

UCLA

UCLA Electronic Theses and Dissertations

Title

Harnessing Network Data to Address Scientific Challenges

Permalink

<https://escholarship.org/uc/item/1jm9x4zw>

Author

Dewey, George

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Harnessing Network Data to
Address Scientific Challenges

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Epidemiology

by

George Dewey

2024

© Copyright by

George Dewey

2024

ABSTRACT OF THE DISSERTATION

Harnessing Network Data to
Address Scientific Challenges

by

George Dewey

Doctor of Philosophy in Epidemiology

University of California, Los Angeles, 2024

Professor Akihiro Nishi, Chair

Complex systems in science can be represented using networks and can be better understood by studying network-based principles that bridge different scientific disciplines. This dissertation showcases three studies that exemplify the use of network data to answer distinct scientific questions. Chapter 1 introduces network concepts and briefly discusses the scientific background and rationale for each of the three studies that make up the dissertation. Chapter 2 evaluates the differences between cooperation, defection, and punishment decisions in two series of online network games using decision times and assessed if experimental time pressure could shift the distribution of decision types away from punishment and toward cooperation; results show that punishment was slower than either cooperation or defection and that experimental time pressure did not reduce the frequency of punishment comparing games with and without time pressure.

Chapter 3 uses an agent-based framework to simulate a network intervention to increase the lung cancer screening rate in the United States and tests two modifications to the framework with the potential to further increase the screening rate (allowing agents to adopt the intervention prior to observation and expanding the network size); results show that the intervention increases the screening rate among eligible individuals to 51% after 10 years of simulation (compared to the current screening rate among eligible individuals of 5.8%) and that increasing the network size boosts the screening rate above the rate achieved by the intervention alone. Chapter 4 describes the use of data from a citation network of scientific articles from epidemiology and public health to evaluate potential biases in ChatGPT's ability to rate the validity of a scientific citation; results show that citation-reference pairs that included cited articles from English-speaking or high-income countries were rated higher than pairs with cited articles from non-English-speaking or low-income countries. Chapter 5 summarizes the results of the three studies and describes implications of those results for future research. Overall, this dissertation demonstrates the use of network science principles and network data to explore solutions to three challenges from behavioral science, intervention planning, and artificial intelligence.

The dissertation of George Dewey is approved.

Timothy F. Brewer

Kai-Wei Chang

Catherine Ann Sugar

Akihiro Nishi, Committee Chair

University of California, Los Angeles

2024

Dedication

I dedicate this dissertation to my friends and family who have supported me throughout my academic journey. Thank you for inspiring me to explore and seek knowledge – both now and into the far future.

Table of Contents

Acknowledgements.....	x
Curriculum Vitae.....	xii
Chapter 1: Introduction.....	1
1.1 Introduction to Networks.....	1
1.2 Cooperation and Punishment in Networks.....	3
1.3 Network Interventions.....	4
1.4 Networks of Scientific Citations.....	5
1.5 Dissertation Goals and Rationale.....	7
1.6 References.....	9
Chapter 2: Punishment is Slower than Cooperation or Defection in Online Network Games.....	14
2.1 Abstract.....	14
2.2 Introduction.....	14
2.3 Methods Overview.....	16
2.4 Experiment 1.....	21
2.5 Experiment 2.....	24
2.6 General Discussion.....	26
2.7 Figures and Tables.....	28
2.8 Supplementary Information.....	32
2.9 References.....	39
Chapter 3: The Role of Network Size in Intervention Planning for Increasing the Lung Cancer Screening Rate: An Agent-based Modeling Study.....	44
3.1 Abstract.....	44
3.2 Introduction.....	44
3.3 Methods.....	47
3.4 Results.....	55
3.5 Discussion.....	57
3.6 Figures and Tables.....	62
3.7 Supplementary Information.....	66
3.8 References.....	69
Chapter 4: ChatGPT Reinforces Geographic and Language Biases in a Citation Network of Epidemiologic Literature.....	73
4.1 Abstract.....	73
4.2 Introduction.....	74

4.3 Methods	77
4.4 Results	81
4.5 Discussion	83
4.6 Figures and Tables.....	86
4.7 Supplementary Information.....	91
4.8 References	92
5. Concluding Remarks.....	96

List of Figures

Figure 2.1. Example player’s screenshot without (A) and with (B) time pressure.....	28
Figure 2.2. Punishment is rare and slower than cooperation or defection in Experiment 1.	29
Figure 2.3. Punishment is slower than cooperation or defection, even under time pressure.....	30
Figure 2.4. Time pressure does not change the frequency of mechanism-specific punishment. ..	31
Supplementary Figure S2.1. Example message shown to dropped players in the time pressure condition.	32
Supplementary Figure S2.2. Distributions of decision times for cooperation, defection, and punishment in Experiment 1.	33
Figure 3.1. The observational learning setting (blue) successfully encourages diffusion of LDCT screening behavior compared to the control setting (gold).....	63
Figure 3.2. Seeding the network with agents who have previously been screened does not result in increased diffusion of the intervention.	64
Figure 3.3. Increasing the network size results in increased diffusion of the intervention, but with diminishing returns.	65
Supplementary Figure S3.1. Results of robustness tests for the observational learning setting (simulation setting 2).	68
Figure 4.1. Flow diagram for obtaining the final dataset of citation-reference pairs used for analysis (n = 401,737).....	86
Figure 4.2. ChatGPT-derived validity scores increase when citation-reference pairs include citations or references from English-speaking, high income countries.	87

List of Tables

Supplementary Table S2.1. Multilevel random intercepts models for decision times for all 3 experimental settings.	34
Supplementary Table S2.2: Multilevel logistic random intercepts model for the effect of time pressure on the odds of punishment, cooperation, and defection in Experiment 2 (results for Fig. 2.3A).	35
Supplementary Table S2.3: Multilevel logistic random intercepts model for the effect of time pressure on the odds of individual punishment mechanisms in Experiment 2 (results for Fig. 2.4).	36
Supplementary Table S2.4. Multilevel random intercepts models for decision times to compare individual punishment mechanisms.	37
Supplementary Table S2.5: Multilevel random intercepts models for the effect of having been punished in the prior round on punishment decision times in Experiment 1.....	38
Table 3.1. The 13 parameters manipulated in the agent-based simulations.....	62
Supplementary Table S3.1. Calibration results for the control setting simulations.	67
Table 4.1. Examples of citation-reference pairs and their validity scores generated by GPT.	88
Table 4.2. Characteristics of the citing papers and cited papers in our sample citation network of epidemiologic articles.	89
Table 4.3. Random effects linear regression results for the associations between cited paper characteristics and the ChatGPT-generated validity scores.	90
Supplementary Table S4.1. Random effects linear regression results for the associations between cited paper characteristics and the ChatGPT-generated validity scores including network characteristics.....	91

Acknowledgements

I would like to express my gratitude to my advisor and committee chair, Akihiro Nishi, for all his support and mentorship during my time as a doctoral student at UCLA. I greatly appreciate the time and effort Dr. Nishi has spent working with me to develop my academic, technical, and personal skills and to establish myself as an independent researcher. I will be grateful for all the doors he opened for me for many years to come and I hope to maintain our strong relationship for the rest of my career.

Thank you to the members of my doctoral committee – Timothy Brewer, Catherine Sugar, and Kai-Wei Chang – for providing outstanding research support, instruction, and career advice during my time working on this dissertation. Your guidance and scientific expertise are much appreciated.

Chapter 2 of this dissertation is an adapted version of:

Dewey G, Ando H, Ikesu R, Brewer TF, Goto R, Nishi A. (2024). Punishment is Slower than Cooperation or Defection in Online Network Games. Under review.

I would like to acknowledge the contributions of my co-authors as follows: GD and AN designed the project. GD and HA conducted experiments. HA provided programming support. RI conducted preliminary data analysis and assisted in early manuscript preparation. GD, HA, RI, TB, RG, and AN analyzed the findings and wrote the manuscript.

I am also grateful to the members of my cohort in the UCLA Department of Epidemiology – Pat Arena, Angie Barrall, and Alex Moran – whose enduring friendship helped me through tough times and inspired me to strive for greatness.

I would also like to thank the Department of Epidemiology and its staff members, especially Joy Miller, who have assisted me and provided valuable guidance throughout my time at UCLA as a student in both the master's and doctoral programs. I also acknowledge financial support from the Department of Epidemiology via the Dean's Leadership Grant, Department of Epidemiology Graduate Fellowship, and HEALRISE Scholarship as well as from the UCLA Fielding School of Public Health High Impact Data Initiative.

Lastly, I am eternally grateful for the lasting encouragement from my parents, Peter and Rose, who have provided me with the opportunity to succeed through their great sacrifices. Thank you for supporting me my entire life – my accomplishments would not have been possible without your ongoing love and support.

Curriculum Vitae

2009-2013	B.A., Biological Sciences, University of Chicago
2015-2017	M.P.H., Epidemiology, UCLA
2015-2017	UCLA Department of Epidemiology Dean's Leadership Grant
2016-2017	Epidemiology Intern, Los Angeles County Department of Public Health, Division of Chronic Disease and Injury Prevention
2018	Los Angeles County Department of Public Health SAS Users Individual Award
2019	Los Angeles County Department of Public Health SAS Users Group Award
2019-2021	UCLA Department of Epidemiology Graduate Fellowship
2019-Present	Graduate Student Researcher, UCLA Department of Epidemiology
2021	UCLA Fielding School of Public Health High Impact Data Initiative Award
2022	UCLA Department of Epidemiology HEALRISE Scholarship

PUBLICATIONS

1. **Dewey G**, Ando H, Ikesu R, Brewer TF, Goto R, Nishi A. (2024). Punishment is Slower than Cooperation or Defection in Online Network Games. Under review.
2. **Dewey G**, et al. (2024). The Role of Network Size in Intervention Planning for Increasing the Lung Cancer Screening Rate: An Agent-based Modeling Study. In preparation.
3. **Dewey G**, et al. (2024). ChatGPT Reinforces Geographic and Language Biases in a Citation Network of Epidemiologic Literature. In preparation.
4. **Dewey G**, Ando H, Goto R, Lu M, Miura A, Nishi A. (2023). Hits and misses of "How do you feel right now?": A validation study using the Positive and Negative Affect Schedule (PANAS). In preparation.

5. Dai J, Nishi A, Tran N, Yamamoto Y, **Dewey G**, Ugai T, Ogino S. (2021). Revisiting Social MPE: An Integration of Molecular Pathological Epidemiology and Social Science in the New Era of Precision Medicine. *Expert Review of Molecular Diagnostics*.
6. Nishi A, **Dewey G**, Endo A, Neman S, Iwamoto SK, Ni MY, Tsugawa Y, Iosifidis G, Smith JD, Young SD. (2020). Network Interventions for Managing the COVID-19 Pandemic and Sustaining Economy. *Proceedings of the National Academy of Science (PNAS), U S A*.
7. **Dewey G**, Wickramasekaran R, Kuo T, Robles B. (2017). Associations between sodium knowledge and health behaviors: results of an internet panel survey in Los Angeles County. *Preventing Chronic Disease*.
8. Wickramasekaran R, Robles B, **Dewey G**, Kuo T. (2017). Evaluating the potential health and revenue impact of a 100% healthy vending machine nutrition policy at a large agency in Los Angeles County, 2013 – 2015. *Journal of Public Health Management and Practice*.

Chapter 1: Introduction

1.1 Introduction to Networks

Networks underpin many of the complex systems that scientists seek to understand (1-3). For example, the spread of the SARS-CoV-2 virus around the world that started the COVID-19 pandemic involved a diverse ensemble of systems that can be modeled as a single network or as a combination of multiple networks. These include the systems and methods governing transmission of the infection (4-6) in addition to those describing the mobility of healthy and infected individuals and the effects of non-pharmaceutical interventions on mobility (7-9). Researchers have also used networks to analyze the spread of both truthful and misleading information about COVID-19 vaccines and treatments through social media platforms (10, 11). These systems (transmission networks, mobility networks, and information networks) are composed of different elements and have entire disciplines built around understanding their characteristics, issues, and the data they generate. However, using network science, scientists can use a set of global principles that describe how individual components of systems behave and interact with each other to evaluate each of these network types using the same fundamental rules. These principles allow scientists to condense information that can describe entire complex systems into network-based data structures.

Fundamentally, network data describes both the characteristics of individual components of a network and the information that quantifies the relationships between those components. In network data, individual components are referred to as *nodes*: nodes represent actors within the network and each node can be characterized by node-specific attributes. Relationships between nodes are called *ties* or *edges*: ties can either quantify a relationship between nodes or be used to describe the characteristics of the relationship. For example, *social networks* are a type of network

that will be familiar to most readers, describing the connections between individuals that engage with each other. In social networks, nodes represent individual humans; characteristics of these nodes could include network members' names, their political affiliation, or their current health status. Ties in social networks could therefore indicate the strength of the friendship or physical distance between two individuals in the network.

In addition to describing individual components of networks such as node-level and tie-level data, network datasets can also include summary measures that describe characteristics of the network in aggregate. The most common of these network metrics are measures of *centrality*, which serve to identify important nodes in the network, such as influential individuals in a business, super-spreaders in a transmission network, or key players in a network game. The simplest form of centrality is the *degree* (and by extension, the *degree distribution* of the network). The degree of a node is the number of ties it has to other nodes – nodes with higher degree have greater centrality since they are connected to more of the network at large. Consequently, the degree distribution is the probability distribution of individual node degrees over the entire network. Another common centrality measure is the *betweenness centrality*, which measures the probability of each node to be on the shortest path between two other nodes; nodes that are on many such paths have higher betweenness and are therefore more central in the network.

In summary, networks allow us to better understand systems which are difficult to understand if analysis only includes information about the individual components of the system. Information from networks can be condensed into network data, which includes information about the individual nodes of the network, relationship data that quantifies node connections, as well as measurements of the relationships between individual nodes or ties and the other nodes or ties in the network. The following three sections describe three examples of scientific areas where the

blend of individual- and group-level characteristics encapsulated by network data provide valuable insights: studying decision-making in dynamic experiments using networks, network intervention planning, and the science of science.

1.2 Cooperation and Punishment in Networks

Cooperation in humans exceeds what would be expected from an evolutionary perspective (12-15). People often choose to behave altruistically, benefiting complete strangers while gaining little or nothing for themselves (16, 17); this altruistic cooperation is often considered to be a defining characteristic of humanity (18). Cooperation is frequently contrasted with punishment: behavior that involves paying a cost to harm others. Prevailing theories about the rationale behind punishment decisions generally link such choices to their influence on future cooperation. One theory suggests that punishment is part of a cycle that maintains cooperation in human societies by reducing the fitness of free-riders to convert them into future cooperators (19, 20). An alternative theory suggests that punishment is an equalizing mechanism which allows people with less wealth to take resources away from those with more wealth to reduce inequality (21-24).

To investigate these theories, researchers from a range of fields including economics, psychology, and evolutionary biology, have turned to laboratory-based experiments such as the public goods game (PGG) or prisoner's dilemma (PD) game to evaluate why and how people choose to cooperate or defect (choosing not to affect others). One aspect of these games that has proven fruitful in the investigation of cooperation has been researchers' ability to directly measure the decision times associated with each type of choice. In these experimental games, decision times reflect a combination of the experimental parameters as well as the conditions of the individual participants (25). Under conditions in which participants are allowed to make decisions without

external time pressure, slower decisions reflect the occurrence of decision conflict (26-28), in which participants' individual preferences do not match the conditions in which decisions are made.

The importance of understanding the environmental conditions of cooperation or defection decisions make experiments using network-based approaches an obvious fit for this field of study. By arranging experimental participants in network structures, researchers can measure the speed of decisions while simultaneously evaluating the environment in which those decisions were made by recording both the decision speed of a focal participant and the decision speeds of participants connected to the focal participant in the experimental network. For example, research has shown that the frequency and speed at which people choose to cooperate or defect is influenced by the social environment of the actor (29) as well as by cultural norms which change how likely people are to be cooperative with strangers (30). Adaptation of the network-based experimental frameworks that were used in these studies could add punishment to the existing cooperation and defection options to evaluate the relationships between these three behaviors, especially in the context of decision times and decision conflict.

1.3 Network Interventions

Valente's framework of network interventions (31) has been a vital asset to community leaders and public health professionals who develop interventions to promote behavior change. Network interventions rely on network features to succeed: for example, the most basic and most common type of network intervention approach (the *individuals* approach) asks intervention planners to identify important individuals within a network and use them as opinion leaders. Successful individual-approach interventions must therefore identify central individuals in their

respective networks who are connected to not only other well-connected individuals but also less-connected pockets who may benefit more from the intervention. Such interventions have been shown to be effective in promoting a variety of preventive health behaviors including drug use cessation (32), HIV prevention (33-35), and cancer screening (36, 37).

An alternative to the individuals approach is the *induction* approach, which encourages network members to interact with each other to promote diffusion of the intervention behavior or product. Induction-approach interventions are like word-of-mouth marketing campaigns (38) in that they begin with a few people supporting the intervention change who spread the behavior or product to others in their network, which is followed by repetition of this process and eventual propagation of the intervention throughout the combined network of the participants. Like individual-approach interventions, induction-based network interventions have also been found to be effective in spreading a variety of preventive health behaviors (39).

The role of network data in planning network interventions is clear. However, solutions to challenges to planning network interventions, such as requiring identification of peer leaders or calculating the needed network size or network architecture to ensure the success of a network intervention are still unclear. Effective use of network data from networks of different constructions, especially through simulation and evaluation of networks in the planning stages of interventions, should provide vital insights into the development of future network interventions.

1.4 Networks of Scientific Citations

Scientific citations are central to the scientific process: researchers give credit to other researchers who inspired their work or those who previously detailed the scientific principles that support the researchers' current work (40). The scientific study of citations and citation behavior

(known as *scientometrics*) has provided insights into the psychology (41, 42) and motives (43, 44) behind citations and has been used to produce many of the citation metrics such as impact factor (45) and h-index (46) that label the profiles of academic researchers today. With the advent of the Internet and platforms such as Web of Science and PubMed, researchers have been able to construct expansive network studies that detail the citation history and patterns of citation for many scientific domains (47-49). In scientific citation networks, nodes represent individual authors, academic institutions, or scientific articles, with connections to other authors (networks where nodes represent authors or institutions are sometimes referred to as *collaboration networks*) or articles representing individual citations. Analysis of these citation networks have improved the scientific community's understanding of biases in academic publishing, hiring, and promotion (50-53) and can be a useful tool for scientists to “self-regulate” any unconscious biases in their own citation behavior.

Recent developments in the field of language modeling and the publicization of large language model (LLM) platforms such as OpenAI's ChatGPT have led researchers to strongly consider the use of these models in the research process. Of particular interest is the use of language models to serve as “research assistants” which can compare information described in one scientific article to information described in others and verify if references connecting multiple articles are accurate. However, the output generation process of models like ChatGPT has been shown to be biased (54-56); these biases are generally considered to arise from the characteristics of the data used to train the models (primarily English-language text from Western nations) (57).

With these issues in mind, an evaluation of ChatGPT using a citation network labeled with paper characteristics that ChatGPT could be biased against (such as the language or authorship or country of origin of individual papers) should be fruitful. By comparing ChatGPT's performance

on a research task using different elements from a complex network dataset, future studies could improve our understanding of ChatGPT's biases and increase our awareness of potential downfalls of using the model and similar tools in the research process.

1.5 Dissertation Goals and Rationale

This dissertation highlights three studies (which each correspond to one chapter of the dissertation) which aim to answer the following scientific questions from the three domains introduced previously:

- How do punishment and cooperation decisions differ in dynamic online network games, especially in the context of decision times? (Chapter 2)
- How do variations in network parameters affect the outcomes of a planned network intervention in agent-based simulations? (Chapter 3)
- How does bias influence ChatGPT's performance on a research task that evaluates data from a network of citations? (Chapter 4)

Each study in this dissertation was based on a network dataset that was curated from a data source that was constructed to identify solutions to these questions and to address testable hypotheses that provide supporting evidence for the solutions. Chapter 2's data is derived from two series of online experiments hosted on Amazon Mechanical Turk that were played by players from around the world. Meanwhile, Chapter 3 details network simulations of an intervention to improve the lung cancer screening rate in the United States and explains the effects of modifying the network size and network structure on the intervention's outcomes. Finally, Chapter 4 utilizes

a citation network of scientific articles from public health and epidemiology to evaluate potential biases in ChatGPT's ability to determine the validity of citation-reference pairs.

1.6 References

1. Strogatz SH. Exploring complex networks. *Nature*. 2001;410(6825):268-76.
2. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network Motifs: Simple Building Blocks of Complex Networks. *Science*. 2002;298(5594):824-7.
3. Barabási AL, Posfai MÁ. *Network Science*: Cambridge University Press; 2016.
4. World Health Organization. Coronavirus disease (COVID-19): How is it transmitted? 2021 [Available from: <https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-how-is-it-transmitted>].
5. Hâncean M-G, Perc M, Lerner J. Early spread of COVID-19 in Romania: imported cases from Italy and human-to-human transmission networks. *Royal Society Open Science*. 2020;7(7):200780.
6. Karaivanov A. A social network model of COVID-19. *PLOS ONE*. 2020;15(10):e0240878.
7. Galeazzi A, Cinelli M, Bonaccorsi G, Pierri F, Schmidt AL, Scala A, et al. Human mobility in response to COVID-19 in France, Italy and UK. *Scientific reports*. 2021;11(1):13141.
8. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*. 2021;589(7840):82-7.
9. Linka K, Peirlinck M, Sahli Costabal F, Kuhl E. Outbreak dynamics of COVID-19 in Europe and the effect of travel restrictions. *Computer methods in biomechanics and biomedical engineering*. 2020;23(11):710-7.
10. Cinelli M, Quattrocioni W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. *Scientific reports*. 2020;10(1):1-10.
11. Muric G, Wu Y, Ferrara E. COVID-19 vaccine hesitancy on social media: building a public Twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR public health and surveillance*. 2021;7(11):e30642.
12. Boyd R, Richerson PJ. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2009;364(1533):3281-8.
13. Axelrod R, Hamilton WD. The evolution of cooperation. *Science*. 1981;211(4489):1390-6.
14. Melis AP, Semmann D. How is human cooperation different? *Philos Trans R Soc Lond B Biol Sci*. 2010;365(1553):2663-74.

