gone awry—a symptom rather than the ultimate cause of AD.

This shift in perspective offers new hope for the millions affected by AD worldwide, given the new therapeutic avenues that it opens. AD may represent a web of interconnected pathological processes, requiring multi-target therapeutic approaches addressing immune dysregulation, protein aggregation, and metabolic dysfunction simultaneously. Given the breadth of these processes, investigators and physicians should pursue integrative approaches focusing on the development of therapeutics targeting multiple aspects of AD simultaneously.

As research continues, our understanding of the brain, its defense mechanisms, and the delicate balance between protection and pathology evolves in fascinating ways, reminding us that even in well-studied diseases, unexpected doors to discovery remain unopened.

References

Abramov, A. Y., & Duchen, M. R. (2005). The role of an astrocytic NADPH oxidase in the neurotoxicity of amyloid beta peptides. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1464), 2309–2314. <https://doi.org/10.1098/rstb.2005.1766>.

Alzheimer's Association. (2024). 2024 Alzheimer's Disease Facts and Figures. *Alzheimer's & Dementia*, 20(5), 3708–3821. <https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13809>.

Alzheimer's Association. (2025). What Is Alzheimer's Disease? Alzheimer's Disease and Dementia; Alzheimer's Association. <https://www.alz.org/alzheimers-dementia/what-is-alzheimers>.

Bloom, G. S. (2014). Amyloid- and tau: the Trigger and Bullet in Alzheimer Disease Pathogenesis. *JAMA Neurology*, 71(4), 505–508. <https://doi.org/10.1001/jamaneurol.2013.5847>.

Chen, X., & Holtzman, D. M. (2022). Emerging roles of innate and adaptive immunity in Alzheimer's disease. *Immunity*. <https://doi.org/10.1016/j.immuni.2022.10.016>.

De Schepper, S., Ge, J. Z., Crowley, G., Ferreira, L. S. S., Garceau, D., Toomey, C. E., Sokolova, D., Rueda-Carrasco, J., Shin, S.-H., Kim, J.-S., Childs, T., Lashley, T., Burden, J. J., Sasner, M., Sala Frigerio, C., Jung, S., & Hong, S. (2023). Perivascular cells induce microglial phagocytic states and synaptic engulfment via SPP1 in mouse models of Alzheimer's disease. *Nature Neuroscience*, 26(3), 406–415. <https://doi.org/10.1038/s41593-023-01257-z>.

Gregersen, N., Bross, P., & Bolund, L. A. (2003). Conformational diseases. *Ugeskrift for Laeger*, 165(8), 801–805. <https://pubmed.ncbi.nlm.nih.gov/12625122/>.

Hefter, D., Ludewig, S., Draguhn, A., & Korte, M. (2020). Amyloid, APP, and Electrical Activity of the Brain. *The Neuroscientist*, 26(3), 231–251. <https://doi.org/10.1177/1073858419882619>.

Hubbard, J. A., & Binder, D. K. (2016). Inflammation. Astrocytes and Epilepsy, 313–342. <https://doi.org/10.1016/b978-0-12-802401-0.00013-2>.

Jorfi, M., Maaser-Hecker, A., & Tanzi, R. E. (2023). The neuroimmune axis of Alzheimer's disease. *Genome Medicine*, 15(1). <https://doi.org/10.1186/s13073-023-01155-w>.

Karran, E., Mercken, M., & Strooper, B. D. (2011). The Amyloid Cascade Hypothesis for Alzheimer's disease: an Appraisal for the Development of Therapeutics. *Nature Reviews Drug Discovery*, 10(9), 698–712. <https://doi.org/10.1038/nrd3505>.

Kepp, K. P., Robakis, N. K., Høiland-Carlsen, P. F., Sensi, S. L., & Vissel, B. (2023). The amyloid cascade hypothesis: an updated critical review. *Brain: A Journal of Neurology*, 146(10), awad159. <https://doi.org/10.1093/brain/awad159>.

Kim, Y. S., Jung, H. M., & Yoon, B.-E. (2018). Exploring glia to better understand Alzheimer's disease. *Animal Cells and Systems*, 22(4), 213–218. <https://doi.org/10.1080/19768354.2018.1508498>.

Mapunda, J. A., Tibar, H., Regragui, W., & Engelhardt, B. (2022). How Does the Immune System Enter the Brain? *Frontiers in Immunology*, 13. <https://doi.org/10.3389/fimmu.2022.805657>

Meier-Stephenson, F. S., Meier-Stephenson, V. C., Carter, M. D., Meek, A. R., Wang, Y., Pan, L., Chen, Q., Jacobo, S., Wu, F., Lu, E., Simms, G. A., Fisher, L., McGrath, A. J., Fermo, V., Barden, C. J., Clair, H. D. S., Galloway, T. N., Yadav, A., Campagna-Slater, V., & Hadden, M. (2022). Alzheimer's disease as an autoimmune disorder of innate immunity endogenously modulated by tryptophan metabolites. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 8(1). <https://doi.org/10.1002/trc2.12283>.

Ramey, G. D., Tang, A., Thanaphong Phongpreecha, Yang, M. M., Woldemariam, S. R., Oskotsky, T. T., Montine, T. J., Allen, I., Miller, Z. A., Nima Aghaeepour, Capra, J. A., & Sirota, M. (2025). Exposure to autoimmune disorders is associated with increased Alzheimer's disease risk in a multi-site electronic health record analysis. *Cell Reports Medicine*, 101980–101980. <https://doi.org/10.1016/j.xcrm.2025.101980>.

Savonije, K., & Weaver, D. F. (2023). The Role of Tryptophan Metabolism in Alzheimer's Disease. 13(2), 292–292. <https://doi.org/10.3390/brainsci13020292>.

Weaver, D. F. (2022). Alzheimer's disease as an innate autoimmune disease (AD 2): A new molecular paradigm. *Alzheimer's & Dementia*. <https://doi.org/10.1002/alz.12789>.