15. Rand DG, Nowak MA. Human cooperation. *Trends in Cognitive Sciences*. 2013;17(8):413-25.
16. Gintis H, Bowles S, Boyd R, Fehr E. Explaining altruistic behavior in humans. *Evolution and Human Behavior*. 2003;24(3):153-72.
17. Fehr E, Fischbacher U. The nature of human altruism. *Nature*. 2003;425(6960):785-91.
18. Grueter CC, Ingram JA, Lewisson JW, Bradford OR, Taba M, Coetzee RE, et al. Human altruistic tendencies vary with both the costliness of selfless acts and socioeconomic status. *PeerJ*. 2016;4:e2610.
19. Fowler JH. Altruistic punishment and the origin of cooperation. *Proc Natl Acad Sci U S A*. 2005;102(19):7047-9.
20. Fehr E, Gächter S. Altruistic punishment in humans. *Nature*. 2002;415(6868):137-40.
21. Bone JE, Raihani NJ. Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior*. 2015;36(4):323-30.
22. Raihani NJ, McAuliffe K. Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biol Lett*. 2012;8(5):802-4.
23. Raihani NJ, Thornton A, Bshary R. Punishment and cooperation in nature. *Trends Ecol Evol*. 2012;27(5):288-95.
24. Raihani NJ, Bshary R. Punishment: one tool, many uses. *Evolutionary Human Sciences*. 2019;1.
25. Evans AM, Rand DG. Cooperation and decision time. *Curr Opin Psychol*. 2019;26:67-71.
26. Castro Santa J, Exadaktylos F, Soto-Faraco S. Beliefs about others' intentions determine whether cooperation is the faster choice. *Scientific Reports*. 2018;8(1):7509.
27. Diederich A. Decision making under conflict: Decision time as a measure of conflict strength. *Psychonomic bulletin & review*. 2003;10(1):167-76.
28. Evans AM, Dillon KD, Rand DG. Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas. *J Exp Psychol Gen*. 2015;144(5):951-66.
29. Nishi A, Christakis NA, Evans AM, O'Malley AJ, Rand DG. Social Environment Shapes the Speed of Cooperation. *Sci Rep*. 2016;6:29622.
30. Nishi A, Christakis NA, Rand DG. Cooperation, decision time, and culture: Online experiments with American and Indian participants. *PLOS ONE*. 2017;12(2):e0171252.

31. Valente TW. Network interventions. *Science*. 2012;337(6090):49-53.
32. Booth RE, Lehman WE, Latkin CA, Brewster JT, Sinitsyna L, Dvoryak S. Use of a peer leader intervention model to reduce needle-related risk behaviors among drug injectors in Ukraine. *Journal of Drug Issues*. 2009;39(3):607-25.
33. Latkin CA, Donnell D, Metzger D, Sherman S, Aramrattna A, Davis-Vogel A, et al. The efficacy of a network intervention to reduce HIV risk behaviors among drug users and risk partners in Chiang Mai, Thailand and Philadelphia, USA. *Soc Sci Med*. 2009;68(4):740-8.
34. Jaganath D, Gill HK, Cohen AC, Young SD. Harnessing Online Peer Education (HOPE): Integrating C-POL and social media to train peer leaders in HIV prevention. *AIDS Care*. 2012;24(5):593-600.
35. Latkin CA. Outreach in natural settings: the use of peer leaders for HIV prevention among injecting drug users' networks. *Public Health Rep*. 1998;113 Suppl 1(Suppl 1):151-9.
36. Maxwell AE, Jo AM, Crespi CM, Sudan M, Bastani R. Peer navigation improves diagnostic follow-up after breast cancer screening among Korean American women: results of a randomized trial. *Cancer Causes & Control*. 2010;21:1931-40.
37. Allen JD, Stoddard AM, Mays J, Sorensen G. Promoting breast and cervical cancer screening at the workplace: results from the Woman to Woman Study. *American Journal of Public Health*. 2001;91(4):584.
38. Buttle FA. Word of mouth: understanding and managing referral marketing. *Journal of strategic marketing*. 1998;6(3):241-54.
39. Hunter RF, de la Haye K, Murray JM, Badham J, Valente TW, Clarke M, et al. Social network interventions for health behaviours and outcomes: A systematic review and meta-analysis. *PLoS Med*. 2019;16(9):e1002890.
40. Nicolaisen J. Citation analysis. *Annual review of information science and technology*. 2007;41(1):609-41.
41. Haslam N, Ban L, Kaufmann L, Loughnan S, Peters K, Whelan J, et al. What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics*. 2008;76(1):169-85.
42. Bavelas JB. The social psychology of citations. *Canadian Psychological Review/Psychologie Canadienne*. 1978;19(2):158.
43. Bornmann L, Daniel HD. What do citation counts measure? A review of studies on citing behavior. *Journal of documentation*. 2008;64(1):45-80.

44. Erikson MG, Erlandson P. A taxonomy of motives to cite. *Social studies of science*. 2014;44(4):625-37.
45. Garfield E. The impact factor. *Current contents*. 1994;25(20):3-7.
46. Bornmann L, Daniel HD. What do we know about the h index? *Journal of the American Society for Information Science and technology*. 2007;58(9):1381-5.
47. Soteriades ES, Falagas ME. A bibliometric analysis in the fields of preventive medicine, occupational and environmental medicine, epidemiology, and public health. *BMC Public Health*. 2006;6:1-8.
48. Goodrum AA, McCain KW, Lawrence S, Giles CL. Scholarly publishing in the Internet age: a citation analysis of computer science literature. *Information Processing & Management*. 2001;37(5):661-75.
49. Robins RW, Gosling SD, Craik KH. An empirical analysis of trends in psychology. *American Psychologist*. 1999;54(2):117.
50. Huang J, Gates AJ, Sinatra R, Barabasi AL. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proc Natl Acad Sci U S A*. 2020;117(9):4609-16.
51. Van den Besselaar P, Sandström U. Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PloS one*. 2017;12(8):e0183301.
52. Van Leeuwen TN, Moed HF, Tijssen RJ, Visser MS, Van Raan AF. First evidence of serious language-bias in the use of citation analysis for the evaluation of national science systems. *Research Evaluation*. 2000;9(2):155-6.
53. Callaham M, Wears RL, Weber E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Jama*. 2002;287(21):2847-50.
54. Motoki F, Pinho Neto V, Rodrigues V. More human than human: measuring ChatGPT political bias. *Public Choice*. 2023;198(1-2):3-23.
55. Wan Y, Pu G, Sun J, Garimella A, Chang K-W, Peng N. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:231009219*. 2023.
56. Fang X, Che S, Mao M, Zhang H, Zhao M, Zhao X. Bias of AI-generated content: an examination of news produced by large language models. *Sci Rep*. 2024;14(1):5224.

57. OpenAI. Is ChatGPT Biased? 2024 [Available from: <https://help.openai.com/en/articles/8313359-is-chatgpt-biased>].
58. Navarro-Prado S, Schmidt-RioValle J, Montero-Alonso MA, Fernández-Aparicio Á, González-Jiménez E. Unhealthy Lifestyle and Nutritional Habits Are Risk Factors for Cardiovascular Diseases Regardless of Professed Religion in University Students. *Int J Environ Res Public Health*. 2018;15(12).
59. Patiño-Masó J, Gras-Pérez E, Font-Mayolas S, Baltasar-Bagué A. [Cocaine abuse and multiple use of psychoactive substances in university students]. *Enferm Clin*. 2013;23(2):62-7.
60. Bhengu KN, Naidoo P, Singh R, Mpaka-Mbatha MN, Nembe N, Duma Z, et al. Immunological Interactions between Intestinal Helminth Infections and Tuberculosis. *Diagnostics (Basel)*. 2022;12(11).
61. Garedew-Kifelew L, Wondafrash N, Feleke A. Identification of drug-resistant Salmonella from food handlers at the University of Gondar, Ethiopia. *BMC Res Notes*. 2014;7:545.
62. Campbell P, Jordan KP, Smith BH, Scotland G, Dunn KM. Chronic pain in families: a cross-sectional study of shared social, behavioural, and environmental influences. *Pain*. 2018;159(1):41-7.
63. Bair MJ, Robinson RL, Katon W, Kroenke K. Depression and pain comorbidity: a literature review. *Arch Intern Med*. 2003;163(20):2433-45.
64. Raymond-Lezman JR, Riskin SI. Benefits and Risks of Sun Exposure to Maintain Adequate Vitamin D Levels. *Cureus*. 2023;15(5):e38578.
65. Holick MF. Sunlight and vitamin D for bone health and prevention of autoimmune diseases, cancers, and cardiovascular disease. *Am J Clin Nutr*. 2004;80(6 Suppl):1678s-88s.

Chapter 2: Punishment is Slower than Cooperation or Defection in Online Network Games

2.1 Abstract

Punishment serves as a balancing force that dissuades people from acting selfishly, which complements cooperation as an essential characteristic for the prosperity of human societies. Past studies using economic games with two options (cooperation and defection) reported that cooperation decisions are generally faster than defection decisions and that time pressure possibly induces human players to be more intuitive and thus cooperative. However, it is unclear where punishment decisions sit on this time spectrum. Therefore, we recruited human players and implemented two series of online network games with cooperation, defection, and punishment options. First, we find that punishment decisions are slower than cooperation or defection decisions across both game series. Second, we find that imposing experimental time pressure on in-game decisions neither reduces nor increases the frequency of punishment decisions, suggesting that time pressure may not directly interact with the mechanisms that drive players to choose to punish.

2.2 Introduction

Punishment is widely observed in various forms: recent and ongoing examples include the gun violence epidemic in the United States (1, 2), the waging of war under the guise of security or using punishment as a just cause for war (3-6), institutional responses to crimes (7), and anti-prevention behavior during the COVID-19 pandemic (8-10). Games are no exception to this rule; for example, in competitive, multiplayer games, players may choose to sacrifice some of their own chances of victory to substantially reduce the chances of victory for their opponents. In all these scenarios, punishment involves paying a cost to harm others and serves as an important counterpart

to cooperation (paying a cost to benefit others), a characteristic that has been integral to the formation of human societies (11-16). However, in addition to the costs associated with choosing to punish, punishers incur societal and emotional penalties: institutions that punish in unjust or cruel ways are labeled tyrannical (17), while individuals who punish others make themselves a target for retaliation or reduced reputation (18, 19). This divide between punishment's potentially important relationship with cooperation and the negatives associated with choosing to punish contributes to a lack of clarity about the mechanisms that drive punishment. Researchers have generally proposed that punishment serves to sustain cooperation through negative reinforcement: prominent theories suggest that punishment promotes the development of cooperation in social networks by reducing the payoff of defectors and converting them into cooperators (20-22) or that punishment acts as a balancing mechanism that allows people to reduce perceived inequality by allowing poorer individuals to spend a small amount to take a larger amount away from their richer neighbors (23-27).

Decision times in experimental games have been used in several studies as a tool to understand cooperation (28-32). Employing an experimental approach in which study participants could only choose between cooperation decisions or defection decisions (not paying anything and not affecting others), one study found that choosing cooperation is faster than choosing defection when participants have many cooperative neighbors but slower when one's neighbors are less cooperative (30). This finding may be the result of "decision conflict" (33): a mismatch between the current state of connected neighbors (the "social environment") and the participant's intended choice, eliciting a feeling of conflict, which slows down decision-making speed; this conflict may compound the general sense that defection may provoke more feelings of conflict than cooperation. Other studies (28, 29) found that participants' intuition and social heuristics may

prefer cooperation over defection, and therefore, when time pressure (limiting the amount of time available for participants to make decisions) was imposed, participants might be more likely to mobilize intuition over deliberation, resulting in more frequent cooperation. However, these findings have not been clearly reproduced in follow-up studies (34-36).

In this study, we examine the unexplored relationship between punishment and decision time using two experiments consisting of online network games to better understand how punishment decision-making differs from decision-making that leads to cooperation or defection. In Experiment 1, we compared the decision times of punishment to those of cooperation and defection. Here, we hypothesized that decisions involving punishment would take longer than those involving cooperation or defection in general (regardless of decision mismatches) because choosing to punish requires players to decide to pay an immediate cost, give up any potential fitness gains from reciprocal cooperation in the short term, and prepare for potential retaliation from those who are punished. Then, in Experiment 2, we determined if experimental time pressure could alter the distribution of players' decisions by reducing the occurrence of punishment and increase the occurrence of cooperation. By minimizing the amount of time available to make decisions, we would expect players to be inclined to make “intuitive” choices – that is, to opt to cooperate and not to punish.

2.3 Methods Overview

a) Experimental Design

We implemented two series of repeated public goods game (PGG) by adapting an online, network-based framework (30, 37) previously used to study cooperation and decision time by introducing a punishment option. The games allow players to interact with a dynamic group of

other players over time and strategize ways to improve their status based on the decisions and statuses of others. Players were recruited from around the world using Amazon Mechanical Turk (mTurk) between March 2018 and December 2018 for Experiment 1 and between March 2023 and May 2023 for Experiment 2. The experiments were approved by and performed according to guidelines and regulations set by the UCLA Office of Research Administration (UCLA IRB#16-001920). Informed consent was obtained online from all participants. At the end of all games, accumulated in-game wealth was converted at a rate of 2,000 points to 1 USD; players were also compensated USD 3 for participating in the games. Recruited players were assigned to dynamic networks and asked to interact in the PGG over 15 rounds. Each network was generated by arranging each game session's players into a Erdős-Renyi random graph in which 30% of all possible ties were present at the start of the experiment.

Players were first invited to play two practice rounds to introduce them to the game format and layout of elements within the experimental platform (**Fig. 2.1**). Once these two practice rounds were completed, all players in the game were randomly allocated a high or low quantity of in-game points and asked to interact with each other. Game sessions that did not recruit enough players from mTurk did not progress to the practice round stage and were closed to further entry; players who previously completed a game (session) were prohibited from participating in future sessions.

At the start of each game round, players entered the decision phase, where they were given one of three options: cooperation (the player pays 50 points per connected player to have all connected players gain 100 points), defection (the player pays nothing, not affecting other players), and punishment (the player pays 50 points per connected player to have all connected players lose 100 points). No institutional/central punishment (38) or sanctioning (39) occurred. Each option

was indiscriminate: players could not choose to interact with some connected players and not interact with others and the option that was selected resulted in the appropriate change to all connected players. Once players made their decisions, they were shown the decisions made by other connected players in the decision phase and an updated tally of in-game points for the focal player and any connected players was shown. In Experiment 1, sessions were randomly assigned to have neighboring players' wealth be visible or invisible (with the update at the end of the decision phase mirroring this randomization); however, statistical analysis (not shown) implied that there was no effect of wealth visibility on the frequency of punishment or decision times, allowing us to keep wealth visible in all sessions of Experiment 2. The effect of wealth visibility on other outcomes such as subjective well-being has been reported elsewhere (without examining or including decision time in analyses) (40).

Players were then allowed to update their network connections by answering a series of questions in the rewiring phase. In this phase, 30% of all possible ties in the network were randomly selected by breadboard. For each extant tie, one player from the tie pairing was chosen at random and asked if they wanted to break the tie. If the player chose to break the tie, the connection was broken for the next round; if the player chose not to break the tie, the tie was retained. For each non-extant tie, both players in the potential tie were asked if they wanted to make a new connection; if and only if both players agreed, the tie was constructed and carried over to the next game round. The networks were then reconstructed based on the results of the rewiring phase and the subsequent round began until all 15 rounds of the game were completed.

b) Decision Time

For this study, we defined decision time (a.k.a. response time) of the cooperation-defection-punishment decision-making as the time interval between two timestamps which were recorded by breadboard: the first timestamp recorded the appearance of the screen which allowed players to choose between cooperation, defection, or punishment (shown in **Fig. 2.1**), while the second timestamp recorded when players clicked one of the buttons representing cooperation, defection, or punishment.

c) Punishment Mechanisms

To guide our analysis of the different rationales for punishment decisions in our games, we classified punishment decisions into four categories based on prevailing theorized punishment mechanisms in the literature (20, 25-27, 41). These mechanisms are not mutually exclusive and can overlap.

First, we classified punishment for copying or retaliation as punishment that occurred when at least one connected player in the prior round chose to punish. Punishment of this type may be beneficial because copying others' behavior and learning from copying is known to be advantageous regardless of the reason for the copied behavior (42, 43). Retaliation occurs when the punisher chooses to punish punishers from the last encounter to dissuade them from punishing in the future (41, 44-46). Revenge is a related concept which pairs retaliation with an emotional argument that punishers should be punished in return. Unfortunately, our experimental setting cannot precisely distinguish between these related mechanisms and therefore we combine punishment decisions that could be classified as supported by either mechanism into one single category.

Second, we defined punishment for negative reinforcement as punishment that occurred when the cooperation rate among connected players in the prior round was less than 50%. In our experimental setting, players paid 50 units for each connecting player when choosing to cooperate. As a result, the cost-benefit ratio of choosing cooperation was 0.5 (as the cost of cooperation is greater than the expected benefits since connected players could choose either to defect or punish); this means that cooperation would decay very quickly if the cooperation rate fell below 50%. Therefore, punishing connected neighbors may contribute to long-term reciprocal cooperation in social networks when the networks are not very cooperative (20, 21, 47, 48).

Third, we defined punishment for inequality aversion as punishment that occurred when a focal player's wealth was less than the average wealth of connected players in the prior round. Since the cost to punish is on average lower than the expected loss from being punished, the degree of economic inequality (the difference of in-game points) is reduced when poorer players punish connected richer players. Therefore, players could be motivated to punish by inequality aversion (25-27, 49).

Fourth and finally, punishment decisions that did not meet any of the above three definitions were classified as unclassified or de novo punishment, possibly reflecting a base inclination to choose punishment. It is possible that some players wanted to "try out" the punishment option even if they were not poor and connected to mainly cooperative neighbors.

d) Statistical Analysis

The structure of the game sessions and the arrangement of players into networks meant that multiple observations were made for a single player across multiple rounds in each session. We account for this hierarchical data structure using multilevel random intercepts models, utilizing R

version 4.3.1 (50) and the lme4 statistical package. Detailed results from these multilevel models are listed in the Supplementary Materials.

While prior work (28, 30, 51) utilized a log10-transformation when analyzing decision times because decision times are generally only left-bounded by zero, time data from games involving time pressure would also be right-bounded by the time limit. As a result, we chose to omit the transformation to keep model estimates from limited and non-limited data on the same scale.

2.4 Experiment 1

a) Results

In Experiment 1, 719 unique players (mean: 14.9/game, range: 9-25/game) made 9,776 decisions of cooperation, defection, or punishment. Cooperation was chosen 4,878 times (49.4%, 95% confidence interval for proportion [CI]: 48.4-50.5%), defection was chosen 4,336 times (43.9%, 95% CI: 42.9-45.0%), and punishment was chosen 562 times (6.6%, 95% CI: 5.6-7.7%) (**Fig. 2.2A**). The mean degree (the number of connections a player had at any moment) in Experiment 1 was 5.9 (range: 0-17). At the end of 15 rounds, the mean accumulated wealth across all players was 1,584 in-game points (equivalent to USD 0.79). Therefore, combined with the USD 3 participation award, players in Experiment 1 obtained USD 3.79 on average.

We found that that punishment decisions (mean = 7.0 seconds, 95% CI: 6.2-7.8 seconds) are slower than cooperation decisions (mean = 5.3 seconds, 95% CI: 5.0-5.5 seconds, p of punishment vs. cooperation from random intercepts model: 0.004) and defection decisions (mean = 5.8 seconds, 95% CI: 5.6-6.1 seconds, p of punishment vs. defection: 0.032) (**Fig. 2.2B**; **Supplementary Table S2.1**).

Punishment due to copying or retaliation encompassed 1.4% (149 instances) of all decisions in Experiment 1 (**Fig. 2.2C**) and the mean decision time of such decisions was 5.5 seconds (95% CI: 4.4-6.7 seconds) (**Fig. 2.2D**), which tended to be the fastest among the four punishment categories (p vs. negative reinforcement = 0.66, p vs. inequality aversion = 0.16, p vs. unknown punishment = 0.02, p vs. all three other categories = 0.01). Punishment due to negative reinforcement comprised 2.9% (307 instances) of all decisions (**Fig. 2.2C**) with a mean decision time of 6.3 seconds (95% CI: 5.3-7.3 seconds) (**Fig. 2.2D**), while punishment due to inequality aversion included 2.7% (286 instances) of all decisions (**Fig. 2.2C**) with a mean decision time of 7.0 seconds (95% CI: 5.9-8.2 seconds) (**Fig. 2.2D**). Unclassified punishment encompassed 1.4% (154 instances) of all decisions (**Fig. 2.2C**) with a mean decision time of 8.1 seconds (95% CI: 6.3-9.9 seconds) (**Fig. 2.2D**); these decisions tended to be longer than punishment for inequality aversion ($p = 0.73$) and longer than punishment decision times for the combination of the three abovementioned mechanisms ($p = 0.04$). Furthermore, having a punisher among one's network neighbors in the previous round (that is, having been punished in the previous round) was associated with a 2.4 second decrease in punishment decision time ($p = 0.009$, **Supplementary Table S2.5**).

b) Discussion

Experiment 1 allowed us to determine how decision times associated with different mechanisms of punishment relate to one another to better understand why punishment was on average slower than either cooperation or defection. The relative quickness of punishment for copying or retaliation compared to the other punishment mechanisms may reflect players needing less time to justify choosing to punish if connected players have punished previously, especially

since punishment was relatively infrequent compared to cooperation and defection and therefore players were not frequently connected to others who had punished before. Punishment for negative reinforcement was slower compared to punishment for copying or retaliation; this difference may be a product of players needing additional time to choose to punish to stop a perceived decay in cooperation in their social network. Increased decision time associated with inequality aversion is also understandable since punishing richer neighbors is a valid in-game strategy which takes a more self-interested viewpoint; players choosing this strategy may need more time to contemplate about their economic position in relation to other players in the session as opposed to the overall wealth of the group (which would apply more to the two previously mentioned mechanisms for punishment). We speculate that unclassified punishment having the longest average decision time may reflect the degree of feelings of decision conflict and may be triggered by feelings of remorse for taking points away from other players.

In summary, the small differences in decision time over the four punishment categories did not give us conclusive evidence about the specific mechanisms that slow down the punishment decision-making process. However, we identified the tendency for the decision times of punishment decisions to increase as the complexity or lack of clarity related to the driving mechanisms increased. For example, in copying or retaliation punishment, some proportion of connecting players are no longer cooperative and thus may deserve to be punished. Because our experimental setting requires players to make a decision that will affect all connected players, choosing to punish to copy or as retaliation will require players to punish both punishers and cooperators; such a situation could elicit feelings of conflict.

2.5 Experiment 2

Since Experiment 1 showed that punishment decisions are slower than cooperation or defection decisions, we hypothesized that setting a time limit on decision-making could reduce the number of punishment decisions (Hypothesis 1). However, we also considered an alternative hypothesis: since most of the punishment decisions that occurred in Experiment 1 were linked to postulated mechanisms, the number of punishment decisions would be unchanged unless the context that led to the punishment decision was addressed (Hypothesis 2).

To reconcile these two hypotheses, we developed Experiment 2, a second experimental series of 50 games which were randomly divided into 25 games with a time pressure setting (the TP+ setting) (**Fig. 2.1B**), in which player decisions had to be made in 3 seconds or less, and 25 other games with no time pressure (the TP- setting) (**Fig. 2.1A**). In the TP+ setting, if players did not confirm their choice within the time limit, the system automatically repeated their decision from the previous round; if players missed two decision-making opportunities due to inactivity, they were dropped from the remainder of the session (**Supplementary Fig. S2.1**). We chose to implement time pressure in Experiment 2 as a three-second limit on decision-making; we based our selection of three seconds on the distribution of decision times in Experiment 1 (**Supplementary Fig. S2.2**).

a) Results

In Experiment 2 (across both TP+ and TP- settings), 739 players (mean: 14.8/game, range: 8-20/game) made 10,654 decisions. Cooperation was chosen 4,185 times (39.3%, 95% CI: 38.3-40.3%), defection was chosen 5,790 times (54.4%, 95% CI: 53.4-55.4%), and punishment was chosen 679 times (6.4%, 95% CI: 5.4-7.4%). The mean degree across sessions in Experiment 2

was 5.7 (range: 0-16). At the end of 15 rounds, the mean accumulated wealth was 935 in-game units (equivalent to USD 0.47); with the USD 3 participation award, the average reward for participants in Experiment 2 was USD 3.47.

In the TP+ setting, punishment was chosen 338 times out of 5,407 decisions (6.3%, 95% CI: 4.8-7.7%), while in the TP- setting, punishment was chosen 341 out of 5,247 decisions (6.5%, 95% CI: 5.1-7.9%). While there was no substantial difference in the frequency of punishment comparing the TP+ and TP- setting ($p = 0.475$) (**Fig. 2.3A**), all decision-making was faster in the TP+ games as expected (**Fig. 2.3C**), suggesting that Experiment 2 was successfully implemented.

We then evaluated if time pressure reduced the frequencies of any of the three justifiable mechanisms of punishment that we highlighted in Experiment 1. Results show that the frequency of the punishment decisions due to copying or retaliation did not change between the TP- and TP+ settings (1.5% and 1.7% respectively, $p = 0.893$), the frequency of punishment for negative reinforcement did not change (3.5% and 3.5%, $p = 0.634$), the frequency of punishment for inequality aversion did not change (3.7% and 3.6%, $p = 0.440$), and the frequency of unclassified punishment did not change (1.3% and 1.1%, $p = 0.645$) (**Fig. 2.4**).

b) Discussion

Although we successfully predicted that punishment would be slower than cooperation or defection, we did not observe the hypothesized reduction in punishment associated with time pressure, as game sessions from the TP+ and TP- settings in Experiment 2 had a similar frequency of punishment decisions (6.3% vs. 6.5%, respectively). Furthermore, we found no substantial difference in frequency for any of the mechanisms for punishment when comparing game sessions with and without time pressure. We were particularly interested in observing a potential decrease

in the frequency of punishment for negative reinforcement; previous studies have linked deliberation to the degradation of cooperation with others (52-54), suggesting that limiting deliberation could potentially dissuade players from choosing the punishment option in favor of cooperation. However, we did not observe this decrease, a result which falls in line with other studies (36, 55, 56) that limit their endorsement of the link between intuition and cooperation (and in our view, the link between deliberation and non-cooperation, which includes punishment).

In summary, time pressure did not substantially reduce the frequency of punishment decisions in general, potentially because prohibiting slow decisions does not address the specific mechanisms that drive punishment decisions. Since our investigation of external time pressure only addresses Hypothesis 1, further research to address Hypothesis 2 should be conducted by addressing each of the individual mechanisms for punishment in an experimental manner and evaluating how the distribution of decision-making changes.

2.6 General Discussion

Using two series of online, network-based public goods games, we found that on average, players spend more time when choosing to punish compared to choosing to cooperate or defect. Follow-up experimentation based on this result evaluated if experimental time pressure could reduce the occurrence of punishment in favor of cooperation; no such reduction occurred.

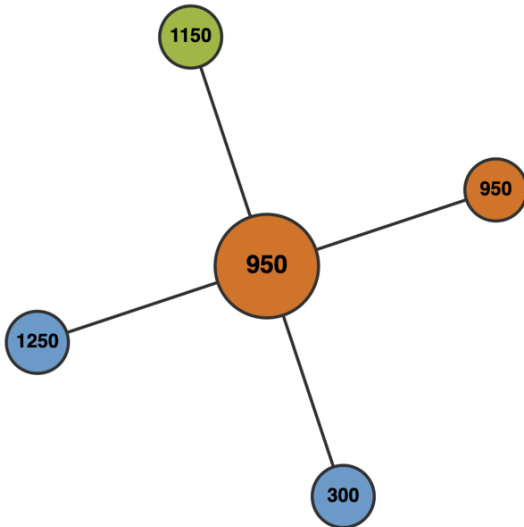
Our experimental efforts to determine if time pressure could reduce the occurrence of punishment were partially inspired by recent violent and damaging trends in the United States and around the world that reflect a widespread inclination to punish. We aimed to identify a practical method to suppress punishment or behavior that harms others in a social context; however, we were unable to do so. Reducing the occurrence of punishment is a difficult task which requires addressing the source of the initial motivation to punish.

Our study has several limitations that could limit the translation of our results to other settings in social and behavioral sciences. First, we did not measure several cultural, emotional, or psychological characteristics that could influence the relationship between punishment decision-making and decision speed in our experiments. Recent efforts analyzing the relationship between decision times and cooperation have explored how factors such as belief about other individuals' intentions (57), personal time preferences (58), and age (59) influence cooperation decision-making. Furthermore, recent work has found that the reputation of punishers decreases when they punish quickly, but when the punishment was slow, the punishers' reputation instead increased (60). Future studies should aim to evaluate these each of these characteristics in detail as they relate to punishment and decision times.

Second, we kept the payoff structure consistent across experimental settings with and without time pressure to evaluate the independent effect of time pressure. Our experimental payoff structure did not penalize punishers any more than cooperators were incentivized; however, this may not reflect realistic interactions in human society because choosing punishment is generally perceived as negative and may incur greater than expected costs for the punisher. It is possible that modifying the payoff structure to be more extreme (i.e., paying a greater cost to inflict the same level of harm shown in Experiments 1 and 2) could return greater differences in experimental participants' decision-making profile.

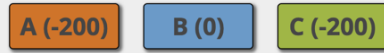
2.7 Figures and Tables

A



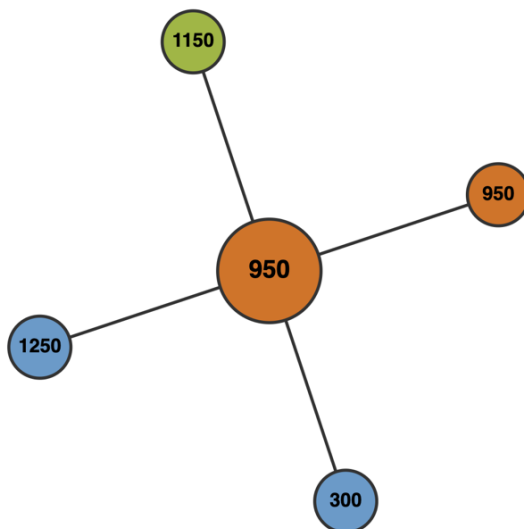
- A** If you **choose A**, you **pay 50 points** for each player you are connected to and each of them **gains 100 points**.
- B** If you **choose B**, you do not pay any points and do not change the points of the players you are connected to.
- C** If you **choose C**, you **pay 50 points** for each player you are connected to and each of them **loses 100 points**.

Each player you are connected to has the same choice. Regardless of your choice, for each of them that chooses A, you **gain 100 points**.



remaining time : 00:06

B



- A** If you **choose A**, you **pay 50 points** for each player you are connected to and each of them **gains 100 points**.
- B** If you **choose B**, you do not pay any points and do not change the points of the players you are connected to.
- C** If you **choose C**, you **pay 50 points** for each player you are connected to and each of them **loses 100 points**.

Each player you are connected to has the same choice. Regardless of your choice, for each of them that chooses A, you **gain 100 points**.

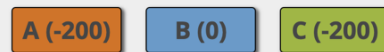


Figure 2.1. Example player's screenshot without (A) and with (B) time pressure. The focal player is represented by the larger, central circle highlighted in orange, while the smaller surrounding circles represent connected players in the same game session (the colors of the circles indicate the choice of each player in the round prior). In the setting with time pressure (B), a horizontal bar appeared on the player's screen showing the remaining time the player had to make their decision (this setting only appeared in Experiment 2). In the setting without time pressure (A), no time pressure was implemented, and no bar

appeared; panel A is representative of all sessions in Experiment 1. Values in the circles represent players' in-game points.

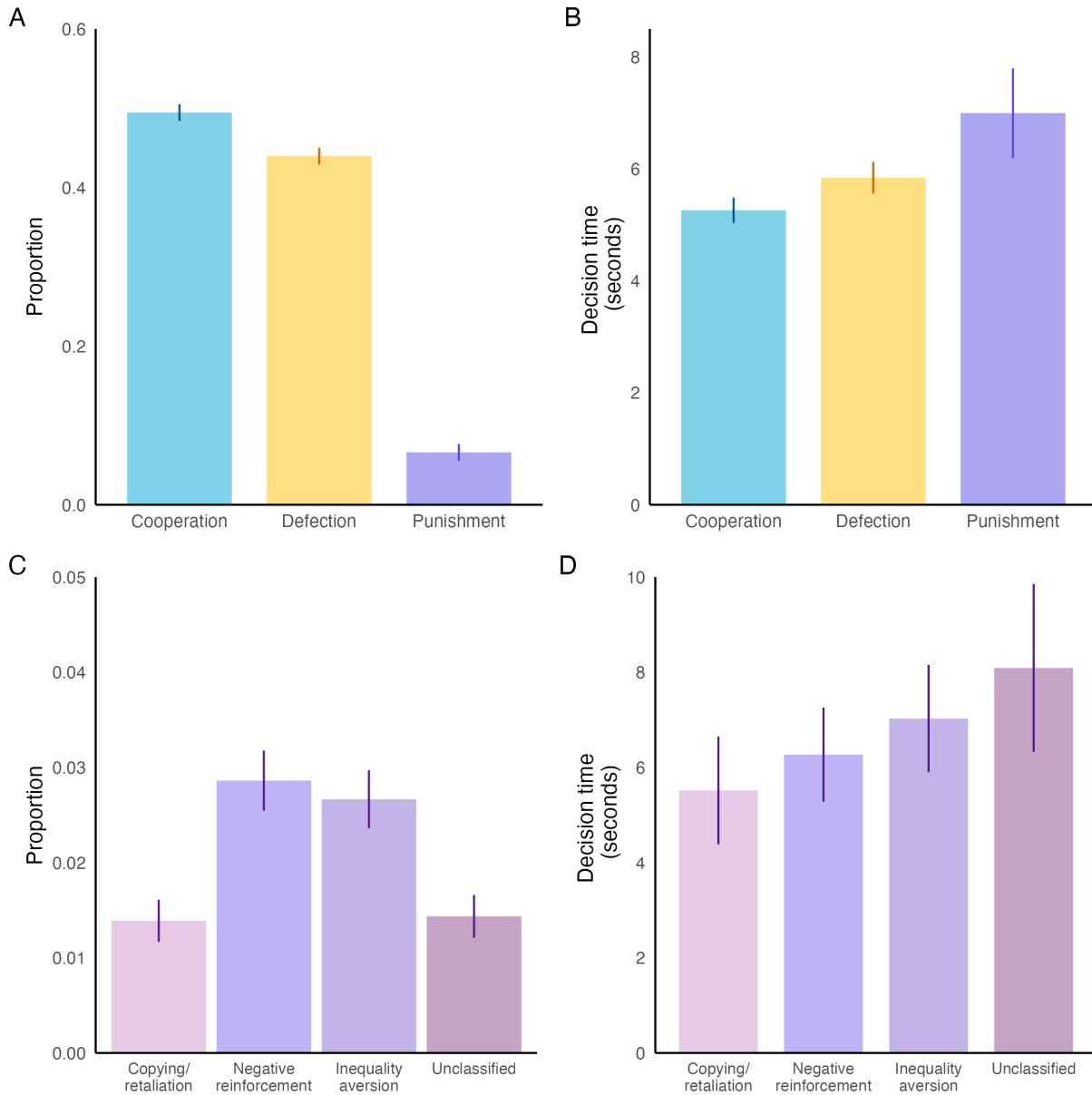


Figure 2.2. Punishment is rare and slower than cooperation or defection in Experiment 1. A: Cooperation and defection are more common. Punishment is relatively uncommon (6.6%). **B:** On average, punishment decisions took longer to make compared to defection or cooperation. **C:** The proportion of decisions reflecting each of the mechanisms of punishment. The total proportion represented in panel C exceeds the actual proportion of punishment decisions in the experiment because the mechanism categories can overlap. **D:** Mean decision times for each punishment mechanism. The mechanisms display a tendency to slow down as the complexity of the punishment decision increases. Bars indicate 95% confidence intervals of proportions or decision time.

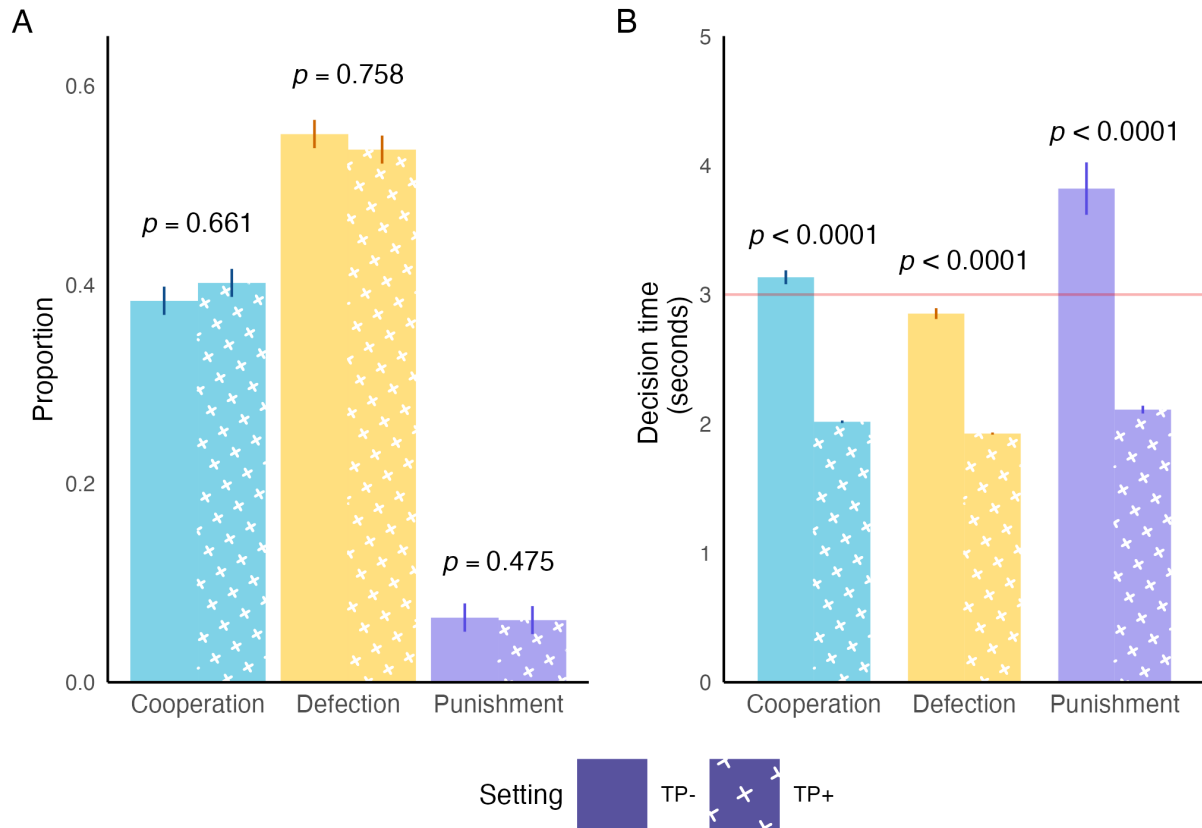


Figure 2.3. Punishment is slower than cooperation or defection, even under time pressure. **A:** There was no substantial difference in the frequencies of each decision type comparing sessions with time pressure (TP-; left, bars without crosses) and sessions with time pressure (TP+; right, bars with white crosses). **B:** Punishment was slower than cooperation and defection in both the TP- and TP+ settings. The red line indicates the three-second time pressure boundary. Error bars represent 95% confidence intervals of proportions or means.

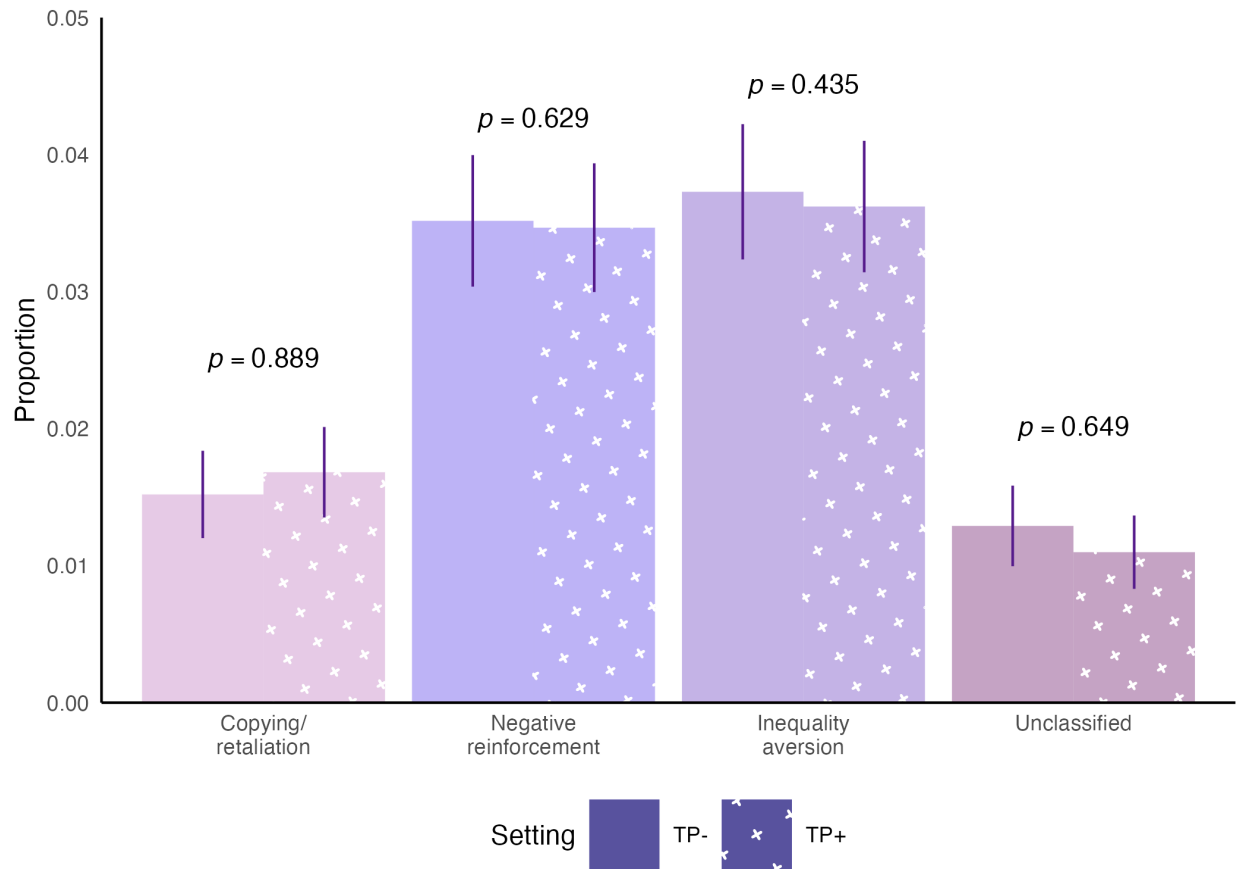
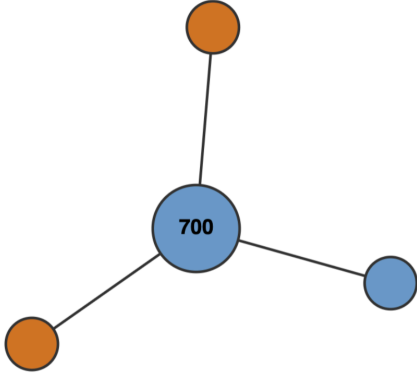


Figure 2.4. Time pressure does not change the frequency of mechanism-specific punishment. In Experiment 2, we observed no substantial difference in the proportion of punishment that reflects each of the competing hypotheses for punishment mechanisms with the implementation of time pressure. Because the punishment motivation categories overlap, the total proportion of the bars exceeds the true proportion of punishment in the experiment. P-values are from random intercepts mixed effects models testing differences between the TP- (left, bars without crosses) and TP+ (right, bars with white crosses) settings. Bars represent 95% confidence intervals of proportions.