Wei, D. C., & Morrison, E. H. (2019, September 9). Histology, Astrocytes. Nih.gov; StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK545142/>.

World Health Organization. (2025, March 31). Dementia. World Health Organization; World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/dementia>.

Delivery. *Nature News*, 18, 358–378.

Xue, Q, et al. (2016). Mir-9 and Mir-124 synergistically affect regulation of dendritic branching via the AKT/GSK3 pathway by targeting rap2a. *Scientific Reports*, 6(1).

Zhang, S., et al. (2021). The risks of Mirna Therapeutics: In A drug target perspective. *Drug Design, Development and Therapy*, 15, 721–733.

FEATURES

Aristotle's Genome: From the Origins of Form to Modern Code

Image courtesy of Ernst Haeckel.

Krishna S. Rajagopal '28

Form and Becoming

Creation holds immense power. Parallel to the construction of this very sentence, where letters, once meaningless, gain shape and meaning through composition, biological creation brings form to formlessness. A single cell unfolds into a full being not by chance, but through an ordered logic of development. In both acts, form emerges not by accident, but by virtue of a shaping principle embedded in the process.

Long before the discovery of the double helix, Aristotle sought to define the principle of developmental organization. In his treatise *On the Generation of Animals*, he proposed that development is not the result of the rote accumulation of biological parts, but rather a gradual actualization of life from matter, an unfolding governed by intelligible structure and internal purpose (Aristotle, 1943). Though scientifically flawed in its

specifics, Aristotle's model offered a conceptual structure so coherent that it endured for over two thousand years.

The operation of biology today is understood within a different metaphysical framework. It speaks of gene regulation, epigenetics, and morphogen gradients in favor of *anima*, the soul, or *telos*, purpose. But the logic of developmental unfolding remains. Beneath the language of signaling cascades and transcription factors lies a structure of thought that still echoes Aristotle's words.

The Embryo and the Soul

Human development begins when sperm meets egg, the formation of the zygote. It rapidly divides, forming a 32-cell *morula* ('mulberry'), quickly becoming a hollow blastocyst, then an embryo that flattens and folds in on itself, laying the groundwork for all major organ systems in the process. Every stage appears highly ordered, preordained, and governed by a hidden script (Gee, 2004).

Aristotle observed development in dissected chicken embryos, and he proposed that this order was not imposed externally, but emerged internally: male semen, he claimed, carries not a miniature human but the *form*, the *organizing principle* that shapes and animates matter. Thus, the female provides the raw material, the *potentiality* that is sculpted and *actualized* by the male semen, the agent of form (Aristotle, 1943). Aristotle considered this form, the highest actualization of human potential, to be the soul.

To Aristotle, the soul is not a separate spirit from or supernatural addition to life, but the animating principle of life itself: the actualization of a body's potential, inherent and inseparable from its physical substance. In *De Anima*, he outlines three levels of soul—nutritive (growth and metabolism), sensitive (sensation and motion), and rational (thought and will)—not as compartments, but as sequential functions that emerge as the body becomes capable of sustaining them (Aristotle, 1981). The nutritive soul arises with metabolic activity, the sensitive soul with sensory and motor structures, and the rational soul with the formation of higher cognitive faculties. The soul does not precede or impose itself on the body; it emerges naturally as the body is structured to support it. This vision is grounded in Aristotle's broader hylomorphic theory: matter (*hylê*) contains the potential for form (*morphê*), but only when that matter is properly organized does the form become actual. Thus, to possess a soul is not to harbor a separate essence—it is to be a living body configured in such a way that life becomes possible.

“While modern biology has shed the metaphysical weight of *telos*, it still describes development in ways that presume order, sequence, and direction”

Aristotle also proposed that this actualization happens *teleologically*: that is, in accordance with a final natural cause. A seed becomes a tree because it is supposed to; it grows toward what it is naturally meant to be. Teleology is thus the internal, implicit orientation of development toward an end. In the embryo, this end is the mature human. While modern biology has

shed the metaphysical weight of *telos*, it still describes development in ways that presume order, sequence, and direction. We now call this teleonomy, an apparent purposiveness without final cause—but the conceptual echo of Aristotle's philosophy remains (Mayr, 1982).

Form, Not Fragments

Some aspects of Aristotle's biology were wrong. Perhaps most notably, he misunderstood the role of the female in reproduction. Applying temperamental and social ideas to biology, Aristotle asserted that the female matter was a 'passive', nurturing receptacle for the 'active' male form, and that female matter was composed of congealed menstrual blood (Aristotle, 1943). Evidently, his model was shaped by observational limitations and embedded gender biases. But despite these flaws, his critiques of alternative developmental theories were strikingly incisive and progressive.

The dominant theory of Aristotle's time was the theory of *pangenesis*, developed by thinkers such as Pythagoras, Anaxagoras, and Empedocles. Pangenesis



Figure 1. The stages of embryonic development as drawn by Aristotelian natural philosophers (Rueff, 1554).

supposed that semen extracts ‘essences’ from all parts of the human body, little pieces of information in the form of microscopic body parts contributed by both mother and father to form a child. Aristotle dismissed it with characteristically razor-sharp logic: if parents contributed bits of each part, why don’t children inherit their injuries? Why not two hearts, two livers? Why not hybrid monstrosities? Nor, he observed, does the embryo appear to develop by simple accumulation (Aristotle, 1943). Aristotle even offered an anecdote foreshadowing Mendelian inheritance: a Greek woman had married an Ethiopian man. Their daughter was not dark-skinned, but her grandson was, suggesting that some traits skip generations (Aristotle, 1943).