2.8 Supplementary Information

You will be dropped in: 00:22



Practice round 1 of 1

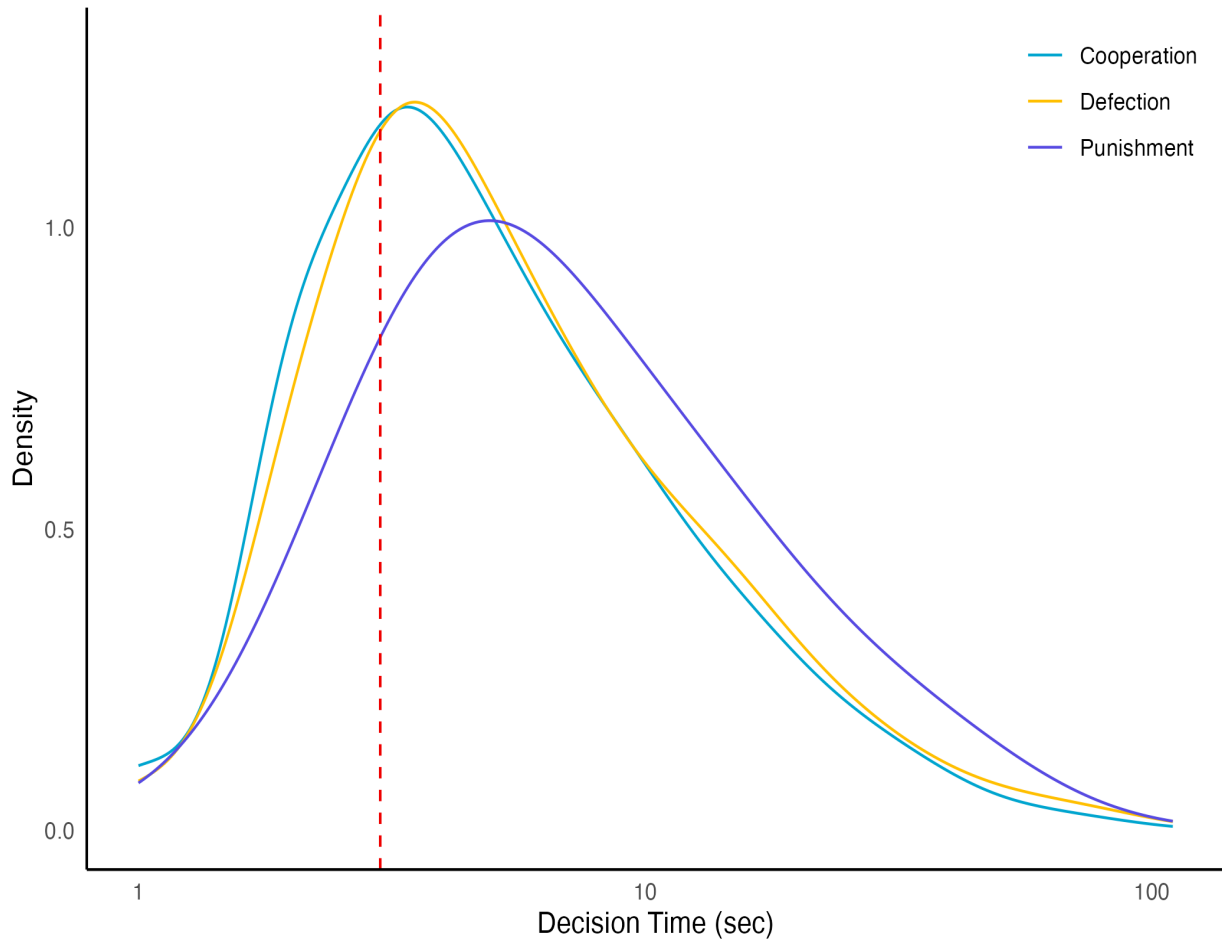
*These rounds will not change your score.
Your score will be reset before the game starts.*

Last round 2 player(s) you are connected to paid 50 each to contribute a total of 200 points to you. The player(s) also contributed to everyone else they are connected to.

Next

You have been dropped for being idle.

Supplementary Figure S2.1. Example message shown to dropped players in the time pressure condition. Players in the TP+ condition who did not click on a button within the allotted time were first given a warning (and their choice from the previous round was repeated). If players did not click in two different rounds, they were dropped from future rounds of the experiment and were shown the message “You were dropped for being idle.”



Supplementary Figure S2.2. Distributions of decision times for cooperation, defection, and punishment in Experiment 1. The median decision times for cooperation and defection in Experiment 1 are close to three seconds (the red dashed line) while the median decision time for punishment is noticeably to the right of the red line.

	Exp. 1	Exp. 2, TP-	Exp. 2, TP+
N	9,776	5,247	4,066
Fixed Effects			
Defection vs. cooperation	0.30411 (0.23036) [p = 0.18684]	-0.273991 (0.106943) [p = 0.01048]	-0.051397 (0.01966) [p = 0.0089]
Punishment vs. cooperation	1.21569 (0.42149) [p = 0.00393]	0.47002 (0.147547) [p = 0.00145]	0.115546 (0.02799) [p = 0.000037]
Round	-0.16708 (0.01916) [p < 0.0001]	-0.020730 (0.006745) [p = 0.00213]	-0.006124 (0.00109) [p < 0.0001]
Intercept	6.79184 (0.30462) [p < 0.0001]	3.329949 (0.122798) [p < 0.0001]	2.090189 (0.023485) [p < 0.0001]
Random Effects			
Player-level variance	11.789 (3.434)	1.58559 (1.2592)	0.07370 (0.27147)
Game-level variance	1.824 (1.350)	0.08172 (0.2859)	0.00315 (0.05613)
Residual variance	64.809 (8.050)	4.38568 (2.0942)	0.08420 (0.29018)

Supplementary Table S2.1. Multilevel random intercepts models for decision times for all 3 experimental settings. The reference category for decision type was cooperation. Standard errors for fixed effects and standard deviations for random effects are shown in parentheses. P-values are shown in square brackets.

	Punishment	Cooperation	Defection
N	10,654	10,654	10,654
Fixed Effects			
Time Pressure	0.77967 (0.39389, 1.54330) [p = 0.475]	1.30413 (0.39811, 4.27207) [p = 0.661]	0.80540 (0.20294, 3.19637) [p = 0.758]
Round	0.97390 (0.95276, 0.99551) [p = 0.182]	0.9530 (0.93636, 0.96993) [p < 0.00001]	1.05696 (1.03725, 1.07704) [p < 0.00001]
Intercept	0.02495 (0.01506, 0.04132) [p < 0.00001]	0.44260 (0.18905, 1.03621) [p = 0.0604]	1.03491 (0.38632, 2.77243) [p = 0.94558]
Random Effects			
Player-level variance	6.908 (2.6283)	27.00 (5.196)	35.06 (5.921)
Game-level variance	0.641 (0.8006)	2.42 (1.556)	3.392 (1.842)

Supplementary Table S2.2: Multilevel logistic random intercepts model for the effect of time pressure on the odds of punishment, cooperation, and defection in Experiment 2 (results for Fig. 2.3A). Estimates shown are exponentiated and represent odds ratios. 95% confidence intervals for fixed effects odds ratios and standard deviations of random effects are shown in parentheses. P-values are shown in square brackets.

	Copying/ retaliation	Negative reinforcement	Inequality aversion	Unclassified
N	10,654	10,654	10,654	10,654
Fixed Effects				
Time Pressure	0.93220 (0.34538, 2.51606) [p = 0.889]	0.83141 (0.39301, 1.75886) [p = 0.629]	0.76709 (0.39435, 1.49212) [p = 0.435]	0.85826 (0.44474, 1.6524) [p = 0.649]
Round	1.02502 (0.98436, 1.0580) [p = 0.269]	1.03644 (1.00987, 1.063471) [p = 0.0069]	1.02228 (0.99594, 1.04932) [p = 0.098]	0.83194 (0.79264, 0.87319) [p < 0.00001]
Intercept	0.00575 (0.00269, 0.01228) [p < 0.00001]	0.01049 (0.00594, 0.01853) [p < 0.00001]	0.01042 (0.00626, 0.01733) [p < 0.00001]	0.02098 (0.01249, 0.03523) [p < 0.00001]
Random Effects				
Player-level variance	2.722 (1.650)	4.903 (2.214)	6.9448 (2.6374)	3.919 (1.9798)
Game-level variance	2.117 (1.455)	1.007 (1.004)	0.4017 (0.6338)	0.358 (0.5984)

Supplementary Table S2.3: Multilevel logistic random intercepts model for the effect of time pressure on the odds of individual punishment mechanisms in Experiment 2 (results for Fig. 2.4). Estimates shown are exponentiated and represent odds ratios. 95% confidence intervals for fixed effects odds ratios and standard deviations of random effects are shown in parentheses. P-values are shown in square brackets.

	Exp. 1
N	9,776
Fixed Effects	
Negative reinforcement	0.5307 (1.1866) [p = 0.655]
Inequality aversion	1.5698 (1.1200) [p = 0.162]
Unclassified	3.7815 (1.6083) [p = 0.019]
Round	-0.1209 (0.1023) [p = 0.2377]
Intercept	6.2481 (1.5349) [p < 0.0001]
Random Effects	
Player-level variance	23.47 (4.845)
Game-level variance	11.29 (3.360)
Residual variance	64.44 (8.028)

Supplementary Table S2.4. Multilevel random intercepts models for decision times to compare individual punishment mechanisms. The reference category for decision type was punishment for copying/retaliation. Standard errors for fixed effects and standard deviations for random effects are shown in parentheses. P-values are shown in square brackets.

	Exp. 1
N	508
Fixed Effects	
Last punished	-2.4355 (0.9306) [p = 0.0091]
Round	-0.1489 (0.1003) [p = 0.1384]
Intercept	9.1691 (1.1074) [p < 0.00001]
Random Effects	
Player-level variance	23.89 (4.887)
Game-level variance	10.43 (3.229)
Residual variance	64.10 (8.006)

Supplementary Table S2.5: Multilevel random intercepts models for the effect of having been punished in the prior round on punishment decision times in Experiment 1. Standard errors for fixed effects and standard deviations for random effects are shown in parentheses. P-values are shown in square brackets.

2.9 References

1. Thompson AJ. Gun violence in the United States: a public health epidemic. *Public Health-Social and Behavioral Health: IntechOpen*; 2012.
2. Mueller KL, Lovelady NN, Ranney ML. Firearm injuries and death: A United States epidemic with public health solutions. *PLOS Global Public Health*. 2023;3(5):e0001913.
3. Kemp KW. Punishment as Just Cause for War. *Public Affairs Quarterly*. 1996;10(4):335-53.
4. Stohl M. The Global War on Terror and State Terrorism. *Perspectives on Terrorism*. 2008;2(9):4-10.
5. Luban D. War as Punishment. *Philosophy & Public Affairs*. 2011;39(4):299-330.
6. Troianovski A. Why Vladimir Putin Invokes Nazis to Justify His Invasion of Ukraine. *New York Times*. 2022 March 17, 2022.
7. Beccaria C, Parzen J. *On Crimes and Punishments and Other Writings*. Thomas A, Ballerini L, Ciavolella M, editors: University of Toronto Press; 2008.
8. Johnson NF, Velásquez N, Restrepo NJ, Leahy R, Gabriel N, El Oud S, et al. The online competition between pro- and anti-vaccination views. *Nature*. 2020;582(7811):230-3.
9. Prieto Curiel R, González Ramírez H. Vaccination strategies against COVID-19 and the diffusion of anti-vaccination views. *Scientific Reports*. 2021;11(1):6626.
10. Taylor S, Asmundson GJG. Negative attitudes about facemasks during the COVID-19 pandemic: The dual importance of perceived ineffectiveness and psychological reactance. *PLoS One*. 2021;16(2):e0246317.
11. Axelrod R, Hamilton WD. The evolution of cooperation. *Science*. 1981;211(4489):1390-6.
12. Boyd R, Richerson PJ. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2009;364(1533):3281-8.
13. Melis AP, Semmann D. How is human cooperation different? *Philos Trans R Soc Lond B Biol Sci*. 2010;365(1553):2663-74.
14. Rand DG, Nowak MA. Human cooperation. *Trends in Cognitive Sciences*. 2013;17(8):413-25.
15. Burkart JM, Allon O, Amici F, Fichtel C, Finkenwirth C, Heschl A, et al. The evolutionary origin of human hyper-cooperation. *Nature Communications*. 2014;5(1):4747.

16. Apicella CL, Silk JB. The evolution of human cooperation. *Current Biology*. 2019;29(11):R447-R50.
17. Charles Louis de Secondat BdM. Book VI: Consequences of the Principles of Different Governments with Respect to the Simplicity of Civil and Criminal Laws, the Form of Judgments, and the Inflicting of Punishments. In: Stewart P, editor. *The Spirit of Laws*, 1748.
18. Krasnow MM, Cosmides L, Pedersen EJ, Tooby J. What Are Punishment and Reputation for? *PLOS ONE*. 2012;7(9):e45662.
19. Jordan JJ, Kteily NS. How reputation does (and does not) drive people to punish without looking. *Proceedings of the National Academy of Sciences*. 2023;120(28):e2302475120.
20. Fehr E, Gächter S. Cooperation and Punishment in Public Goods Experiments. *The American Economic Review*. 2000;90(4):980-94.
21. Fehr E, Gächter S. Altruistic punishment in humans. *Nature*. 2002;415(6868):137-40.
22. Gächter S, Renner E, Sefton M. The Long-Run Benefits of Punishment. *Science*. 2008;322(5907):1510-.
23. Masclet D, Villeval M-C. Punishment, inequality, and welfare: a public good experiment. *Social Choice and Welfare*. 2008;31(3):475-502.
24. Houser D, Xiao E. Inequality-seeking punishment. *Economics Letters*. 2010;109(1):20-3.
25. Raihani NJ, McAuliffe K. Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biol Lett*. 2012;8(5):802-4.
26. Raihani NJ, Bshary R. Punishment: one tool, many uses. *Evolutionary Human Sciences*. 2019;1.
27. Deutchman P, Bračić M, Raihani N, McAuliffe K. Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evolution and Human Behavior*. 2021;42(1):12-20.
28. Rand DG, Peysakhovich A, Kraft-Todd GT, Newman GE, Wurzbacher O, Nowak MA, et al. Social heuristics shape intuitive cooperation. *Nat Commun*. 2014;5:3677.
29. Evans AM, Dillon KD, Rand DG. Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas. *J Exp Psychol Gen*. 2015;144(5):951-66.
30. Nishi A, Christakis NA, Evans AM, O'Malley AJ, Rand DG. Social Environment Shapes the Speed of Cooperation. *Sci Rep*. 2016;6:29622.

31. Evans AM, Rand DG. Cooperation and decision time. *Curr Opin Psychol.* 2019;26:67-71.
32. Alós-Ferrer C, Garagnani M. The cognitive foundations of cooperation. *Journal of Economic Behavior & Organization.* 2020;175:71-85.
33. Diederich A. Decision making under conflict: Decision time as a measure of conflict strength. *Psychonomic bulletin & review.* 2003;10(1):167-76.
34. Tinghög G, Andersson D, Bonn C, Böttiger H, Josephson C, Lundgren G, et al. Intuition and cooperation reconsidered. *Nature.* 2013;498(7452):E1-2; discussion E-3.
35. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour.* 2018;2(9):637-44.
36. Bouwmeester S, Verkoeijen PP, Aczel B, Barbosa F, Bègue L, Brañas-Garza P, et al. Registered replication report: Rand, greene, and nowak (2012). *Perspectives on Psychological Science.* 2017;12(3):527-42.
37. Nishi A, Shirado H, Rand DG, Christakis NA. Inequality and visibility of wealth in experimental social networks. *Nature.* 2015;526(7573):426-9.
38. Baldassarri D, Grossman G. Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences.* 2011;108(27):11023-7.
39. Gülerk Ö, Irlenbusch B, Rockenbach B. The Competitive Advantage of Sanctioning Institutions. *Science.* 2006;312(5770):108-11.
40. Nishi A, German CA, Iwamoto SK, Christakis NA. Status invisibility alleviates the economic gradient in happiness in social network experiments. *Nature Mental Health.* 2023.
41. Wolff I. Retaliation and the role for punishment in the evolution of cooperation. *J Theor Biol.* 2012;315:128-38.
42. Rendell L, Boyd R, Cownden D, Enquist M, Eriksson K, Feldman MW, et al. Why Copy Others? Insights from the Social Learning Strategies Tournament. *Science.* 2010;328(5975):208-13.
43. van Schaik CP, Burkart JM. Social learning and evolution: the cultural intelligence hypothesis. *Philos Trans R Soc Lond B Biol Sci.* 2011;366(1567):1008-16.
44. Dreber A, Rand DG. Retaliation and antisocial punishment are overlooked in many theoretical models as well as behavioral experiments. *Behavioral and Brain Sciences.* 2012;35(1):24-.

45. Nikiforakis N. Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*. 2008;92(1):91-112.
46. Janssen MA, Bushman C. Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology*. 2008;254(3):541-5.
47. Boyd R, Richerson PJ. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*. 1992;13(3):171-95.
48. Boyd R, Gintis H, Bowles S, Richerson PJ. The evolution of altruistic punishment. *Proc Natl Acad Sci U S A*. 2003;100(6):3531-5.
49. Bone JE, Raihani NJ. Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior*. 2015;36(4):323-30.
50. R Core Team. R: A language and environment for statistical computing. 4.3.0 ed. Vienna, Austria: R Foundation for Statistical Computing; 2023.
51. Rand DG, Greene JD, Nowak MA. Spontaneous giving and calculated greed. *Nature*. 2012;489(7416):427-30.
52. Bear A, Rand DG. Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*. 2016;113(4):936-41.
53. Everett JAC, Ingbreetsen Z, Cushman F, Cikara M. Deliberation erodes cooperative behavior — Even towards competitive out-groups, even when using a control condition, and even when eliminating selection bias. *Journal of Experimental Social Psychology*. 2017;73:76-81.
54. Rand DG. Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation. *Psychol Sci*. 2016;27(9):1192-206.
55. Kvarven A, Strømmland E, Wollbrant C, Andersson D, Johannesson M, Tinghög G, et al. The intuitive cooperation hypothesis revisited: a meta-analytic examination of effect size and between-study heterogeneity. *Journal of the Economic Science Association*. 2020;6(1):26-42.
56. Verkoeijen PPJL, Bouwmeester S. Does Intuition Cause Cooperation? *PLOS ONE*. 2014;9(5):e96654.
57. Castro Santa J, Exadaktylos F, Soto-Faraco S. Beliefs about others' intentions determine whether cooperation is the faster choice. *Scientific Reports*. 2018;8(1):7509.
58. Lie-Panis J, André J-B. Cooperation as a signal of time preferences. *Proceedings of the Royal Society B: Biological Sciences*. 2022;289(1973):20212266.

59. Nava F, Margoni F, Herath N, Nava E. Age-dependent changes in intuitive and deliberative cooperation. *Scientific Reports*. 2023;13(1):4457.
60. Maeda K, Kumai Y, Hashimoto H. Potential influence of decision time on punishment behavior and its evaluation. *Front Psychol*. 2022;13:794953.

Chapter 3: The Role of Network Size in Intervention Planning for Increasing the Lung Cancer Screening Rate: An Agent-based Modeling Study

3.1 Abstract

Low-dose computed tomography (LDCT) has been the recommended screening test for lung cancer by both the Centers for Disease Control and Prevention and the United States Preventive Services Task Force since 2013. However, as of 2021, only 5.8% of eligible individuals in the United States have received LDCT screening. To find methods to address this low adoption rate, we developed an agent-based simulation framework to model a network intervention program that promotes observational learning between peers which would increase the LDCT screening rate in a community of individuals susceptible to lung cancer. Using this framework, we evaluated if seeding the network with agents who adopt LDCT screening behavior before intervention initiation or increasing the number of agents in the intervention population could increase the screening rate beyond the rate achieved by the intervention alone. Results show that while seeding the network with adopter agents does not increase the screening rate, increasing the number of agents in the network does increase the screening rate until the network size becomes sufficiently large. In conclusion, network intervention planners should consider expanding the size of their intervention programs to secure a strong enough induction force to generate a diffusion cascade of the intervention behavior.

3.2 Introduction

Annual lung cancer screening using low-dose computed tomography (LDCT) has been recommended since 2013 for individuals aged 55 to 74 with 30 or more pack-years of smoking

history who currently smoked or had quit smoking in the last 15 years, with a recent shift to a recommendation for individuals aged 50 to 80 with a 20 pack-year smoking history (1, 2). While there is a substantial range of clinical evidence (3-7) for the effectiveness of LDCT in increasing the detection of lung cancers and reducing mortality, adoption of LDCT screening among eligible individuals in the United States remains low. As of 2021, only 5.8% of an estimated 14.2 million eligible patients have received LDCT under CDC and US Preventive Task Force recommendations (1, 8). Furthermore, LDCT screening rates vary widely across geographic regions, ranging from less than 1% in California to over 16% in Massachusetts (9, 10).

There is a sharp contrast between the rapid diffusion of other preventive behaviors such as physical distancing (11, 12) and mask-wearing (13) during the COVID-19 pandemic and the slow diffusion of LDCT screening. Two factors that complement the urgency associated with the rapid onset of COVID-19 can explain why diffusion of LDCT screening is slow compared to the spread of physical distancing and mask-wearing. First, since the proportion of individuals in the US eligible for LDCT is low (4.2% (9)) and the current screening rate is low (5.8% (9)), it is unlikely that eligible individuals would interact with each other without intervention. Conversely, both physical distancing and mask-wearing were widely prevalent during the COVID-19 pandemic since almost all individuals were susceptible to COVID-19 infection. Second, LDCT screening behavior is not public; people do not know if friends or family members in their social network are interested in LDCT screening or receive the screening unless the topic is discussed. In contrast, both physical distancing and mask-wearing are public behaviors whose adoption did not need to be explicitly shared with friends or family to be known.

We therefore developed a simulation framework using a social network of eligible individuals in which an intervention displayed LDCT screening behavior publicly and shared the behavior with other eligible individuals. This network intervention design falls into the induction approach from Valente's network interventions framework (14). The induction approach aims to motivate individuals to implicitly or explicitly interact with others in their social network to promote the diffusion of new or beneficial behaviors, information, or products (15, 16). These interventions have been proven to be effective in various domains, including the improvement of social and emotional well-being (17), academic citation behavior (18), and job-seeking (19-21). As people tend to use observational learning (22) – observing, analyzing, and adopting the behaviors of their network connections to modify their own behavior, it should follow that network interventions that motivate individuals to take advantage of observational learning (for example, asking people to share their voting behavior with their friends and acquaintances (23) should accelerate the spread of the intervention behavior. This is supported by a recent systematic review (24) which identified 19 intervention programs using the induction approach to generate influence or elicit behavioral change in targeted social networks, with varying pre-intervention rates of the target behavior (median: 38.5%, range: 1–83%) and varying sizes of communities of interest (median: 854, range: 89–9,042).

Induction-based interventions have demonstrated their ability to promote various behaviors in different socioeconomic and cultural contexts (25-28). However, planning these interventions is difficult when the pre-intervention rate of the target behavior is low. In such a scenario, the number of individuals in the social network who adopt the intervention behavior is insufficient to promote a diffusion cascade of the intervention behavior. When this is the case (e.g., in the case of LDCT where the pre-intervention rate of the behavior is approximately

5.8%), observational learning is unlikely to occur because sources from which LDCT screening behavior can be observed are uncommon. We therefore assessed two potential solutions to resolve this issue in the network intervention planning stage: 1) proactively asking a percentage of study participants to adopt the target behavior as seed individuals (forced early adopters) to boost its diffusion, and 2) expanding the size of the social network in which intervention participants share information about their behavioral choices.

In the present study, we focused on increasing the LDCT screening rate in the US. Using agent-based simulations embedding network interventions via the induction approach, we aimed to examine the consequences of seeding the network with agents who adopt the LDCT screening behavior before the intervention initiation and expanding the total number of study participants. We initially hypothesized that the post-intervention screening rate would be lower at lower levels of initial seeding of the target behavior and in networks representing smaller study participant sizes.

3.3 Methods

To develop our agent-based simulation framework, we first identified and calibrated a set of simulation parameters (the *control* setting) that would generate results which approximated the currently low rate of LDCT screening in the United States of 5.8%. We used 1,000 iterations of the simulation under the control setting to model a network of 200 agents over 40 quarters (or 10 years) in which agents behaved without being influenced by other agents. This setting confirmed that in the absence of any network intervention, the screening rate in our simulated networks remained constant at or near the current LDCT screening rate. Next, we developed three different simulation settings that included the induction-based intervention and observed

how the LDCT screening rate changed in comparison to the control setting. In the *observational learning* setting, agents were allowed to share information with each other through their strong and weak network ties to influence each other agents' behavior. In the *adding seeds* setting, we selected a percentage of agents to adopt the intervention behavior before introducing the intervention to the network. Lastly, in the *modifying network size* setting, we changed the size of the network to be smaller or larger than the default value of 200 agents and deployed the intervention (without adding seeds). We analyzed the outcomes of each simulation setting by calculating the median screening rate for each year of observation and the corresponding 95% quartile ranges. We also evaluated if each setting led to diffusion of LDCT screening behavior by comparing the screening rate in year 1 of the simulations and the screening rate in year 10 of the simulations.

1. *Setup of the agent-based simulations (Control setting)*

Agents. We constructed an agent-based simulation framework that models a human social network in which the target behavior rate approximates the current rate of LDCT screening in the United States. Agents in the simulation represent individuals eligible for annual LDCT screening who are recruited into an intervention that aims to increase the LDCT screening rate in the agents' community. For reference, the US Preventive Services Task Force recommends "annual screening for lung cancer with low dose computed tomography (LDCT) in adults aged 50 to 80 years who have at least a 20 pack-year smoking history" (1). We assumed that all agents had the same sociodemographic characteristics and that individual agents were permitted to interact with any other agent in the network.

We conceptualized the simulation as a scenario in which we work with a community intervention planner who manages the health of a community with a population size approximating the size of an average US-based elementary school zone. Based on the current population of the US (335 million as of January 2024 (29) and the number of public elementary schools (approximately 70,000 as of 2021) (30), one school zone is estimated to have 4,750 individuals. Since the number of LDCT-eligible individuals in the US is estimated to be 14.2 million (9) (4.2% of the US population), an elementary school zone, on average, is estimated to include approximately $4,750 \times 0.042 = 200$ eligible individuals. We therefore used 200 individuals as the default value for parameter 1 [P1] of the simulations.

Environment (network structure). We placed the 200 agents into a single social network (i.e., the number of networks = 1 [P2]), meaning that none of the agents were isolated and each agent was on at least one path with another agent in the network. Agents were weakly linked to all other agents (weak ties) such that each agent had a degree (the number of network ties per agent) of 199. In this manner, we assumed that each agent had some level of knowledge of the behavior of all other agents, for example, through word-of-mouth, engagement in the community, or via social media/the internet, and thus was only weakly influenced by any specific agents' behavior.

We also modeled influence on agents from close friends and family members (strong ties) using the Watts-Strogatz small-world model (31), using an initial neighborhood size (the number of neighbors in the base lattice structure) of 4 [P3] and a rewiring probability of 0.2 [P4]. In this arrangement, each agent had 4 neighbors in a base circle lattice structure, which is then rewired, resulting in variations in the number of neighbors and reducing the average network distance between any two agents in the network. In real-world terms, this value suggests that each agent

in our network of eligible individuals had 4 close friends or family members who were also eligible to be screened. While this may suggest strong homophily between eligible participants, past research has shown that individuals considering cancer screening or individuals at higher risk for cancer strongly consider information from homophilic network ties when making health decisions (32, 33). Agents connected by strong ties were not simultaneously connected with a weak tie to avoid overlaps. In summary, each agent had a different number of weak and strong ties that exerted peer influence within networks of 200 agents. One network was prepared for each iteration of the simulations.

Behavior. Each simulation progressed over a period of 10 years, divided into 4 quarters each (a total of 40 quarters [Quarter 1 to 40]), with each quarter making up a single time step in the simulation). We run the simulation for 12 additional quarters (3 years) prior to observation to allow the simulation's behavior to stabilize. In the control setting, agents' behavior in each time step was programmed to only be influenced by the agents' own decision-making. The control setting represents the current situation of lung cancer screening where scheduling appointments for screening and actively going to a doctor's office or clinic for screening is driven solely by personal intentions (without observational learning or peer influence).

Agents were programmed to schedule LDCT screening every quarter if they had not previously been screened. Agents who received the screening were prevented from scheduling new appointments for 3 quarters after their last screening. To inform the initial set of parameters for the simulation, we assumed a baseline quarterly rate of appointment (the rate at which individuals, with no influence from network ties or from intervention, choose to voluntarily schedule an LDCT appointment) of 2.0% [P5], a conversion rate from scheduling a screening

appointment to going to the scheduled appointment of 75% [P6], and that when an appointment is missed, no follow-up scheduling of the appointment occurs for any following quarters. Given these parameters, we hypothesized that the median rate of LDCT screening in our simulated population would stabilize at approximately the rate of lung cancer screening in the US of 5.8% (as $[1 - (1 - (0.02 \times 0.75))^4] \approx 0.058$).

Our model also considered changes in behavior associated with waning interest in screening; agents who have been screened in the last four quarters (i.e., screened in the last year) have their chance to make an appointment set to 55% [P7]. We used this value to estimate the effect of adherence to screening guidelines once agents get screening in the year prior; we based this value on the estimated rate of LDCT adherence as reported by Lopez-Olivo and colleagues (34). Agents who did not have a screening in the last year have their chance of scheduling an LDCT appointment reduced accordingly [P8] (see **Supplementary Information** for details on the value of P8, which is not manually manipulated but is calculated as a function of P7). Based on these conditions and our assumptions based on the reported rate of LDCT in the US, we aimed to model no progress in the diffusion of LDCT screening over 10 years of the simulation under the control setting and a constant LDCT screening rate of around 5.8%.

Calibration. Using the parameters described above (P1 – P8), we calibrated our baseline simulation setting to approximate the LDCT screening rate of 5.8% by identifying sets of parameters that resulted in simulations that achieved median screening rates over 90% of the 40 quarters of the simulation of $5.8 \pm 0.5\%$ (5.3% – 6.3%). Our main target for calibration was the baseline appointment rate (P5) We selected a range of possible values for P5 centered around 2.0% and conducted 1,000 simulations for each value of P5. We varied P5 from 0.5% to 3.5% in

increments of 0.1% (a total of 30 values) and compared the performance of each set of 1,000 simulations. Based on our hypothesis for a constant LDCT rate in the control setting, we predicted the correctly calibrated simulation should result in approximately $200 \times 10 \times 0.058 = 116$ instances of LDCT screening over the 10 years of the simulation. Using this criterion, we identified that a P5 value of 3.4% resulted in simulations with a median of 116 instances of screening (95% quantile range for median: 115-117, **Supplementary Table S3.1**).

2. *Adding Observational Learning to the Simulations (Observational learning setting)*

In simulation setting 2 (the *observational learning* setting), we modeled an intervention where agents in the same social network are notified of others' LDCT screening behavior in the past quarter at the start of every new quarter using a peer-to-peer communication system representing a social media platform and/or text messaging. The intervention encourages agents to use observational learning to adopt a new behavior based on information learned from strong or weak ties. In this setting, agents have a 200% increased chance [P9] to schedule an LDCT screening when a strong tie schedules a screening in the prior quarter. For example, given the calibrated appointment rate of 3.4%, if one strong tie of a focal agent schedules LDCT screening in Quarter t , the chance of the focal agent to schedule LDCT screening in Quarter $t+1$ is boosted from 3.4% to 10.2% (since the context-dependent rate is equal to the baseline rate of making an LDCT screening appointment multiplied by the boost from the strong tie (200%); $0.034 \times (1+2) = 10.2\%$). Similarly, a focal agent who received screening in the last year whose appointment rate was set to 24.15% by P7 would have their appointment rate in Quarter $t+1$ boosted from 24.15% to 72.45% by a single strong tie who also schedules LDCT in Quarter t . If multiple strong ties schedule LDCT within the same time step, the aggregate effect is constrained by a

90% discount rate [P10]; for example, given a focal agent with four strong ties in which all four tie agents receive LDCT screening in Quarter t , the probability the focal agent schedules an LDCT screening is boosted from 3.4% to 20.7% ($0.034 \times (1 + 2 + 2 \times 0.9 + 2 \times 0.9^2 + 2 \times 0.9^3) = 20.7\%$). Under this intervention setting, agents are also notified when their weak ties schedule LDCT; weak ties contribute a 50% increased chance [P11] to schedule LDCT (a quarter of the effect of strong ties). This effect of weak ties is similarly discounted at a rate of 90% [P12].

3. *Addressing the Challenge of Observational Learning in Two Ways*

We anticipated that the observational learning setting would result in a higher LDCT screening rate than the control setting over the course of the simulations; however, it is possible that observational learning alone would not promote rapid diffusion of the intervention behavior. One potential cause of this slow diffusion is a low occurrence of observational learning events: since the average number of strong ties (close friends or family members) is 4, most agents are not connected to other agents who have adopted LDCT screening at the start of the simulations (on average, each agent is expected to be connected to $4 \times 0.058 = 0.23$ agents who adopt the behavior at the start of the intervention). Furthermore, under the control setting, each network of 200 agents would only include on average $200 \times 0.058 \approx 12$ agents who adopt LDCT screening behavior. In this context, the number of weak ties per agent in this scenario may not be sufficient to generate a strong enough trend of behavioral change for us to observe.

To address this challenge, we explored two solutions that network intervention planners could apply. In the first solution, we modified the intervention program so that the intervention behavior was seeded in the community through outreach before the observation period began. In this setting, we randomly assign a percentage of agents [P13] to get screened prior to widespread

deployment of the intervention (i.e., force a certain percentage of agents to receive LDCT at Quarter 0), which enhances observational learning as soon as the intervention period begins (“LDCT screening seeds”). Just as in the observational learning setting, we measured the LDCT screening rate at the end of each simulation and the median screening rate over the course of the simulations. This method combines the induction approach utilized in the observational learning setting with the individuals approach of Valente’s network intervention framework (14), which identifies important individuals in social networks to use as facilitators of behavioral change. In the second solution, instead of adding seeds prior to the interventions, we modeled scenarios where network intervention planners modified the capacity of the intervention to include different sizes of networks, ranging from networks of size 50 to networks of size 5,000. In this setting, the social networks for the intervention are modified to include fewer or more weak ties compared to the control setting depending on the size of the network being evaluated, while the number of strong ties remains the same. We use this setting to observe the relationship between network size and the LDCT screening rate, again measuring LDCT screening rate at the end of each simulation period and tracking the median LDCT screening rate over time.

4. *Quantitative Analysis of Simulation Results*

We ran 1,000 iterations of the simulation for each set of simulation parameters. For each simulation, we calculated the median screening rate of agents in the simulations across each quarter of the simulation (Quarters 1 – 40) and the 95% quartile ranges for the median screening rate in each year. We compared the effects of the three intervention settings (simulation settings 2, 3, and 4) to the control setting by comparing the end-of-simulation screening rates and their corresponding quartile ranges.

5. Robustness tests

We also conducted a series of robustness tests to evaluate the effects of modifying individual simulation parameters under the observational learning setting. To do so, we carried out 1,000 iterations of the simulation for discrete values in the parameter ranges listed in **Table 3.1** while keeping the values of the other parameters constant (equal to those used in the setup for simulation setting 2). We then recorded the median screening rate for each year for each parameter value and compared the effects of adjusting each parameter's value on the end-of-simulation screening rate.

3.4 Results

Observational Learning Setting vs. Control Setting. We found that in the observational learning setting (simulation setting 2), the median screening rate increased from 21.5% in year 1 (95% quantile range [QR]: 21.0 – 22.0%) to 51.0% in year 10 (95% QR: 50.5% – 51.0%). There was no increase in the control setting (setting 1) (5.5% in year 1 and year 10 for the control setting) (**Figures 3.1A and 3.1B**). This corresponds to the observational learning setting influencing 91 additional agents to get LDCT screening at year 10 compared to the control setting ($200 \times 0.51 = 102$ agents for the observational learning setting vs. $200 \times 0.055 = 11$ agents for the control setting). We can also compare the total number of screening events across the 10 years of the simulation for the observational learning setting and the control setting; using this comparison, we see that the observational learning setting results in approximately 1,119 screening events compared to only 110 screening events for the control setting. We observed a steep increase in

the LDCT screening rate in the first three years after the introduction of the intervention, followed by smaller increases in screening rate over the remainder of the simulations.

Adding seeds (Solution 1). We observe that changing the percentage of seeding agents shifts the median screening rate at the start of the simulations but does not affect the median screening rate achieved in year 10. We found no difference in year 10 screening rate comparing simulations in networks where there were no LDCT screening agents (median rate = 51.0%, 95% QR: 50.5 – 51.5%) (**Fig. 3.2A and 3.2B, dark green**) and in networks where 50% of agents were assigned to be LDCT screening agents (median rate = 51.0%, 95% QR: 50.5% – 51.5%) (**Fig. 3.2A and 3.2B, gold**).

Modifying network size (Solution 2). From simulations evaluating different network sizes, we observe that the largest increase in the year 10 screening rate occurs when increasing the network size from 50 agents to 100 agents (from 35.0% to 45.0%). We also observe that each subsequent addition of 50 agents results in a smaller net increase in year 10 median screening rate (**Figure 3.3B**), with the increases being minimized after the network size exceeds approximately 1,000 agents.

Robustness tests. In our robustness tests, we found that increases in the Watts-Strogatz neighborhood size, observational learning strength (for both strong and weak ties) and discount rate (again, for both strong and weak ties) corresponded to increases in the year 10 median screening rate. These results are consistent with the roles of each of these parameters in the simulation; by increasing the neighborhood size, we increase the number of strong ties each

agent has and therefore increase the chances that agents will use observational learning and adopt LDCT screening behavior. Similarly, increasing the strength of the intervention and increasing the discount rate (thus increasing the added effect of each additional tie who adopts the intervention behavior) should correspond to more agents adopting LDCT screening behavior. These robustness tests did not substantially change our major results and interpretations.

3.5 Discussion

In this study, we used agent-based simulations to demonstrate that a network intervention can successfully increase the rate of lung cancer screening in a community of eligible individuals. We also tested two methods (adding seed individuals and modifying the network size) that could theoretically increase the screening rate above the rate resulting from the intervention alone. We found that increasing the network size in which the intervention is deployed improves the increase in screening rate generated by the intervention. However, we also found that seeding the intervention behavior prior to the simulation period results in different starting levels of the screening rate but does not result in higher screening rates at the end of the simulations.

Using our agent-based simulation framework, we identified an intervention scenario that would increase the median LDCT screening rate in our simulated community above the national average LDCT rate of 5.8% to approximately 51.0% after 10 years of intervention. Ideally, public health professionals planning an LDCT screening intervention would want all the intervention participants to adopt and maintain a regular screening regimen. This aligns with the idea that through observational learning, all network members would eventually be exposed to the intervention behavior and be more likely to adopt it. However, we observed that the

screening rate in our simulations which included the intervention increased rapidly over the first few years of the intervention program but increased at a much slower rate once approximately 45.0% of the agents were getting LDCT screening every year.

One plausible explanation for this slowing down can be attributed to the behavioral logic of agents in our simulation. Agents were programmed to reduce their chances to get LDCT screening if they did not receive the screening in the past year to reflect barriers against screening behavior (for example, cost of the procedure or lack of awareness leading to low interest (35, 36)) but increased their chances to get LDCT screening if they did receive the screening in the past year. This dichotomy would likely lead to two sub-sections of the network developing over time – a section of the network who get regular screening and maintain high interest and willingness to get screened and another section that is entrenched in their refusal. As a result of the relatively small size of the network, it is possible that groups of strong ties could become static in their makeup, with clusters of willing screeners in one group and clusters of screening refusers in the other group, preventing the overall screening rate from exceeding the 50% mark we observed in simulation setting 2.

Our investigations into adjustments to the intervention simulation set-up that could increase the screening rate past the rate achieved by the intervention alone were somewhat fruitful. We hypothesized that simulation setting 3 (adding seeding agents) could increase the screening rate boost generated by the intervention by using the LDCT screening seeds as influential agents to start observational learning earlier in the simulation, resulting in more gains in screening rate at the end of the simulation period. However, we found that the year 10 screening rates were similar across all tested values of the proportion of LDCT screening seeds. This result suggests that in networks where most individuals are highly connected, it is difficult

to identify key network nodes from which diffusion cascades can spread. It would thus be reasonable for intervention planners to consider the layout of their network of interest before deploying interventions that rely on the network structure to take hold. This result also encourages careful consideration of the costs and benefits of utilizing individuals within the network compared to induction-based approaches in which only a few individuals need to adopt the intervention behavior to begin a diffusion cascade.

We also expected that simulation setting 4 (changing network size) could increase the screening rate past the rate observed in simulation setting 2. We obtained two main findings. First, even under the smallest tested network size ($N = 50$), the year 10 screening rate exceeded the year 10 screening rate of the control setting. Second, we identified that gains in median screening rate associated with increasing the network size diminish as the network size increases. Since we observed that the intervention increased the median screening rate in a small network of 50 agents and that larger networks experienced a greater increase in median screening rate until the network size reaches approximately 1,000 agents, we can be relatively confident that our intervention was successfully implemented in the simulations. Our observation of a curve (**Fig. 3.3B**) that could be used by intervention planners to perform calculations akin to the power calculations conducted before experimental studies to determine if the network intervention can successfully diffuse shows promise. Future work should evaluate different types of interventions and intervention characteristics which could be used to develop a series of formulas that could be used for network intervention planning calculations.

Our study has several limitations. Like all agent-based simulation studies, we make several assumptions regarding the behavior and environment of our agents to mimic the real-world scenario of eligible patients receiving exposure to LDCT screening from friends and

acquaintances. We especially must approximate the “strength” of the intervention to convert individuals resistant to screening into those willing to make clinic appointments to eventually receive LDCT screening. Underestimating the strength of the intervention and requiring more network neighbors to adopt the behavior before observational learning can take hold would result in slower diffusion of the screening rate and a need for a longer period of observation. As seen in the robustness tests (**Supplementary Fig. S3.1**), we observe that at lower levels of observational learning strength (for both strong ties and weak ties), the median screening rate at year 10 does not reach the 51% figure we identified in simulation setting 2. We also observe that increases in appointment rate past the 10% mark do not correlate with increases in the year 10 median screening rate, suggesting that diminishing returns should be a consistent concern when planning network interventions.

Additionally, although we designed our simulations to include agents were programmed to reduce their chances to get the screening if they had not been screened in the last year, prior research suggests that participants who are already willing to get cancer screening such as LDCT screening may be more willing than random members of the population to adhere to guidance from clinicians or other health professionals about health behaviors or lifestyle (37, 38). Finally, we do not model sociodemographic factors or intervention-related factors that could change the behavior of intervention participants in real-world scenarios such as duration of the intervention program or required frequency of interacting with research staff that could limit individuals’ willingness to participate in an intervention program, interact with healthcare professionals or clinical staff, or listen to advice from friends, family members, or other types of acquaintances to change their own behavior (39-41).

3.6 Figures and Tables

Parameter		Calibrated Values (Range for Robustness Tests)
<i>Agents</i>		
P1	Number of LDCT-eligible agents in a target community	200 (50 – 5,000)
<i>Environment (network structure)</i>		
P2	Number of networks	1
P3	Watts-Strogatz neighborhood size	4 (2 – 30)
P4	Watts-Strogatz rewiring probability	0.2 (0.1 – 0.4)
<i>Behavior</i>		
P5	Rate of appointment	3.4% (1 – 20%)
P6	Conversion rate	75% (50 – 90%)
P7	Chance of screening after being screened in last year*	24.15% (10 – 80%)
P8	Chance of screening after missing screening in past year	Varies as a function of P7
<i>Intervention</i>		
P9	Observational learning effect – strong ties	2 (0.5 – 4)
P10	Discount rate for strong ties	0.9 (0 – 1)
P11	Observational learning effect – weak ties	0.5 (0.125 – 0.75)
P12	Discount rate for weak ties	0.9 (0 – 1)
P13*	Percentage of LDCT screening seed agents	0 (0 – 50%)

Table 3.1. The 13 parameters manipulated in the agent-based simulations. Parameters listed under the Agents, Environment, and Behavior headings are used in the control setting and the three settings that include the intervention. Values and ranges were determined using literature review of cancer-related and health or economics-based behavioral research involving diffusion of innovations as well as discussion with clinicians specializing in lung cancer screening. *P13 is only used in simulation setting 3.