Crucially, Aristotle concluded that the embryo does not contain pre-formed parts but instead develops gradually through the realization of form in matter. This became the basis for epigenesis: the theory that an organism’s structure unfolds progressively over time, not all at once. Unlike pangenesis or preformationism, epigenesis accounts for emergence—the formation of

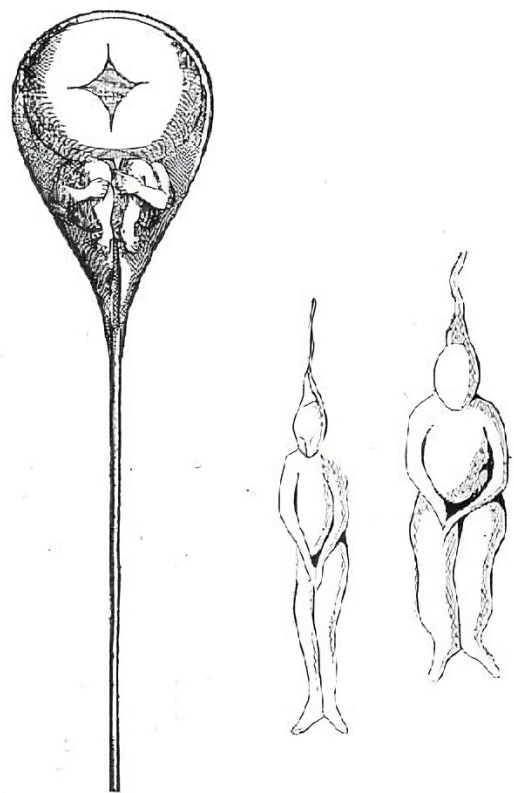


Figure 2. An illustration of a homunculus inside a sperm cell (Hartsoeker, 1695).

new structures stage-by-stage in an embryo rather than miniature replication (Gee, 2004).

Persistence of a Paradigm

Aristotle’s epigenetic framework of development remained dominant until the 17th century, when the rudimentary microscope revealed sperm and egg, and new theories emerged: spermism and ovism (Gee, 2004; Mayr, 1982). Spermists argued that the entire form of a human was located in the male seed—an image literalized in contemporary diagrams of *homunculi*, little preformed humans. Ovists, conversely, claimed that the fully-formed human was contained in the egg, awaiting activation.

Spermism and ovism were both subtheories of *preformationism*, a spiritual descendant of Pythagorean-Anaxagorean pangenesis. Here, the embryo was not shaped over time, but it was simply *unfolded*, already in its final form. Some imagined that all future generations were pre-formed and stacked inside the first woman, awaiting sequential birth. This preformationist idea of infinitesimal nesting dolls was not just poor biology; it was poor philosophy (Gee, 2004; Mayr, 1982). Unlike Aristotle’s model, grounded in potentiality, emergence, and structured becoming, preformationism rejected the developmental logic that gave Aristotelian epigenesis its enduring scientific and philosophical robustness.

It wasn’t until the mid-eighteenth century that Aristotle’s model saw a revival under a different name. The contrarian German scientist Caspar Friedrich Wolff rejected preformationism and returned to the idea that structure emerges throughout development, that the embryo begins without form and gains it step-by-step (Gee, 2004). This was, in effect, the revival of Aristotle’s epigenesis. The British physician William Harvey, more than a century earlier, had already articulated the principle of epigenesis, although he was often mistakenly labeled an ovist for his emphasis on the importance of the egg (Embryology 2020). In 1651, Harvey published his magnum opus, rejecting both the idea that development occurs by unfolding a miniature form and Aristotle’s claim that the embryo arises from coagulated menstrual blood (Harvey, 1651). Through meticulous dissections of pregnant deer, Harvey found no trace of menstrual blood, nor any physical residue of semen in the womb (Gee, 2004). Yet rather than reject the Aristotelian paradigm, Harvey refined it, retaining its language even as he deepened its implications. Additionally dissecting chickens in the spirit of Aristotle, Harvey remarked that the chick’s parts are “not fashioned simultaneously, but emerge in their due succession and order,” and more clearly that “its



Figure 3. Allegorical engraving of animals and humans emerging from an egg bearing the phrase *ex ovo omnia*—“everything from the egg.” (Harvey, 1651).

form proceeds simultaneously with its growth, and its growth with its form” (Harvey, 1651). Harvey famously declared, *ex ovo omnia*—that all animals come from the egg—not as preformation, but as emergence (Harvey, 1651). Harvey did not merely hint at epigenesis; he gave it empirical form.

But two hundred years after Harvey, even Charles Darwin, revolutionary as he was, resurrected a version of pangenesis when pressed for an explanation of heredity. Natural selection explained how traits were filtered over time, but where did those traits come from? Darwin proposed a theory of gemmules, microscopic particles shed from each organ that gathered in the reproductive system (Gee, 2004; Mayr, 1982). This was nothing more than a repackaging of the old, debunked theory of pangenesis, the very theory that Aristotle had refuted *twenty centuries* earlier.

Darwin’s failure to explain heredity contributed to widespread skepticism of his theory. It took the rediscovery of Gregor Mendel’s work and William Bateson’s insistence on variation to reframe the question

of how traits are generated, inherited, and materially encoded—that is, how structured form is preserved across generations without relying on speculative particles or preformed bodies (Gee, 2004). The discovery of DNA revealed that form is not carried in miniature but is encoded more abstractly in sequences, regulated over time, and activated only when certain thresholds are met. Morphogens and transcription factors now orchestrate development with striking precision. They are not Aristotle’s metaphysical forces, but they perform a similar role within a new conceptual framework: guiding the emergence of structure from potential, and revealing that even in molecular code, the logic of form and epigenesis still speaks.

From Form to Code

Epigenesis—development as the gradual actualization of potential—reflects Aristotle’s teleology. His notion of *form* is functionally a proto-genome: an immaterial principle that determines not just what a thing is, but

how it becomes what it is. For Aristotle, form governed the sequence and structure of development from within, just as DNA regulates cellular differentiation over time. Today, gene regulatory networks and morphogen gradients reveal what Aristotle intuited: that becoming is not random, but scripted by code (Gee, 2004).