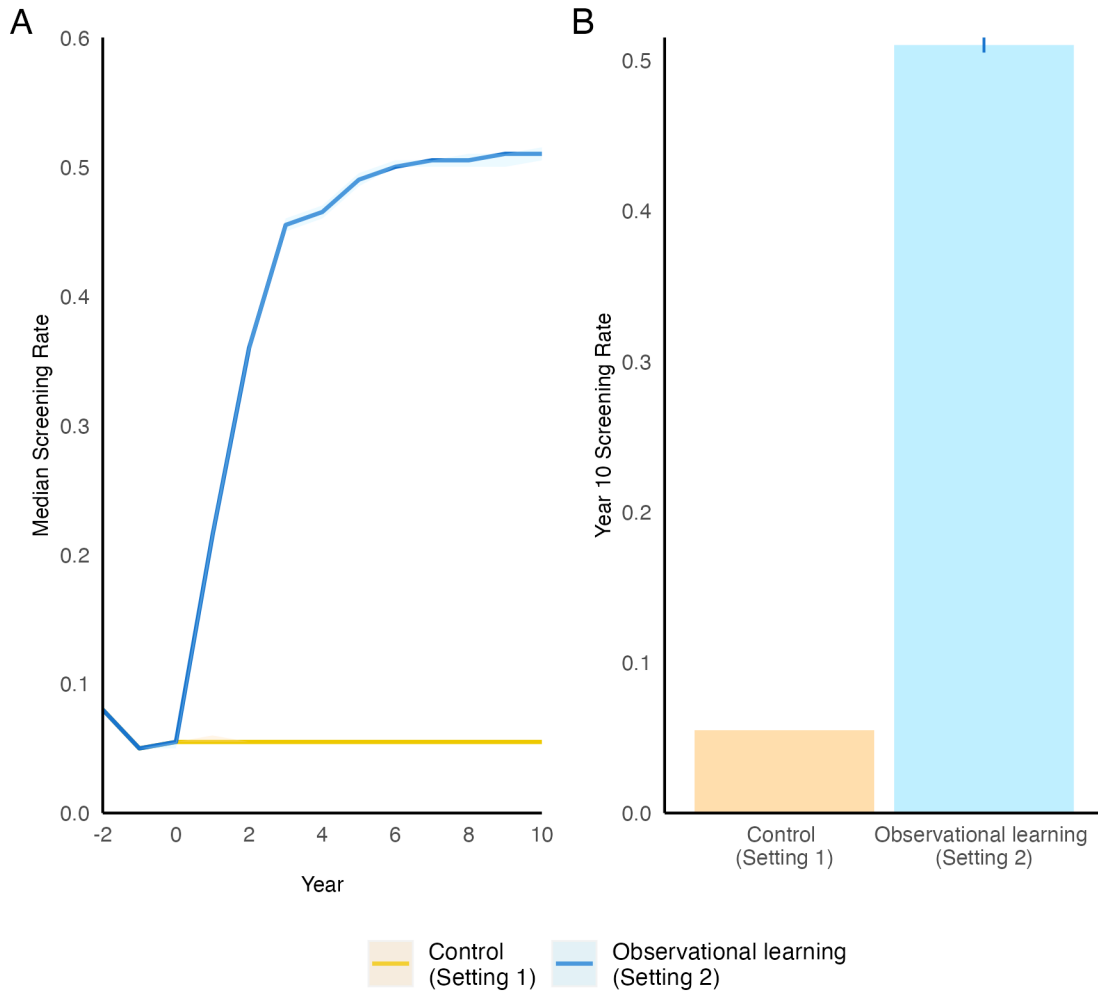


Figure 3.1. The observational learning setting (blue) successfully encourages diffusion of LDCT screening behavior compared to the control setting (gold). **A:** Solid lines represent the median screening rate across 1,000 simulations for each year of the simulation, while shaded regions represent 95% quantile ranges of the median screening rate. In the observational learning setting, we observe an increase in median screening rate of 29.5% from year 1 to year 10 (from 21.5% to 51.0%); there was no increase in the median screening rate under the control setting (5.5% in both year 1 and year 10). Simulations were run for an additional 12 quarters (shown in the figure as years -2, -1, and 0) prior to intervention deployment to allow simulations to stabilize. **B:** Comparison of the year 10 screening rate for the control setting and the observational learning setting. Error bars indicate the 95% quantile range of the year 10 screening rate.

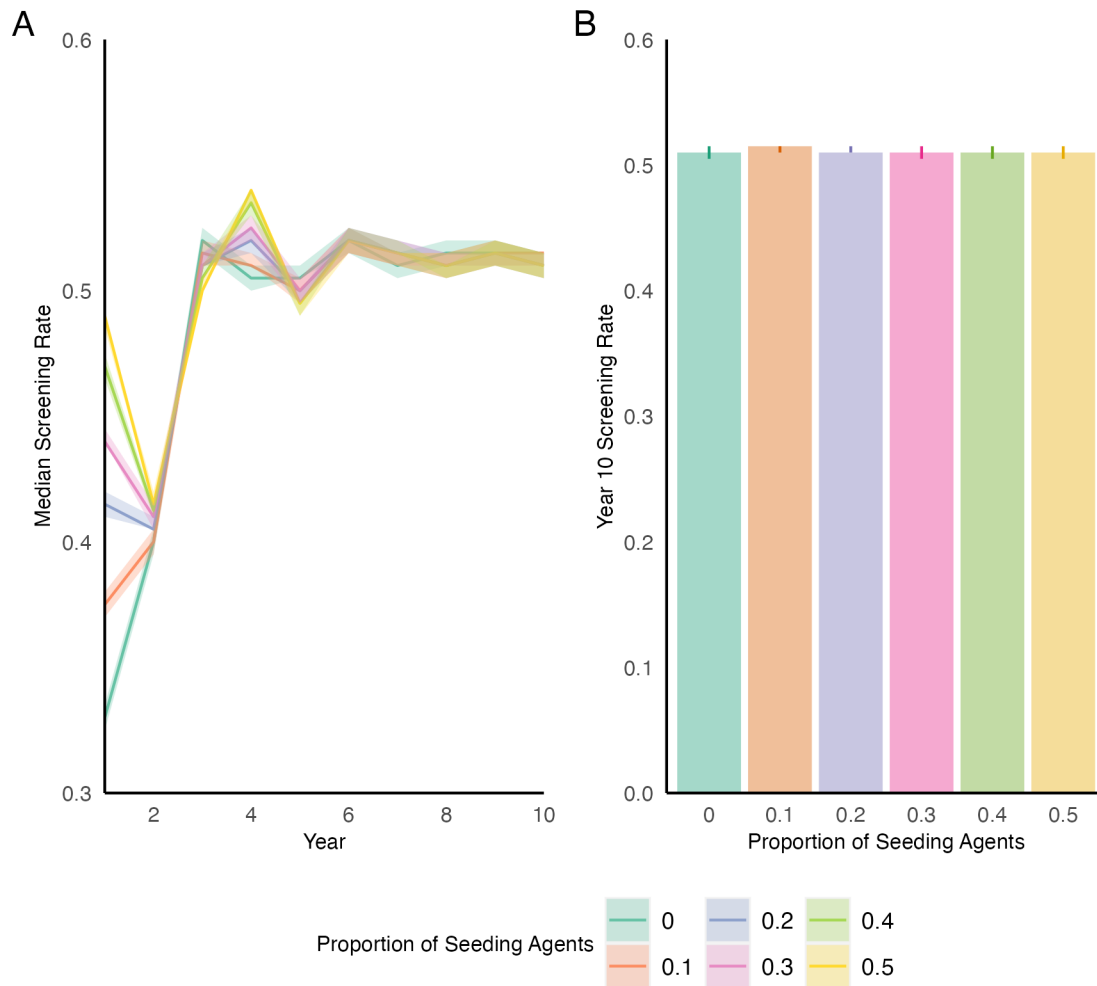


Figure 3.2. Seeding the network with agents who have previously been screened does not result in increased diffusion of the intervention. A: The overlap in lines representing median screening rates for different proportions of seeding agents suggests that increasing the proportion of seeding agents does not spread the intervention behavior to more agents compared to deploying the intervention alone. Lines indicate median screening rates for 1,000 simulations while shaded regions represent 95% quantile ranges of the median screening rate. **B:** Year 10 screening rates for different proportions of seeding agents. Error bars represent 95% quantile ranges of the median screening rate in year 10 of the simulations.

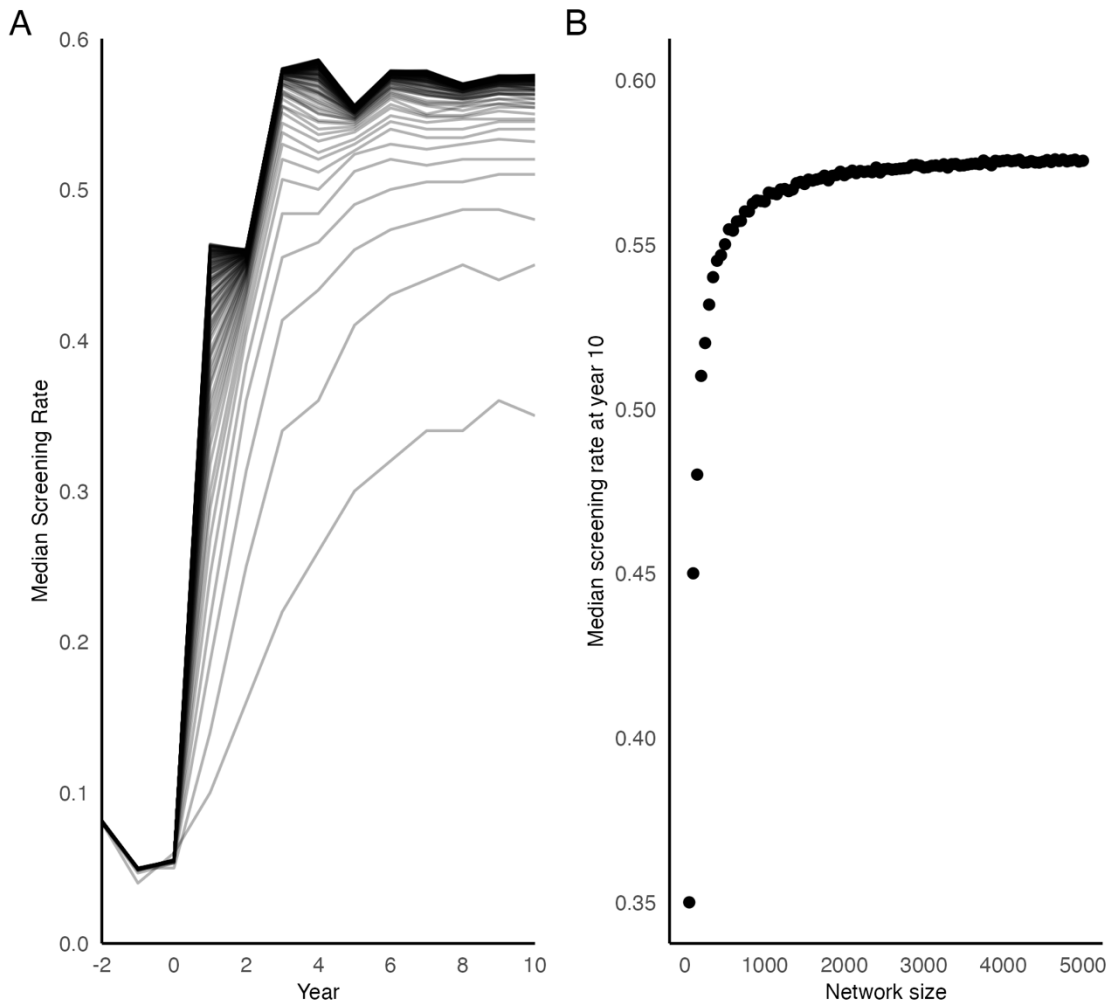


Figure 3.3. Increasing the network size results in increased diffusion of the intervention, but with diminishing returns. A: Median screening rates for 1,000 simulations with different network sizes. Lines represent median screening rates for different network sizes. Overlapping lines result in darker regions. **B.** Scatterplot demonstrating the correlation between network size and the year 10 median screening rate. We observe that as the network size increases, the corresponding increase in the year 10 screening rate decreases, with only a 1.0% difference in screening rate comparing networks of size 1,000 (median rate = 0.565, 95% QR: 0.564 – 0.566) and networks of size 5,000 (median rate = 0.5754, 95% QR: 0.5750 – 0.5762).

3.7 Supplementary Information

Calculating the percent decrease in the appointment rate of agents who were not screened in the previous year

Our simulation's control setting aims to replicate the average rate of LDCT screening in the United States of 5.8%. To allow the model to account for differential interest in screening, we program the simulation to adjust the rate at which agents make appointments for future screening based on their screening status in the past year.

Based on literature evidence (34, 42) and UCLA Health data, we set the overall adherence rate of agents screened in the last year to 55%, which corresponds to a P7 value of 24.15% (as $1 - (1 - (0.2415 \times 0.75))^4 \approx 0.55$). We accordingly modify the appointment rate of agents who were not screened in the prior year to model the effect of decreasing interest in the intervention behavior among agents who do not interact with the screening program while restricting the overall screening rate to remain at 5.8%. We demonstrate the calculation for P8, the appointment rate for agents not screened in the last year, below:

$$58[1 - (1 - (1 - [P7] \times [P6])^4)] + 942[1 - (1 - [P8] \times [P6])^4] = 58$$

$$58 \times 0.55 + 942[1 - (1 - [P8] \times 0.75)^4] = 58$$

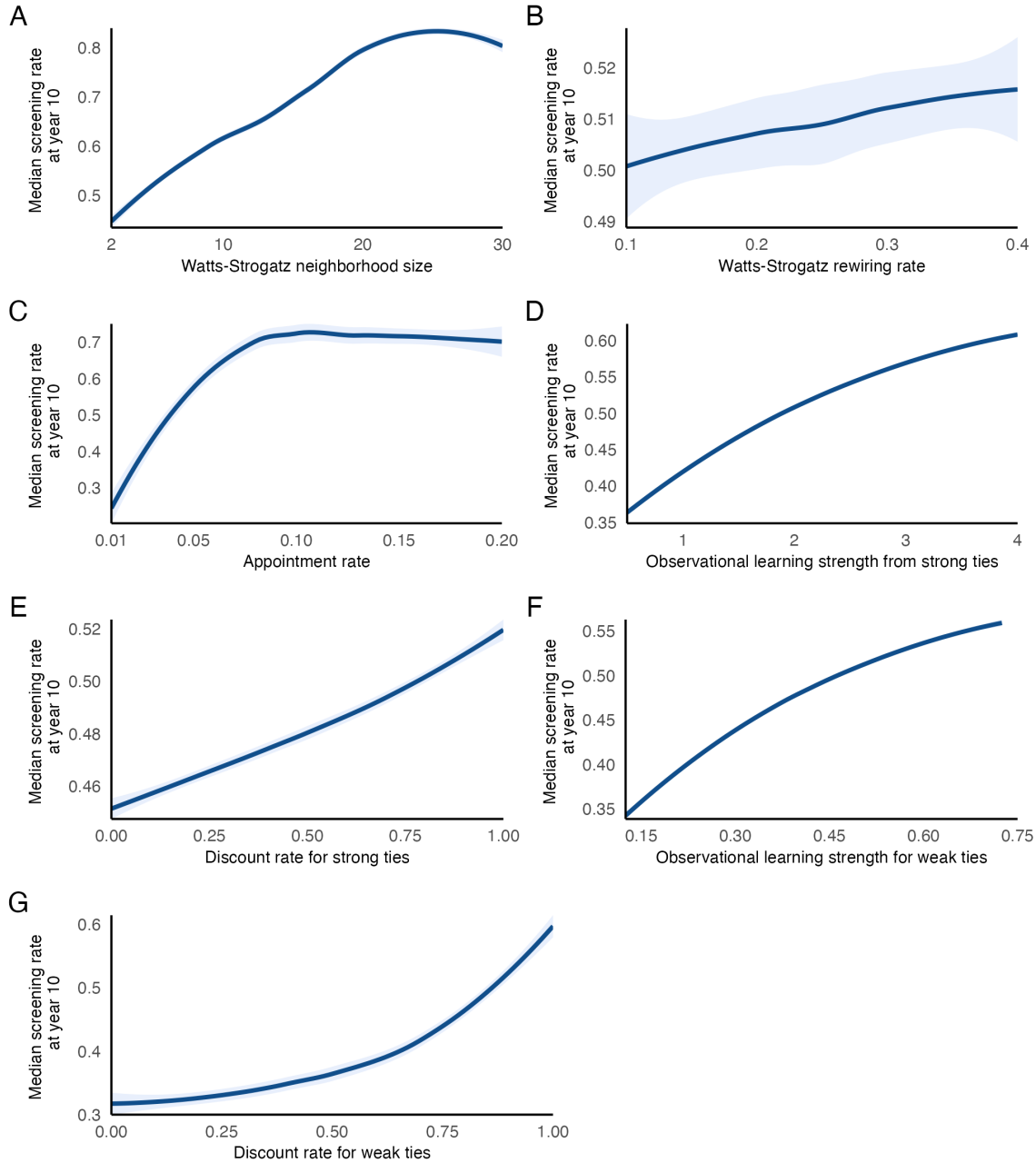
$$31.9 + 942[1 - (1 - 0.75 \times [P8])^4] = 58$$

$$1 - (1 - 0.75 \times [P8])^4 = 0.0277$$

$$P8 = 0.009$$

Appointment Rate	Median Number of LDCT Screenings	95% Quartile Range for Median LDCT Screenings
0.030	112	111 – 113
0.031	114	113 – 115
0.032	114	113 – 115
0.033	115	113 – 116
0.034	116	115 – 117
0.035	117	115 – 118

Supplementary Table S3.1. Calibration results for the control setting simulations. Based on the search method described in the main text, we compared simulations using different values of the appointment rate [P5] by calculating the median number of LDCT screenings observed for each value. The grey highlighted row indicates that a simulation using an appointment rate of 3.4% resulted in our theorized value of 116 LDCT screenings (from $200 \times 10 \times 0.058$) with a tolerance rate of 0.05%.



Supplementary Figure S3.1. Results of robustness tests for the observational learning setting (simulation setting 2). Increasing the neighborhood size [P3] (A), observational learning strength for strong and weak ties [P9 and P11] (D, F), and the observational discount rate for strong and weak ties [P10 and P12] (E, G) are correlated with increases in year 10 median screening rate. We also observe that the highest year 10 screening rates correspond with raising the Watts-Strogatz network size to above 20 agents (A) and that increasing the appointment rate above 10% did not result in substantial increases in the year 10 median screening rate (C). Lines indicate median year 10 screening rates from 1,000 simulations for each setting; shaded regions indicate 95% quartile ranges from the median year 10 screening rate.

3.8 References

1. U.S. Preventive Services Task Force. Lung Cancer: Screening 2021 [Available from: <https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/lung-cancer-screening>].
2. McDowell S. New Lung Cancer Screening Guideline Increases Eligibility 2023 [Available from: <https://www.cancer.org/research/acs-research-news/new-lung-cancer-screening-guidelines-urge-more-to-get-ldct.html>].
3. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395-409.
4. De Koning H, Van Der Aalst C, Ten Haaf K, Oudkerk M. PL02. 05 effects of volume CT lung cancer screening: mortality results of the NELSON randomised-controlled population based trial. *Journal of Thoracic Oncology*. 2018;13(10):S185.
5. Becker N, Motsch E, Trotter A, Heussel CP, Dienemann H, Schnabel PA, et al. Lung cancer mortality reduction by LDCT screening—results from the randomized German LUSI trial. *International journal of cancer*. 2020;146(6):1503-13.
6. Field JK, Vulkan D, Davies MP, Baldwin DR, Brain KE, Devaraj A, et al. Lung cancer mortality reduction by LDCT screening: UKLS randomised trial results and international meta-analysis. *The lancet regional health–Europe*. 2021;10.
7. Pastorino U, Silva M, Sestini S, Sabia F, Boeri M, Cantarutti A, et al. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. *Ann Oncol*. 2019;30(7):1162-9.
8. American Lung Association. Lung Cancer Key Findings 2023 [Available from: <https://www.lung.org/research/state-of-lung-cancer/key-findings>].
9. American Lung Association. State of Lung Cancer 2022 Report. 2022.
10. Fedewa SA, Kazerooni EA, Studts JL, Smith RA, Bandi P, Sauer AG, et al. State Variation in Low-Dose Computed Tomography Scanning for Lung Cancer Screening in the United States. *JNCI: Journal of the National Cancer Institute*. 2020;113(8):1044-52.
11. Zang E, West J, Kim N, Pao C. U.S. regional differences in physical distancing: Evaluating racial and socioeconomic divides during the COVID-19 pandemic. *PLoS One*. 2021;16(11):e0259665.