At the smallest scale, using tools Aristotle could only dream of, we now see that even single cells exhibit a similar process of actualization. In embryogenesis, a single totipotent zygote divides into pluripotent stem cells, which retain the potential to become any tissue but do not function as specialized cells until gene regulatory networks orchestrate their differentiation. A stem cell has its developmental trajectory embedded within its internal genetic program, activated in a highly structured, sequential manner.

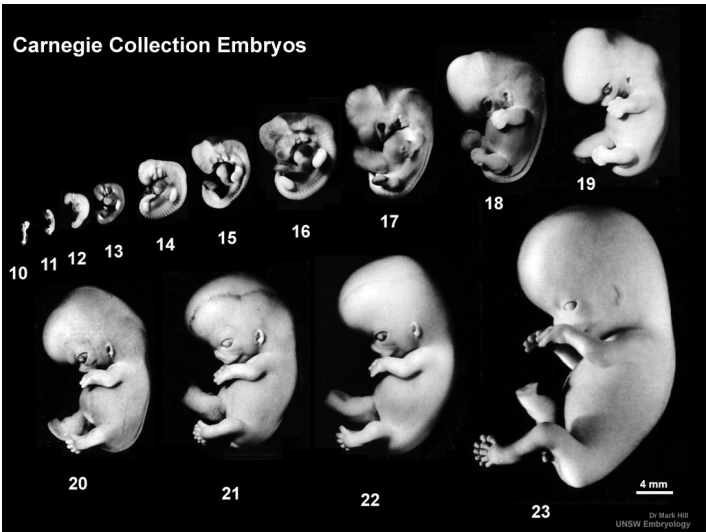


Figure 4. The stages of human embryonic development from zygote through early fetal form (Hill, n.d.).

This programmatic unfolding may also be viewed through the lens of Aristotle’s model of the soul. The vegetative soul corresponds to early cellular division and metabolic activity, as the embryo sustains and grows itself. The sensitive soul parallels the emergence of neuronal networks, as cells take on sensory and motor functions. Finally, the rational soul, which Aristotle considers uniquely human, finds a counterpart in the late-stage development of higher brain function, as the neocortex forms and cognition becomes possible. At each stage, new functions emerge only when the biological structures necessary to sustain them are properly developed—precisely the logic of Aristotle’shylomorphism applied to modern cellular differentiation.

Teleology After Telos

Aristotle could not have predicted genes, yet he understood that matter does not organize itself without a principle. That principle, for him, was the soul, the form of a living body that is actualized only when the material conditions for life are properly arranged.

Modern biology no longer uses the language of souls or final causes. It speaks of feedback loops, thresholds, and probability. Teleology, in the strict sense, is out. Teleonomy—the appearance of purposiveness—is what remains. This shift spares us the need to stretch metaphysics into biology; we no longer need to retrofit Aristotle into morphogen gradients. But something of his vision persists. In tracing the development of an embryo, we still ask how structure arises from undifferentiated matter, how complexity unfolds in time. The explanatory structure of potential realized through ordered steps remains intact. What Aristotle called form, we call code. What he saw as purpose, we see as biological function.

References

Aristotle. (1943). *Generation of animals* (A. L. Peck, Trans.). Harvard University Press. (Original work published ca. 350 BCE)

Aristotle. (1981). *On the soul* (H. G. Apostle, Trans.). Peripatetic Press. (Original work published ca. 350 BCE)

Gee, H. (2004). *Jacob’s ladder: The history of the human genome*. W. W. Norton & Company.

Harvey, W. (1651). *Exercitationes de generatione animalium*. London: William Dugard. Engraving reproduced from historical edition. Retrieved April 4, 2025, from <https://letraslibres.com/wp-content/uploads/2019/02/ovo%20omnia.jpg>

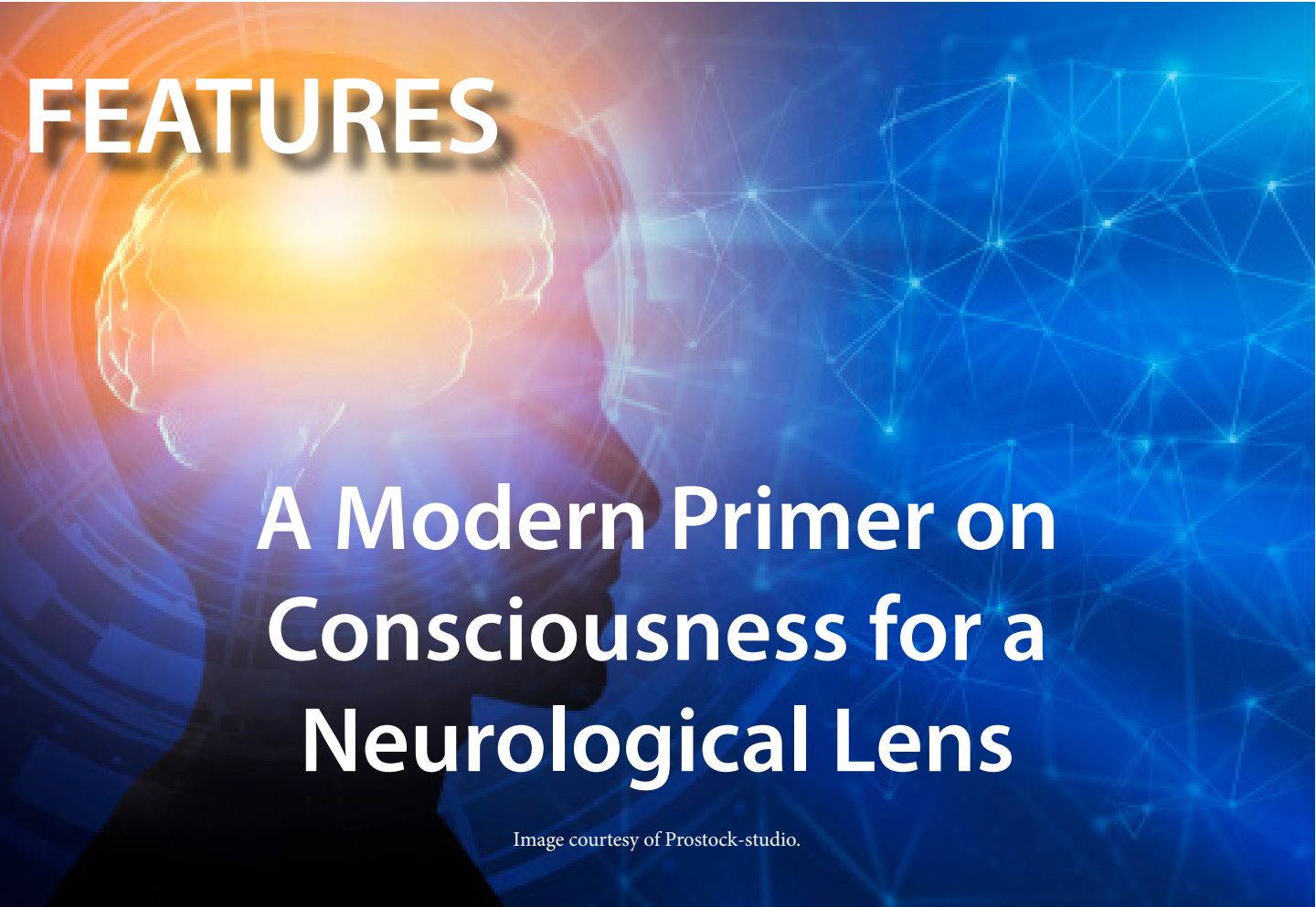
Hartsoeker, N. (1695). *Homunculus in sperm cell* [Illustration]. Wikimedia Commons. Retrieved April 4, 2025, from <https://upload.wikimedia.org/wikipedia/commons/9/94/Preformation.GIF>

Hill, M. A. (n.d.). *Human Carnegie stages 10–23* [Image]. UNSW Embryology. Retrieved April 4, 2025, from https://embryology.med.unsw.edu.au/embryology/images/thumb/3/37/Human_Carnegie_stage_10-23.jpg/800px-Human_Carnegie_stage_10-23.jpg

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Belknap Press of Harvard University Press.

Rueff, J. (1554). *De conceptu et generatione hominis* (W. Haller, Trans.). Zurich: Christopher Froschauer. Woodcuts reproduced in *Making visible embryos*. Department of History and Philosophy of Science, University of Cambridge. Retrieved April 4, 2025, from http://www.sites.hps.cam.ac.uk/visibleembryos/s1_3.html

Embryology. (2020). *Talk: Paper—William Harvey as an embryologist (1897)*. UNSW Embryology. Retrieved April 12, 2025, from [https://embryology.med.unsw.edu.au/embryology/index.php?title=Talk:Paper_-_William_Harvey_as_an_embryologist_\(1897\)](https://embryology.med.unsw.edu.au/embryology/index.php?title=Talk:Paper_-_William_Harvey_as_an_embryologist_(1897))



Avery Mizrahi ’28

Introduction

What does it mean to be human? Efforts to answer this question almost always end up mentioning one key term, popular and numerous in its definitions: consciousness. In day-to-day language, we might mention someone being knocked unconscious. Or we might say we were unconscious of our actions, such as driving home (in this case, interchangeable with unaware). For many philosophers of mind, the most succinct definition of consciousness is “what it is like” to be a human: unified, directed, and self-aware. It is the redness of an apple, or the painfulness of pain (Nagel, 1974).

Scientists attempt to answer the question of human meaning by formulating a biological basis of consciousness, which is paramount to understanding how human experience of life changes over time, occasionally for the worse.

How does our consciousness change with age and neurodegenerative disease (NDD), when we lose the memories or functions that make us, us? How about after traumatic brain injury (TBI), when personality erodes in the face of neuronal loss? Of great relevance is the question of how we can clarify, quantify, and restore consciousness for patients afflicted by ailments such as these. In this article, I will briefly summarize how historical and modern theories of consciousness have been conceptualized and quantified with the goal of better treating neurological disorders.

Historical Foundations of Consciousness

As early as the 5th century BCE, the Buddha’s foundational theory of consciousness described no consistent self, but rather the shifting impermanent instances that make up consciousness (Bodhi, 2000). Later, western philosophers tended to reject this view, and John Locke is credited with postulating consciousness as the consistent awareness of