12. Courtemanche C, Garuccio J, Le A, Pinkston J, Yelowitz A. Strong Social Distancing Measures In The United States Reduced The COVID-19 Growth Rate. *Health Affairs*. 2020;39(7):1237-46.
13. Fischer CB, Adrien N, Silguero JJ, Hopper JJ, Chowdhury AI, Werler MM. Mask adherence and rate of COVID-19 across the United States. *PLOS ONE*. 2021;16(4):e0249891.
14. Valente TW. Network interventions. *Science*. 2012;337(6090):49-53.
15. Zhang J, Centola D. Social Networks and Health: New Developments in Diffusion, Online and Offline. *Annual Review of Sociology*. 2019;45(1):91-109.
16. Valente TW, Vega Yon GG. Diffusion/Contagion Processes on Social Networks. *Health Education & Behavior*. 2020;47(2):235-48.
17. Sandstrom GM, Dunn EW. Social Interactions and Well-Being: The Surprising Power of Weak Ties. *Pers Soc Psychol Bull*. 2014;40(7):910-22.
18. Fronczak A, Mrowinski MJ, Fronczak P. Scientific success from the perspective of the strength of weak ties. *Sci Rep*. 2022;12(1):5074.
19. Granovetter MS. The Strength of Weak Ties. *American Journal of Sociology*. 1973;78(6):1360-80.
20. Krackhardt D. The Strength of Strong Ties : The Importance of Philos in Organizations. 2003 [cited 1/15/2024]. In: *Networks in the Knowledge Economy* [Internet]. Oxford University Press, [cited 1/15/2024]; [0]. Available from: <https://doi.org/10.1093/oso/9780195159509.003.0008>.
21. Granovetter M. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*. 1983;1:201-33.
22. Bandura A, Grusec JE, Menlove FL. Observational Learning as a Function of Symbolization and Incentive Set. *Child Development*. 1966;37(3):499-506.
23. Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, et al. A 61-million-person experiment in social influence and political mobilization. *Nature*. 2012;489(7415):295-8.
24. Hunter RF, de la Haye K, Murray JM, Badham J, Valente TW, Clarke M, et al. Social network interventions for health behaviours and outcomes: A systematic review and meta-analysis. *PLoS Med*. 2019;16(9):e1002890.
25. Latkin CA, Donnell D, Metzger D, Sherman S, Aramrattna A, Davis-Vogel A, et al. The efficacy of a network intervention to reduce HIV risk behaviors among drug users and risk partners in Chiang Mai, Thailand and Philadelphia, USA. *Soc Sci Med*. 2009;68(4):740-8.

26. Latkin CA. Outreach in natural settings: the use of peer leaders for HIV prevention among injecting drug users' networks. *Public Health Rep.* 1998;113 Suppl 1(Suppl 1):151-9.
27. Earp JA, Eng E, O'Malley MS, Altpeter M, Rauscher G, Mayne L, et al. Increasing Use of Mammography Among Older, Rural African American Women: Results From a Community Trial. *American Journal of Public Health.* 2002;92(4):646-54.
28. Bastian LA, Fish LJ, Peterson BL, Biddle AK, Garst J, Lyna P, et al. Proactive recruitment of cancer patients' social networks into a smoking cessation trial. *Contemp Clin Trials.* 2011;32(4):498-504.
29. United States Census Bureau. U.S. and World Population Clock 2024 [Available from: <https://www.census.gov/popclock/>].
30. National Center for Education Statistics. Fast Facts: Educational Institutions 2023 [Available from: <https://nces.ed.gov/fastfacts/display.asp?id=84>].
31. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature.* 1998;393(6684):440-2.
32. Flatt JD, Agimi Y, Albert SM. Homophily and health behavior in social networks of older adults. *Fam Community Health.* 2012;35(4):312-21.
33. Brown D, Oetzel J, Henderson A. Communication networks of men facing a diagnosis of prostate cancer. *J Clin Nurs.* 2016;25(21-22):3266-78.
34. Lopez-Olivo MA, Maki KG, Choi NJ, Hoffman RM, Shih YT, Lowenstein LM, et al. Patient Adherence to Screening for Lung Cancer in the US: A Systematic Review and Meta-analysis. *JAMA Netw Open.* 2020;3(11):e2025102.
35. Coughlin JM, Zang Y, Terranella S, Alex G, Karush J, Geissen N, et al. Understanding barriers to lung cancer screening in primary care. *J Thorac Dis.* 2020;12(5):2536-44.
36. Wang GX, Baggett TP, Pandharipande PV, Park ER, Percac-Lima S, Shepard J-AO, et al. Barriers to Lung Cancer Screening Engagement from the Patient and Provider Perspective. *Radiology.* 2019;290(2):278-87.
37. Stevens C, Vrinten C, Smith SG, Waller J, Beeken RJ. Acceptability of receiving lifestyle advice at cervical, breast and bowel cancer screening. *Preventive Medicine.* 2019;120:19-25.
38. Stevens C, Vrinten C, Smith SG, Waller J, Beeken RJ. Determinants of willingness to receive healthy lifestyle advice in the context of cancer screening. *British Journal of Cancer.* 2018;119(2):251-7.
39. Baldwin DR, Brain K, Quaipe S. Participation in lung cancer screening. *Transl Lung Cancer Res.* 2021;10(2):1091-8.

40. Vallone F, Lemmo D, Martino ML, Donizzetti AR, Freda MF, Palumbo F, et al. Factors promoting breast, cervical and colorectal cancer screenings participation: A systematic review. *Psycho-Oncology*. 2022;31(9):1435-47.
41. Medicine; Io, Council NR. *Fulfilling the Potential of Cancer Prevention and Early Detection*. Curry SJ, Byers T, Hewitt M, editors. Washington, DC: The National Academies Press; 2003. 564 p.
42. Rivera MP, Durham DD, Long JM, Perera P, Lane L, Lamb D, et al. Receipt of Recommended Follow-up Care After a Positive Lung Cancer Screening Examination. *JAMA Netw Open*. 2022;5(11):e2240403.

Chapter 4: ChatGPT Reinforces Geographic and Language Biases in a Citation Network of Epidemiologic Literature

4.1 Abstract

ChatGPT is an online chatbot that offers tremendous potential as a tool to be employed in the scientific research process. However, it is well documented that ChatGPT has the potential to give biased responses which may influence its utility as a research assistant, including preferences for English-language text and text from wealthy, Western countries. Given these biases, it is important for researchers to consider how ChatGPT's biases affect its ability to complete research tasks. Therefore, in this study, we assess the potential biases in ChatGPT's performance on a research task – rating the validity of a reference to the sentence in which it was cited – using a citation network of scientific articles from epidemiology and public health written by authors from over 160 countries around the world. We find that while ChatGPT completes the task for nearly all citation-reference pairs, it gives higher ratings when citations and references are from English-speaking countries and countries with higher incomes. Our findings suggest that ChatGPT reinforces geographic and language biases that stem from its training process and that output from ChatGPT should be carefully considered and verified by humans before it is fully integrated into the research workflow.

4.2 Introduction

ChatGPT, an online chatbot based on the GPT suite of large language models by OpenAI, was released in late 2022 to worldwide attention. Because of its tremendous ability to take in human input and provide contextual responses based on user input, ChatGPT has been a topic of major interest among the scientific community since its release, especially regarding its potential uses in the research process. Researchers quickly recognized the high potential for plagiarism or misinformation associated with using AI models such as ChatGPT to co-author scientific papers, and within months of the release of ChatGPT, major scientific organizations like the journals Nature and Science established editorial policies forbidding human authors from claiming text generated by ChatGPT as their own or crediting ChatGPT with authorship (1, 2). In the year following these announcements, the scientific community has continued to search for ways to leverage ChatGPT's power as a research assistant, but widespread adoption has hindered by several key shortcomings demonstrated in the platform's output.

Fundamentally, ChatGPT's propensity to make errors in various tasks limits its ability to be a comprehensive research tool. For example, while ChatGPT can produce text that is hard to distinguish from human-written text (3), such text has both semantic and stylistic limitations that are more easily recognized (4). ChatGPT is also susceptible to artificial hallucination - asserting that a statement is factual (and stating that the statement is supported by real-world evidence) when in fact it is completely fabricated (5). While these faults are concerning, perhaps the most controversial issue regarding ChatGPT's use in research is the presence of bias in ChatGPT's underlying systems that change responses when given inputs with different characteristics. Researchers have identified that ChatGPT is susceptible to generating output containing biased language (6, 7). Such biased responses are consequences of a combination of the use of human-

created data in model training and development (8) as well as adherence to associations between words that are key to the models' function (9). Public commentary about these biased outputs have spurred ChatGPT's development team to adjust the model and tailor responses generated by the model to avoid biased language at the expense of not providing answers to all user queries (10). With these issues in mind, we suggest that ChatGPT's applications and use in scientific research could be expanded, given careful analysis of its capabilities and a thorough assessment of its weaknesses.

Two specific areas of bias in ChatGPT's functionality are preferences for English language content and preferences for Western (primarily American) culture. ChatGPT's preference for the English language is substantial: it is faster at answering questions and generating requested output in English compared to other languages, even when more recently updated versions of the model are tested (11). Additionally, the model offers adequate performance translating other languages into English (12), but struggles translating English text into non-English languages (13), especially when the target language does not use Latin characters. ChatGPT also demonstrates a strong alignment with American culture, with research showing that English-language input can guide the model to present its American bias even if the input is about non-Western cultures or countries (14).

In a similar vein, biases towards wealthier, Western countries, especially those that use English as a primary language, are well-documented across the scientific landscape. The concentration of academic researchers in nations such as the United States and the United Kingdom and the historical movement of scientists around the world, among other factors, have led to a gap in scientific research that separates these wealthier, Western nations with less developed nations that are less likely to use English as a primary language, measured by metrics

such as scientific research output (15-19), funding for science (20, 21), opinions on research quality (22), barriers to writing and publishing scientific articles (23), and trust in science (24, 25). The imbalances in ChatGPT's responses and among trends in scientific publishing are both examples of the Matthew effect (26), in which those who start with more wealth accumulate wealth at greater rates than those who begin with less. Because most of the world's writing on the Internet is in English (it is estimated that between 55% and 64% of searchable websites on the Internet are in English (27, 28)), ChatGPT's training process was fueled by large amounts of English-language data. As model development accelerated, additional training data was required and was inevitably composed of mostly English text due to the imbalances in the availability of unlabeled and labeled text data. This "rich get richer" phenomenon is amplified in scientific publishing, as English is firmly established as the default language of science across the world even in countries where English is a second or third language for most residents (29).

With these issues in mind, we aim to provide an overview of ChatGPT's performance on a simple research task and evaluate if ChatGPT's responses reinforce the biases toward English-language text and Western, more affluent countries that have been previously documented. Specifically, we assessed ChatGPT's ability to quantify the relationship between sentences from scientific articles relevant to epidemiologists and public health scientists and the citations used to support those sentences. We hypothesized that ChatGPT would be conducive to the citation-evaluation task (as ChatGPT has been reported to be able to solve many types of language-based tasks (30)) and would be able to provide a validity estimate for almost all the sentence-citation pairs we provided. However, we hypothesized that citation-reference pairs with references written in countries that predominantly speak and write in English such as the United States, Canada, and the UK would have higher overall validity scores compared to those with references from countries

that are not English-dominant and that validity scores for references from high income countries would be higher than those from low income countries.

4.3 Methods

Data Collection. We identified 96,459 epidemiology-related research articles published between January 1, 2018, and December 31, 2022, that were indexed in PubMed. To be included in the study dataset, articles had to be tagged in PubMed with the epidemiology Major Topic term and be identified as discussing an epidemiologic study type (specifically cohort studies, case control studies, randomized controlled trials, and/or cross sectional studies) or epidemiologic bias. We identified articles meeting these criteria using the search query “*((epidemiology [majr]) AND (((cohort studies [mh])) OR (case control studies [mh])) OR (randomized controlled trial[mh])) OR (cross sectional studies [mh]) OR (bias [majr]))) NOT (editorial[pt]) NOT (comment[pt]) NOT (review[pt]) NOT (systematic review[pt]) NOT (meta analysis[pt])) AND (2018/01/01: 2022/12/31[dp])) AND (pubmed pmc[sb])*” to serve as the citing papers in our analysis. We excluded articles that were categorized as systematic reviews or meta analyses because references in these article types are fundamentally different to those found in other scientific articles. We simultaneously obtained the full text of articles contained in the PubMed Central Open Access Commercial subset published in the same date range and determined that we had full text data for 42,241 articles that matched our search query (see **Figure 4.1** for details regarding the data collection and curation process). The citing papers contained 1,457,438 separate citations that referred to another paper in the PubMed database, citing a total of 632,130 unique articles. We limited the final dataset used for analysis to only include articles indexed by PubMed to ensure we could obtain article-level data (including the title, abstract and first author

affiliation) for cited papers using PubMed-supported APIs. After consolidation for network data purposes (see **Network Characteristics** below), we retained a final dataset encompassing 401,737 references in 40,936 citing papers, citing a total of 129,063 unique articles.

ChatGPT and Prompting. Data was provided to OpenAI’s GPT-3.5-turbo model in March 2024. This model is an updated version of GPT 3.5 that was used to fine-tune versions of ChatGPT that were first available to public users in 2023 (OpenAI does not provide methods for programmatic interaction with the web browser based ChatGPT platform). We asked the model to estimate the validity of each citation-reference pair by asking it to generate a numeric value from 0-100, using the prompt: “Here is a sentence from a scientific article: [citing sentence]. An article with the abstract [cited abstract] was used as a citation for the sentence. On a scale of 0-100, how valid is the citation? Keep your response as a numeric value.” If no abstract was provided for the cited article in PubMed, the title of the article was substituted instead. To avoid attribution of geographic or linguistic characteristics to either the citing sentences or cited text, we did not give ChatGPT additional information about the authors of the papers or their affiliations. A series of examples of citing sentences and cited abstracts from the dataset and their corresponding scores is shown in **Table 4.1**.

Outcomes and Variables. Our outcome variable was the ChatGPT-generated validity score. Because OpenAI does not publicly release ChatGPT’s source code, we are unable to specifically define or parameterize the algorithm used to generate the scores; however, given that the score values were collected within a short time frame and that it is unlikely that the algorithms informing ChatGPT’s responses would change within that same time frame, our outcome

measure is likely to be representative of ChatGPT's performance on our designated task. Furthermore, we do not define the outcome variable as the true validity of references to the citing sentences (as that would require manual verification of each citation-reference pair in our dataset by scientific experts), but ChatGPT's estimated validity of each citation-reference pair based on the model's combined comprehension of the citing sentence, the cited reference, and the research task input prompt. Assuming the algorithm used to generate the scores is identical for each sentence, our outcome variable measures the model's understanding of scientific language and comparative ability, conditional on the biases and preferences that influence the model's output.

Our primary explanatory variables of interest were defined by similarities and differences between the countries of origin for cited papers and the countries of origin of citing papers. Specifically, we investigated if scores differed comparing citation-reference pairs where the citation and reference were from the same country to those where their country of origin differed; we hypothesized that citation-reference pairs where the citation and reference were from the same country would score higher compared to pairs where the citation and reference were from different countries. We based the country of origin for each paper on the first author's reported affiliation, as first authors are intended to write and conduct most of the research work comprising scientific publications.

In addition to evaluating differences between international and same-nation citation-reference pairs, we also evaluated if English-language usage or the income level of the countries of origin of citation-reference pairs were associated with differences in the validity score. While most articles in PubMed Central are indexed in English (39) and English is the primary language of scientific writing worldwide (29), we assume that author teams around the world have differing levels of comfort and familiarity with English and therefore may use other languages in

their research. We therefore classified each article in our dataset into either English or non-English primary categories based on the article's country of origin. Language data was based on the CIA World Factbook (40); if multiple languages were listed as the official language or lingua franca of the country, we based our English vs. non-English classification on the language that was most prevalent.

Country income data was obtained from the World Bank (41). Following the World Bank classification scheme, we categorized each country into one of four levels: high income, middle-high income, middle-low income, and low income. These categories, when divided between high and middle-high income vs. middle-low and low income, are comparable to the global North vs. global South distinction historically used in other scientometric studies to compare scientific citation behavior (42-44).

Network Characteristics. Network data was analyzed using the igraph R package (version 1.6.0). We obtained network characteristics for both citing and cited articles including their degree (the number of unique citations in each article) and betweenness centrality (how often an article was on the shortest citation path between two other articles). To reduce computational complexity, we excluded citation-reference pairs where the reference only appeared in one citing paper. Because we do not have complete citation information for the cited articles (as they are not all included in the PMC Open Access subset), we utilize only their degree and centrality among articles whose data was collected for this study.

Statistical Analysis. We report proportions, frequencies, and cross-tabulations of our dataset's characteristics. We use linear regressions with random effects to examine associations between the validity score and characteristics of each citation-reference pair. We use the random effects

parameter to account for clustering by citing article. All statistical analyses were performed using R version 4.3.2 (45).

4.4 Results

40,936 citing articles were included in our dataset for analysis (**Table 4.2**). Authorship of the citing papers represented 160 unique countries. China (n = 4,457, 10.9%), the United States (n = 4,040, 9.9%), the United Kingdom (n = 2,502, 6.1%), Australia (n = 2,324, 5.7%), and Ethiopia (n = 2,120, 5.2%) were the five most represented countries among citing papers. 13,993 citing papers (34.2%) were written by English-primary authors. A majority of citing papers were written by authors from high-income countries (n = 23,997, 58.6%), with each decreasing level of income being less represented (middle-high income: n = 8,618, 21.1%; middle-low income: n = 5,244, 12.8%; low income: n = 3,077, 7.5%).

Our dataset contained data on 129,063 cited papers that were indexed in PubMed and had relevant authorship information available in the PubMed database. Authorship of the cited papers represented 167 unique countries. The United States was the most represented country of authorship among cited papers (n = 48,299, 33.3%), followed by the United Kingdom (n = 20,455, 14.1%), China (n = 7,055, 4.9%), Australia (n = 5,006, 3.5%), and Canada (n = 4,891, 3.4%). Unlike the citing papers, most cited papers were written by English-primary authors (n = 86,471, 59.6%). Cited papers were more likely to be written by authors from high-income countries (n = 118,346, 81.6%), with the lower income brackets proportionally less represented (middle-high income: n = 15,405, 10.6%; middle-low income: n = 8,841, 6.1%; low income: n = 2,425, 1.7%).

Since citing papers contained multiple references, we also analyzed the dataset on the citation-reference pair level ($n = 401,737$). Citing papers contained on average 9.8 citations (median: 9, range: 1-101) and cited papers were on average referenced 3.1 times (median: 2, range: 2-374). We found that 322,613 pairs (80.3%) included a citation and reference from the same country. We also observe that among pairs with English-primary citing papers ($n = 134,100$), 26.2% of references were to papers from non-English primary countries, while among pairs with non-English-primary citing papers ($n = 267,637$), 45.7% of references were to papers by authors from non-English-primary countries. We also observe that many citation-reference pairs include citing and cited papers from countries with similar economies ($n = 250,263$, 62.3%). Papers from lower-income economies citing papers from higher-income economies comprised 30.0% of sentence-reference pairs ($n = 120,605$) while sentences from higher-wealth economies citing papers from lower-wealth economies represented the remaining 7.7% of analyzed pairs ($n = 30,869$).

ChatGPT rated most sentences as highly related to their cited references (mean = 84.3, standard error of mean [s.e.m.] = 0.01; median = 85; range = 0-100). In bivariate analyses, we found that pairs with sentences that cited an English primary paper (mean = 84.4, s.e.m. = 0.01) had higher ratings than those that cited non-English primary papers (mean = 84.1, s.e.m. = 0.01) (Fig. 2b). We also found that pairs with citing sentences that were from English-primary papers (mean = 84.4, s.e.m. = 0.01) had on average higher scores than citing sentences from non-English primary papers (mean = 84.2, s.e.m., 0.001) (**Fig. 4.2A**). We also observed differences in rating across pairs with different income levels of both citing and cited countries of origin, with score increases correlating with higher levels of income for both citing and cited countries (**Fig. 4.2C, 4.2D**). We observed no differences in mean validity score comparing pairs where both citation

and reference were from the same country to pairs where the citation and reference were from different countries.

Results from the random effects linear regression models are shown in **Table 4.3**. We found that controlling for the citing paper's English-language status and country income, citing an English-primary paper was associated with a score increase of 0.17 points compared to citing a paper from a non-English primary country and that for each increase in income level of the cited paper's country of origin, the validity score increased by 0.26 points. We also evaluated regression models including the degree and betweenness centrality of the cited paper (**Supplementary Table S4.1**); while both network characteristics of the cited paper were associated with the validity score, the coefficients of the network metrics were an order of magnitude smaller than those of the primary variables of interest.

4.5 Discussion

In this study, we asked ChatGPT to provide ratings on citation-reference pairs from a citation network of epidemiologic articles to evaluate if the linguistic and economic characteristics of the cited paper's country of origin would bias ChatGPT's performance on a simple research task. We found that the ChatGPT ratings increased when citing sentences or cited papers were from countries that spoke English and had higher incomes. We also observed that while network characteristics of the cited paper such as the degree and betweenness centrality were associated with increases in the ChatGPT ratings, the magnitude of these associations were small in comparison to the associations with the cited paper's characteristics.

We also generally observe that the distribution of our sample of epidemiology papers is like that of the general profile of scientific articles. More affluent, English-speaking nations, such as the United States, the United Kingdom, Canada, and Australia were well-represented among citing papers, but substantially more prevalent among cited papers. Interestingly, while many previous studies (42-44) focus on the differences in scientific research output or funding between the global North (which is generally considered to include only wealthy, English-speaking nations) and global South (which includes large nations such as China, Brazil, and countries in South America and Africa), China appears as a substantial contributor of both citing and cited papers in our dataset, perhaps reflecting the country's increased investment in science (46, 47). As we would expect most scientific citation-reference pairs to be accurate to some degree, the generally high validity scores are not unexpected. However, using both the bivariate and regression analyses, we observe that pairs including papers from high income countries and English-speaking countries get higher scores than those that do not, with the cited papers' country of origin contributing larger increases compared to the citing papers' country of origin. As we previously noted, this preference is likely influenced by the characteristics of the tremendous amount of English-language data used to train GPT (and all its variations and derivative models): as of April 2024, OpenAI openly notes that ChatGPT is "skewed toward Western views and performs best in English. (48)" Although OpenAI has put significant effort into reducing the obvious biases of its flagship model, our findings suggest that some preferences persist; this should alert scientists aiming to use ChatGPT to evaluate the relationships between scientific texts to carefully consider the model's output.