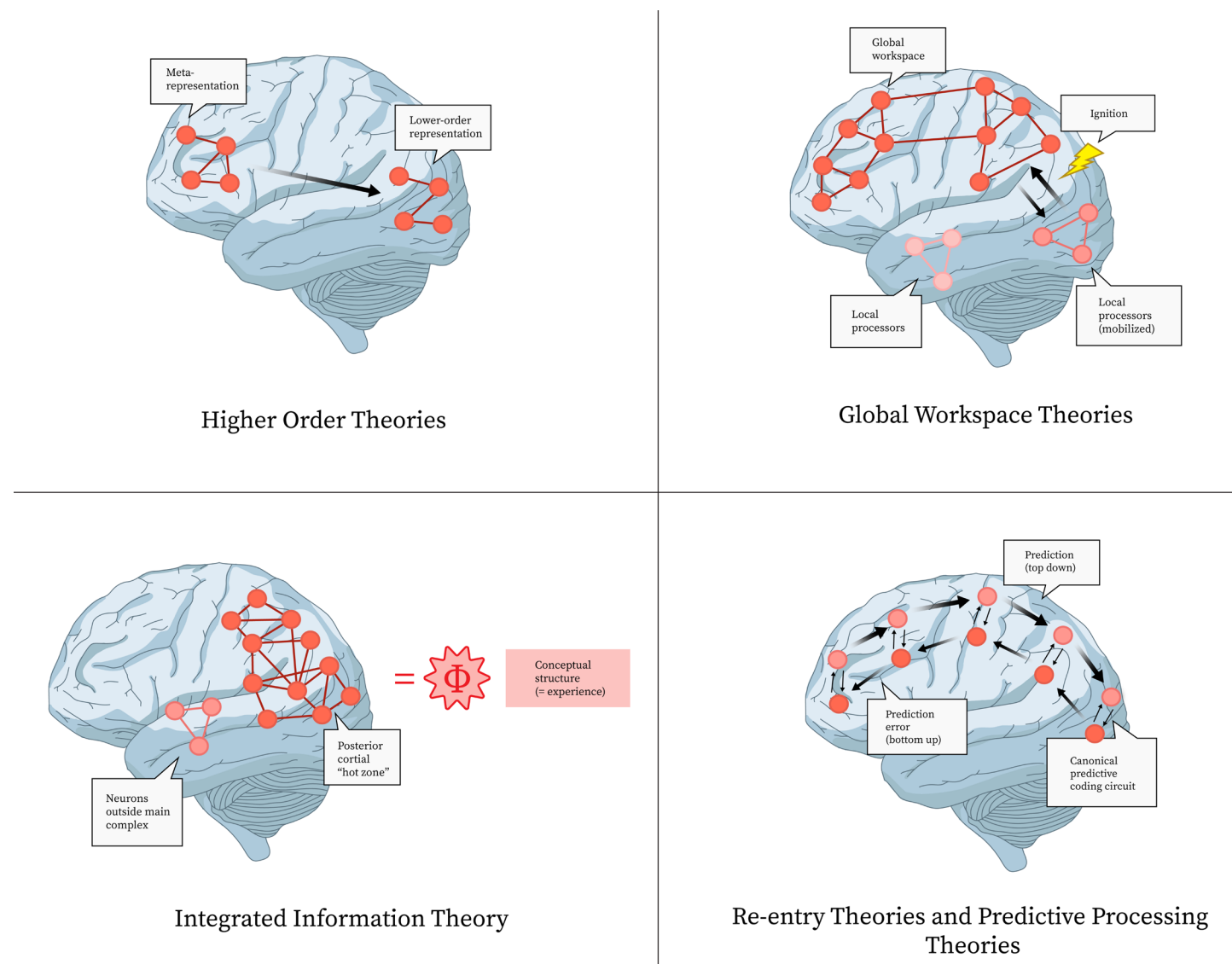


Figure 1. Anatomical networks involved in a few specific theories of consciousness (Adapted from Seth & Bayne, 2021).

thought: "I do say he can not think at any time, waking or sleeping, without being sensible of it" (Locke, 1689). However, as we now know—and as G.W. Leibniz posited as early as 1720—unconscious cognitive processes frequently bypass our conscious awareness; this is the example of being unaware of driving home, or sleeping. In 1787, Immanuel Kant modified this idea, arguing that consciousness is not the awareness or unawareness of thought, but rather the experience of a self situated in the world (Kant, 1787/1998). Kant inspired the Phenomenology movement of the early 20th century, which was characterized by studying the philosophy of lived experience (Smith, 2013). Modern theories elaborate on these philosophical notions in an attempt to create a biological hypothesis of how consciousness arises.

Modern Theories of Consciousness

The theories surrounding how the brain creates a subjective experience are numerous and contradictory. For example, the Attention Schema Theory (AST) argues the perception of self arises from the brain being inundated with information. Because of the sheer amount of information we take in, our attention is selective; therefore, consciousness is our awareness of selective attention (Graziano & Webb, 2015). Awareness of attention creates a sense of self, the "I" in the statement "I am having the conscious experience of this red apple." In the same way, our brain simplifies visual information and creates an internal representation of the outside world, AST argues our brain simplifies attention and creates an internal representation of ourselves. Therefore, AST may be beneficial in comparing how representations

The Entropic Brain Hypothesis

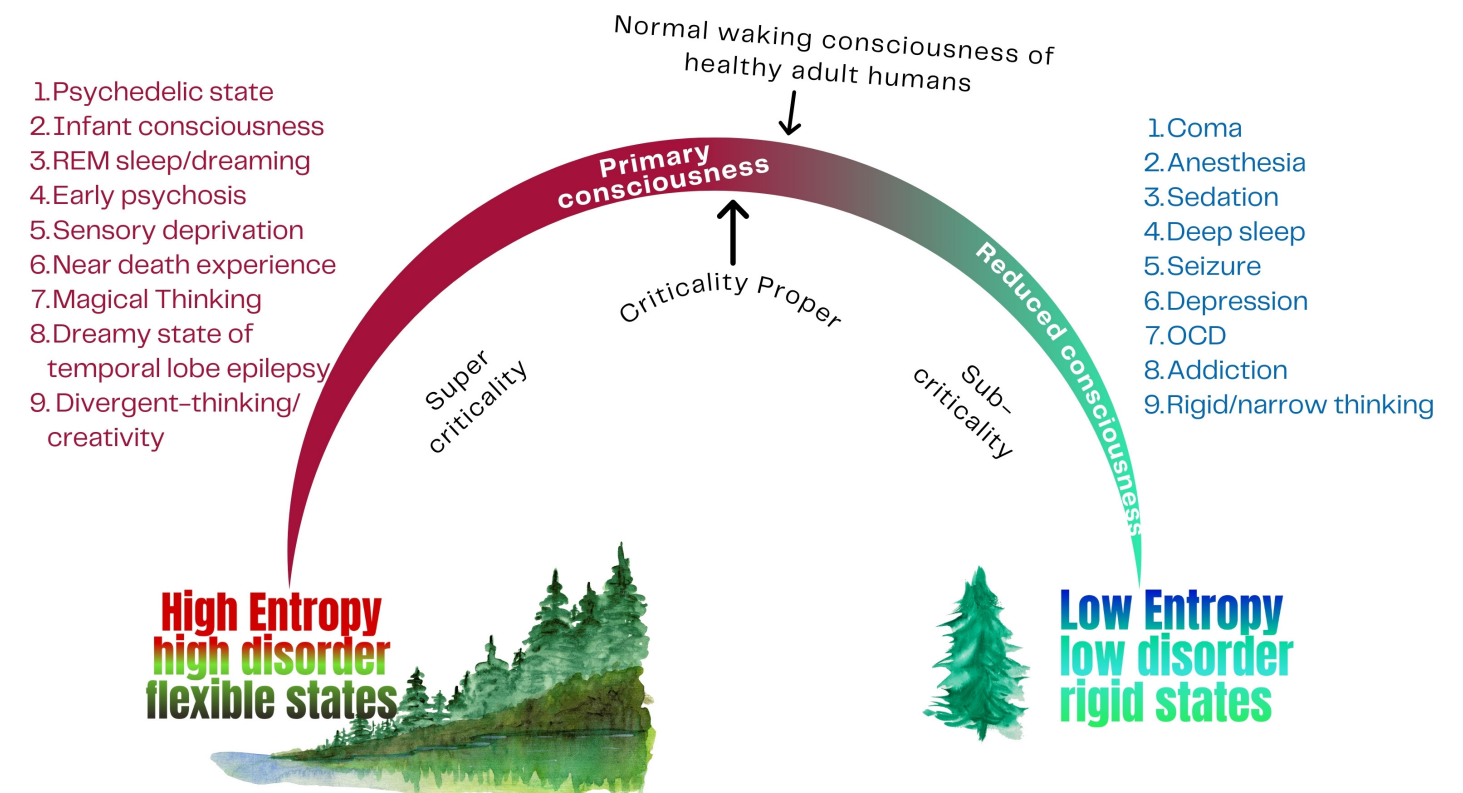


Figure 1. Various cognitive states and disorders may differ in entropic levels (Carhart-Harris, 2014).

of the self shift, from when memory deteriorates in Alzheimer's disease (AD) to when worldview changes in those suffering from Major Depressive Disorder.

A more mathematical approach to consciousness is the Integrated Information Theory (IIT), which proposes that one's degree of phenomenal experience can be translated into a single value (Seth & Bayne, 2021). This value is derived from the connectivity between each node (discrete brain regions) in a network where consciousness arises. IIT may also help identify the neural correlates of consciousness, or specific and minimal neural events that are associated with conscious experience (Dehaene et al., 2011). IIT is hotly debated because opponents state the brain is too complex to create one consistent value that adequately encompasses consciousness. The same researchers have created the more accessible Perturbational Complexity Index, which calculates consciousness values in a variety of states with reduced computational demand from electroencephalogram (EEG) data (Casali et al., 2013).

Another useful hypothesis that provides a cohesive framework for varieties of conscious experience is the Entropic Brain Hypothesis. This notion defines

a conscious state by its entropy, or the flexibility of neural patterns (Carhart-Harris, 2014). While examples of flexible states on the "high entropy" side of the spectrum are psychedelic experiences or rapid eye movement (REM) sleep, examples of "low entropy" states are comas or Obsessive-Compulsive Disorder.

Other popular theories tend to differ on whether consciousness arises from higher-order networks (Higher Order Theories), a global consciousness hub (Global Workspace Theory), or the brain's attempt to minimize prediction errors (Predictive Processing Theory) (Seth & Bayne, 2021). Though the above theories may influence how we approach the underlying biological mechanisms of consciousness, some argue that theories alone are unhelpful without additional tools of measurement. Therefore, theorists work in conjunction with experiments and technology to measure and quantify consciousness, especially in diseased states.

Measurement and Quantification

A meaningful method for measuring consciousness must reliably detect both consciousness and its

absence. One route taken is the Mirror Test for self-recognition (Dehaene, 2016). This intuitive test often consists of marking the forehead of an animal and seeing whether they wipe away the chalk when looking in the mirror. The theory proposes that if an animal can recognize itself in the mirror, there must be some sense of self-awareness. Interestingly, this behavior does not emerge in humans until 15-

“A meaningful method for measuring consciousness must reliably detect both consciousness and its absence”

24 months of age, and is linked to introspection, mental states, and empathy. At the opposite end of life, experts have observed that self-recognition deteriorates with cognitive decline (Biringer & Anderson, 1992).



Figure 3. A puppy likely doesn’t recognize itself in a mirror (iStock Emma Jocelyn).

As appealing as the recognition test’s binary answer to consciousness might be, it is commonly thought that consciousness is a gradient rather than a question of all-or-none. To understand this spectrum, many neuroscientists have utilized tools such as functional Magnetic Resonance Imaging (fMRI) to look at activity and connectivity in the brain. Most commonly, the default mode networks (DMN), discovered from task-free resting state fMRI, have

been identified as quantifiable features that vary in different states of consciousness (Raichle & Mintun, 2006). However, fMRI is not a perfect tool; it has high spatial resolution, but low temporal resolution (i.e., you can accurately view many regions of the brain, but not over a span of time). By comparison, electroencephalograms (EEG) have the opposite limitations, as a temporally, but not spatially, specific tool of measurement. Combined, these tools create an experimental paradigm to study consciousness both temporally and spatially.

Neurological Implications

Using the quantification and measurements of consciousness, scientists can observe how certain neural states change in the disordered brain, such as in NDDs. For example, early AD is correlated with decreased functional connectivity in the DMN during resting-state fMRI, suggesting an altered consciousness as the disease progresses (Simic et al., 2014).

Additional work surrounds how the brain recovers from coma to the normal function associated with wakefulness. Many potential neuronal activity states are associated with wakefulness, so neuroscientist Diany Calderon speculates that the brain may pass through different activity states as it recovers its normal function (Hudson et al., 2014). As the brain systematically passes through these distinct states of neuronal activity, consciousness may “increase” as the brain recovers, possibly gaining entropy (as postulated by the Entropic Brain Hypothesis). Therefore, understanding how comatose and TBI patients with reduced consciousness recover may provide insight into the levels of human consciousness affected by varied brain states and regions.

Brain Computer Interfaces (BCIs) may even be able to preserve minimally conscious states by following commands and executing external behaviors (Galiotta et al., 2023). Doctors can diagnose and treat disorders of consciousness and locked-in syndrome by utilizing BCIs, which are external devices that detect mental activities like motor imagery, spatial navigation, and arithmetic to allow patients to engage with the world. Advances such as these push us to question how much a physical body, which can be plagued by disease, is needed to interact with the world (Manisha et al., 2022).

“Doctors can diagnose and treat disorders of consciousness and locked-in syndrome by utilizing BCIs, which are external devices that detect mental activities like motor imagery, spatial navigation, and arithmetic to allow patients to engage with the world”

Conclusion

Characterizing consciousness is the fundamental basis of many important questions, whether it be the discontinuity of life support for coma patients, personality shift following brain injury, or even the consciousness of brain organoids and AI models. By theorizing and measuring consciousness, we can understand how treatment affects neurological disorders. Therapeutic pathways ranging from meditation to immune pathway harnessing have an effect on consciousness (Kraemer et al., 2022; Valiukas et al., 2022). Theories, spanning AST to IIT, and tools, like rsfMRI and EEG, may not only provide insight into the effectiveness of treatment, but possibly the identity of the patient themselves. Defining consciousness is no longer just a philosophical debate, but a necessary area of study to cure the ailments of the human brain.

References

Bodhi, Bhikkhu (Trans.). (2000). The connected discourses of the Buddha: A new translation of the Samyutta Nikāya. Wisdom Publications.

Biringer, F., & Anderson, J. R. (1992). Self-recognition in Alzheimer's disease: A mirror and video study. *Journal of Gerontology*, 47(6), P385–P388. <https://doi.org/10.1093/geronj/47.6.p385>

Carhart-Harris, R. L. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, 8, 20. <https://doi.org/10.3389/fnhum.2014.00020>

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198), 198ra105. <https://doi.org/10.1126/scitranslmed.3006294>

Dehaene, S. (2016). Consciousness and the brain: Deciphering how the brain codes our

thoughts. *Neuroscience & Biobehavioral Reviews*, 83, 229–242. <https://doi.org/10.1016/j.neubiorev.2016.01.002>

Dehaene, S., Charles, L., King, J. R., & Marti, S. (2011). Neural correlates of consciousness: Progress and problems. *Neuroscience & Biobehavioral Reviews*, 36(2), 775–792. <https://doi.org/10.1016/j.neubiorev.2011.07.001>

Eliasmith, C. (n.d.). Semantic Pointer Architecture [Slides]. Slides by Xueyang Yao. <https://ndey96.github.io/deep-learning-paper-club/slides/Semantic%20Pointer%20ArchiEcture.pdf>

Galiotta, V., Luo, L., Haugg, A., Kotchoubey, B., & Naci, L. (2023). EEG-based brain–computer interfaces for people with disorders of consciousness: Features and applications. A systematic review. *Frontiers in Human Neuroscience*, 17, 1117653. <https://doi.org/10.3389/fnhum.2023.1117653>

Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, 6, 500. <https://doi.org/10.3389/fpsyg.2015.00500>

Hudson, A. E., Calderon, D. P., Pfaff, D. W., & Proekt, A. (2014). Recovery of consciousness is mediated by a network of discrete metastable activity states. *Proceedings of the National Academy of Sciences of the United States of America*, 111(25), 9283–9288. <https://doi.org/10.1073/pnas.1408296111>

Kant, I. (1998). Critique of pure reason (P. Guyer & A. W. Wood, Trans.). *Cambridge University Press*. (Original work published 1787)

Kraemer, K. M., Jain, F. A., Mehta, D. H., & Fricchione, G. L. (2022). Meditative and mindfulness-focused interventions in neurology: Principles, science, and patient selection. *Seminars in Neurology*, 42(2), 123–135. <https://doi.org/10.1055/s-0042-1742287>

Locke, J. (1689). An essay concerning human understanding. Thomas Basset.

Manisha, M., Raza, H., Cecotti, H., Prasad, G., & Majid, A. (2022). EEG-based brain–computer interfaces for communication and rehabilitation of people with motor impairment: A novel approach of the 21st century. *Frontiers in Neuroscience*, 16, 959339. <https://doi.org/10.3389/fnins.2022.959339>

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.

Richle, M. E., & Mintun, M. A. (2006). Brain work and brain imaging. *Annual Review of Neuroscience*, 29, 449–476. <https://doi.org/10.1146/annurev.neuro.29.051605.112819>

Seth, A. K., & Bayne, T. (2021). Theories of consciousness. *Nature Reviews Neuroscience*, 22(8), 514–526. <https://doi.org/10.1038/s41583-022-00587-4>

Simic, G., Babic, M., Borovecki, F., & Hof, P. R. (2014). Early failure of the default-mode network and the pathogenesis of Alzheimer’s disease. *CNS Neuroscience & Therapeutics*, 20(7), 692–698. <https://doi.org/10.1111/cns.12260>

Smith, D. W. (2013). Phenomenology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition). Stanford University. <https://plato.stanford.edu/entries/phenomenology/>

Tye, M. (2021). Qualia. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). Stanford University. <https://plato.stanford.edu/entries/qualia/>

Valiukas, Z., Ephraim, R., Tangalakis, K., Davidson, M., Apostolopoulos, V., & Feehan, J. (2022). Immunotherapies for Alzheimer’s disease—A review. *Vaccines*, 10(9), 1527. <https://doi.org/10.3390/vaccines10091527>

Van Gulick, R. (2023). Consciousness. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition). Stanford University. <https://plato.stanford.edu/entries/consciousness/>