Our study has several limitations that provide possibilities for future research. First, while our study intended to identify areas of bias in ChatGPT's responses using an unbiased research

task, we recognize the possibility that even providing the input prompt in English may influence ChatGPT. However, given ChatGPT's difficulties with translation between languages, we expect that providing the research task prompt in English offers a good balance of efficiency (allowing the model to rate more citation-reference pairs) and performance (ensuring that the model performs the task as requested). Similarly, we did not evaluate ChatGPT's ability to score citation-reference pairs where either the citing sentence or the reference text were written in non-English languages. Finally, our citation network data does not span the entire citation network of the citing papers since we could only use the full text data of papers in PubMed Central or metadata for both citing and cited papers that was stored in the larger PubMed database. It is possible that papers that are not indexed in PubMed or PubMed Central may be different enough from those we analyzed in ways that would affect ChatGPT's ability or performance in conducting the evaluation.

4.6 Figures and Tables

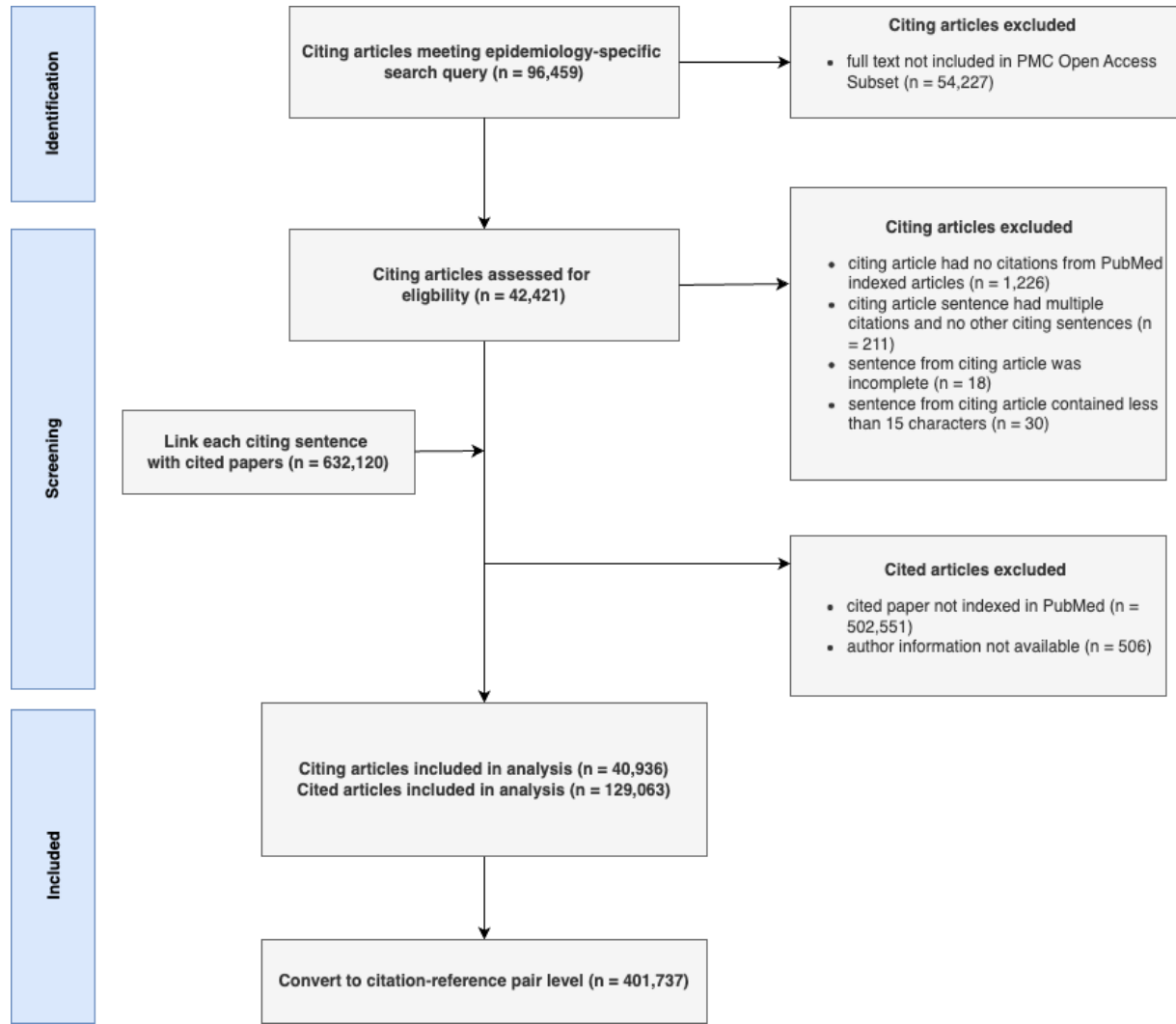


Figure 4.1. Flow diagram for obtaining the final dataset of citation-reference pairs used for analysis (n = 401,737). The final dataset included data from 40,936 citing articles and 129,063 cited articles, combining to include the 401,737 individual citation-reference pairs.

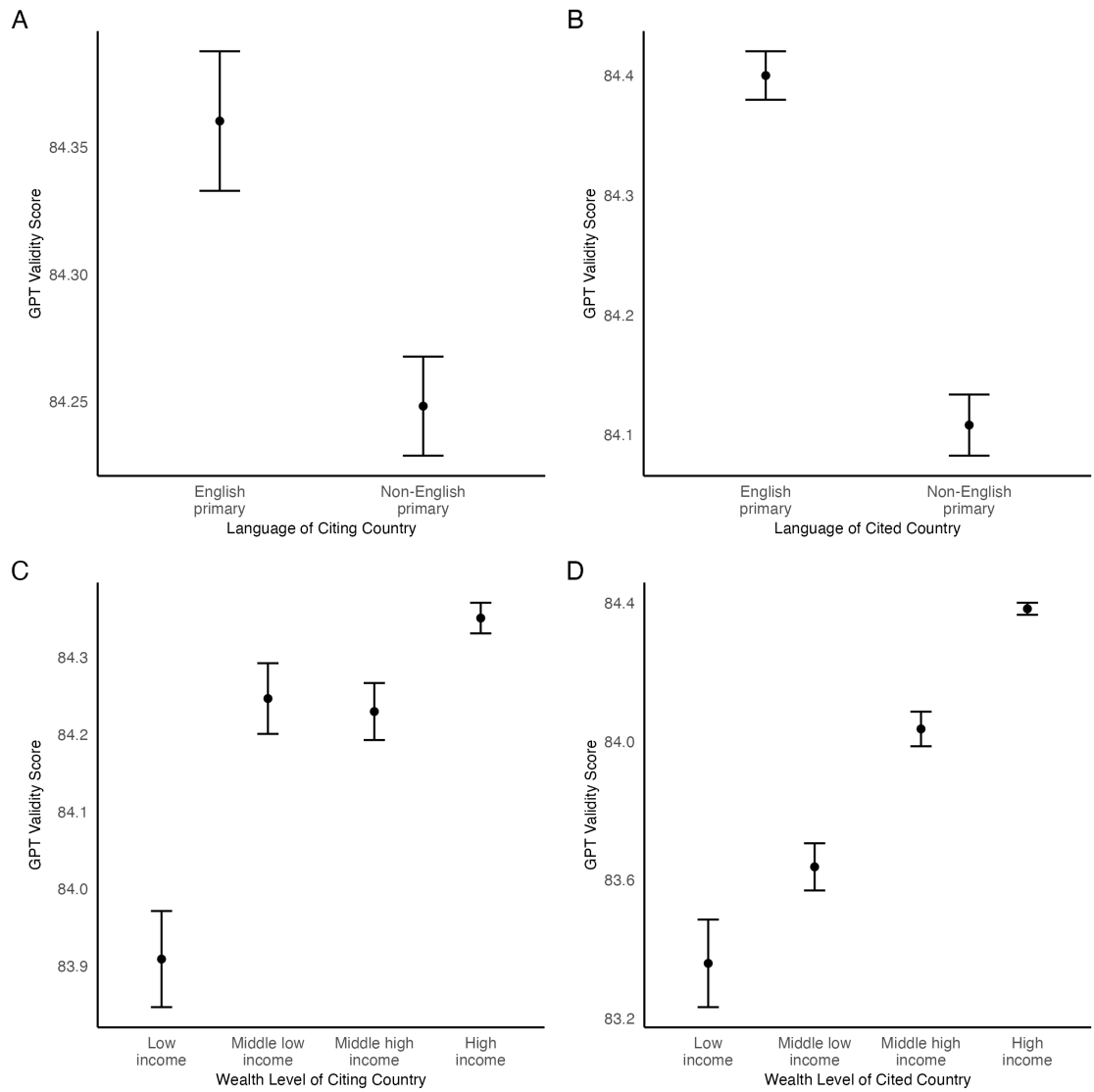


Figure 4.2. ChatGPT-derived validity scores increase when citation-reference pairs include citations or references from English-speaking, high income countries. A and B: We observe higher validity scores when the either the citing paper or cited paper was from a country that primarily speaks English compared to countries that did not primarily speak English. **C and D:** We also find that pairs score highest when either the citing paper or cited paper is from a high income country. Error bars represent 95% confidence intervals of the mean GPT validity scores for each category.

Citing Sentence	Cited Abstract*	GPT Validity Score
<p>In this sense, other studies indicate unhealthy dietary practices in this group, such as rapid weight-loss diets, the omission of certain important food groups, consumption of high-calorie foods and foods with reduced nutritional value and the excessive intake of alcohol, among other toxins (58).</p>	<p>Objectives: To identify the prevalence of cocaine consumption among university students and to analyse the use of other drugs among the regular cocaine consumers (59).</p>	25
<p>The most common intestinal helminthes include Taenia a, Hymenolepis, Ascaris, Strongyloides, Trichuris, Enterobius vermicularis and Hook worm and are usually transmitted from contaminated food, water or environment (60).</p>	<p>Background: Salmonella species are among the most common food borne pathogens worldwide and their infection is one of the major global public health problems. During the last decade, multidrug-resistant Salmonella species have increased to a great deal, especially in developing countries. The prevalence and antimicrobial susceptibility pattern of Salmonella isolates among food handlers at the University of Gondar, Ethiopia, were described in the current investigation (61).</p>	50
<p>Chronic pain is common and creates a significant burden to the individual and society. Emerging research has shown the influence of the family environment on pain outcomes. However, it is not clear what shared factors between family members associate with chronic pain. This study aimed to investigate the family-level contribution to an individual's chronic pain status (62).</p>	<p>Because depression and painful symptoms commonly occur together, we conducted a literature review to determine the prevalence of both conditions and the effects of comorbidity on diagnosis, clinical outcomes, and treatment. The prevalences of pain in depressed cohorts and depression in pain cohorts are higher than when these conditions are individually examined. The presence of pain negatively affects the recognition and treatment of depression (63).</p>	85
<p>Sun exposure for 5 to 15 minutes twice or thrice a week allows the skin to generate about 80% of the vitamin D required by the human body (64).</p>	<p>Most humans depend on sun exposure to satisfy their requirements for vitamin D. Solar ultraviolet B photons are absorbed by 7-dehydrocholesterol in the skin, leading to its transformation to previtamin D3, which is rapidly converted to vitamin D3. Season, latitude, time of day, skin pigmentation, aging, sunscreen use, and glass all influence the cutaneous production of vitamin D3. Once formed, vitamin D3 is metabolized in the liver to 25-hydroxyvitamin D3 and then in the kidney to its biologically active form, 1,25-dihydroxyvitamin D3 (65).</p>	100

Table 4.1. Examples of citation-reference pairs and their validity scores generated by GPT. Each sentence (left column) was used as the first input using the prompt described in the main text, followed by the corresponding abstract (middle column). *For brevity, we include only the first few sentences of each abstract; however, the entire abstract's text was used when generating the scores used in the main analysis.

	Citing Papers (n = 40,936)	Cited Papers (n = 129,063)
Number of countries represented	160	166
English-primary country		
Yes	13,993 (34.2)	77,261 (59.9)
No	26,943 (65.8)	51,802 (40.1)
Income level		
High income	23,997 (58.6)	105,321 (81.6)
Middle high income	8,618 (21.1)	13,637 (10.6)
Middle low income	5,244 (12.8)	7,936 (6.2)
Low income	3,077 (7.5)	2,169 (1.7)
Number of citations*	9.81 [0.03]	3.11 [0.01]
Network metrics		
Degree	29.6 [14.6]	11.6 [31.6]
Betweenness	22.4 [0.27]	20.1 [0.67]

Table 4.2. Characteristics of the citing papers and cited papers in our sample citation network of epidemiologic articles. For categorical variables, percentages are given in parentheses; standard errors of means for continuous variables are given in square brackets. *The number of citations for citing papers is the average number of citations contained in each citing paper; for cited papers, the number is the average number of times the paper is referenced by citing papers.

	Coefficient (Standard error) [p-value]
Fixed Effects	
Intercept	82.97 (0.074) [p < 0.0001]
Cited English-primary paper*	0.17 (0.016) [p < 0.0001]
Cited country Income**	0.26 (0.013) [p < 0.0001]
Citing English-primary paper*	0.05 (0.038) [p = 0.200]
Citing country Income**	0.05 (0.019) [p = 0.009]
Random Effects	
Citing paper-level variance	10.00 (3.162)
Residual variance	18.48 (4.299)

Table 4.3. Random effects linear regression results for the associations between cited paper characteristics and the ChatGPT-generated validity scores. The model controls for the English-language status of the citing paper’s country of origin and the wealth level of the citing paper’s country of origin. We use a random effects term to account for clustering by sentences from the same citing paper. *Reference level is non-English **Reference level is low income.

4.7 Supplementary Information

	Coefficient (Standard error) [p-value]
Fixed Effects	
Intercept	82.96 (0.074) [p < 0.0001]
Cited English-primary paper*	0.15 (0.017) [p < 0.0001]
Cited country Income**	0.26 (0.013) [p < 0.0001]
Citing English-primary paper*	0.05 (0.038) [p = 0.180]
Citing country Income**	0.05 (0.019) [p = 0.01]
Cited paper degree	0.005 (0.0002) [p < 0.0001]
Cited paper betweenness centrality	-0.0002 (0.00002) [p < 0.0001]
Random Effects	
Citing paper-level variance	10.00 (3.162)
Residual variance	18.46 (4.299)

Supplementary Table S4.1. Random effects linear regression results for the associations between cited paper characteristics and the ChatGPT-generated validity scores including network characteristics. The model controls for the English-language status of the citing paper's country of origin and the wealth level of the citing paper's country of origin. We use a random effects term to account for clustering by sentences from the same citing paper. *Reference level is non-English **Reference level is low income.

4.8 References

1. Nature Editorial Board. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*. 2023;613:1.
2. Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379(6630):313-.
3. Casal JE, Kessler M. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*. 2023;2(3):100068.
4. Ma Y, Liu J, Yi F, Cheng Q, Huang Y, Lu W, et al. AI vs. Human -- Differentiation Analysis of Scientific Content Generation. *arXiv*. 2023.
5. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. 2023;15(2):e35179.
6. Fang X, Che S, Mao M, Zhang H, Zhao M, Zhao X. Bias of AI-generated content: an examination of news produced by large language models. *Sci Rep*. 2024;14(1):5224.
7. Wan Y, Pu G, Sun J, Garimella A, Chang K-W, Peng N. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:231009219*. 2023.
8. Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*. 2022;4(3):258-68.
9. Liang PP, Wu C, Morency L-P, Salakhutdinov R. Towards Understanding and Mitigating Social Biases in Language Models. In: Marina M, Tong Z, editors. *Proceedings of the 38th International Conference on Machine Learning; Proceedings of Machine Learning Research: PMLR*; 2021. p. 6565--76.
10. OpenAI. How should AI systems behave, and who should decide? 2023 [Available from: <https://openai.com/blog/how-should-ai-systems-behave>].
11. Jun Y. GPT-4 can solve math problems — but not in all languages 2023 [Available from: <https://www.artfish.ai/p/gpt4-project-euler-many-languages>].
12. Zhu W, Liu H, Dong Q, Xu J, Kong L, Chen J, et al. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:230404675*. 2023.
13. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:230204023*. 2023.

14. Cao Y, Zhou L, Lee S, Cabello L, Chen M, Hershcovich D. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. arXiv preprint arXiv:230317466. 2023.
15. Falagas ME, Michalopoulos AS, Bliziotis IA, Soteriades ES. A bibliometric analysis by geographic area of published research in several biomedical fields, 1995–2003. *Canadian Medical Association Journal*. 2006;175(11):1389-90.
16. Merriman R, Galizia I, Tanaka S, Sheffel A, Buse K, Hawkes S. The gender and geography of publishing: a review of sex/gender reporting and author representation in leading general medical and global health journals. *BMJ Global Health*. 2021;6(5):e005672.
17. Falagas ME, Karavasiou AI, Bliziotis IA. A bibliometric analysis of global trends of research productivity in tropical medicine. *Acta Tropica*. 2006;99(2):155-9.
18. Sweileh WM. Global research output on HIV/AIDS–related medication adherence from 1980 to 2017. *BMC Health Services Research*. 2018;18(1):765.
19. Amano T, González-Varo JP, Sutherland WJ. Languages Are Still a Major Barrier to Global Science. *PLoS Biol*. 2016;14(12):e2000933.
20. Petersen OH. Inequality of Research Funding between Different Countries and Regions is a Serious Problem for Global Science. *Function (Oxf)*. 2021;2(6):zqab060.
21. Nature Editorial Board. Rich countries must align science funding with the SDGs. *Nature*. 2023;621(7979):444.
22. Kowal M, Sorokowski P, Kulczycki E, Żelaźniewicz A. The impact of geographical bias when judging scientific studies. *Scientometrics*. 2021;127(1):265-73.
23. Hanauer DI, Sheridan CL, Englander K. Linguistic Injustice in the Writing of Research Articles in English as a Second Language: Data From Taiwanese and Mexican Researchers. *Written Communication*. 2019;36(1):136-54.
24. Rabesandratana T. These are the countries that trust scientists the most—and the least 2019 [Available from: <https://www.science.org/content/article/global-survey-finds-strong-support-scientists>].
25. Our World in Data. Share of people who trust science, 2020 [Available from: <https://ourworldindata.org/grapher/share-people-trust-science>].
26. Merton RK. The Matthew Effect in Science. *Science*. 1968;159(3810):56-63.
27. Gabriel N, Bhatia A. Lost in Translation: Large Language Models in Non-English Content Analysis. 2023.

28. Foundation IS. What are the most used languages on the Internet? 2023 [Available from: <https://www.isocfoundation.org/2023/05/what-are-the-most-used-languages-on-the-internet/>].
29. Gordin MD. Introduction: Hegemonic Languages and Science. *Isis*. 2017;108(3):606-11.
30. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. *Information Fusion*. 2023;99.
31. Navarro-Prado S, Schmidt-RioValle J, Montero-Alonso MA, Fernández-Aparicio Á, González-Jiménez E. Unhealthy Lifestyle and Nutritional Habits Are Risk Factors for Cardiovascular Diseases Regardless of Professed Religion in University Students. *Int J Environ Res Public Health*. 2018;15(12).
32. Patiño-Masó J, Gras-Pérez E, Font-Mayolas S, Baltasar-Bagué A. [Cocaine abuse and multiple use of psychoactive substances in university students]. *Enferm Clin*. 2013;23(2):62-7.
33. Bhengu KN, Naidoo P, Singh R, Mpaka-Mbatha MN, Nembe N, Duma Z, et al. Immunological Interactions between Intestinal Helminth Infections and Tuberculosis. *Diagnostics (Basel)*. 2022;12(11).
34. Garedeu-Kifelew L, Wondafrash N, Feleke A. Identification of drug-resistant Salmonella from food handlers at the University of Gondar, Ethiopia. *BMC Res Notes*. 2014;7:545.
35. Campbell P, Jordan KP, Smith BH, Scotland G, Dunn KM. Chronic pain in families: a cross-sectional study of shared social, behavioural, and environmental influences. *Pain*. 2018;159(1):41-7.
36. Bair MJ, Robinson RL, Katon W, Kroenke K. Depression and pain comorbidity: a literature review. *Arch Intern Med*. 2003;163(20):2433-45.
37. Raymond-Lezman JR, Riskin SI. Benefits and Risks of Sun Exposure to Maintain Adequate Vitamin D Levels. *Cureus*. 2023;15(5):e38578.
38. Holick MF. Sunlight and vitamin D for bone health and prevention of autoimmune diseases, cancers, and cardiovascular disease. *Am J Clin Nutr*. 2004;80(6 Suppl):1678s-88s.
39. National Library of Medicine. PMC FAQs 2024 [Available from: <https://www.ncbi.nlm.nih.gov/pmc/about/faq/#q16>].
40. CIA World Factbook. CIA World Factbook - Field Listing - Languages 2024 [Available from: <https://www.cia.gov/the-world-factbook/field/languages/>].
41. World Bank. The World by Income and Region 2024 [Available from: <https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html>].

42. Haelewaters D, Hofmann TA, Romero-Olivares AL. Ten simple rules for Global North researchers to stop perpetuating helicopter research in the Global South. *Public Library of Science*; 2021. p. e1009277.
43. Skopec M, Issa H, Reed J, Harris M. The role of geographic bias in knowledge diffusion: a systematic review and narrative synthesis. *Res Integr Peer Rev.* 2020;5:2.
44. Confraria H, Mira Godinho M, Wang L. Determinants of citation impact: A comparative analysis of the Global South versus the Global North. *Research Policy.* 2017;46(1):265-79.
45. R Core Team. *R: A language and environment for statistical computing.* 4.3.0 ed. Vienna, Austria: R Foundation for Statistical Computing; 2023.
46. Mallapaty S. China promises more money for science in 2024 2024 [updated March 8, 2024. Available from: <https://www.nature.com/articles/d41586-024-00695-4#:~:text=The%20government%20will%20spend%20371,by%20China's%20Ministry%20of%20Finance>.
47. Xie Y, Zhang C, Lai Q. China's rise as a major contributor to science and technology. *Proceedings of the National Academy of Sciences.* 2014;111(26):9437-42.
48. OpenAI. Is ChatGPT Biased? 2024 [Available from: <https://help.openai.com/en/articles/8313359-is-chatgpt-biased>.

5. Concluding Remarks

The potential of network science and the use of networks and network data to answer questions in science is seemingly limitless. As awareness of the ubiquity of networks in scientific systems has increased, a multi-disciplinary wave of researchers has come to recognize the need for network methods that capture information about not only the individual elements of the systems they study, but also about the relationships between those components and the position of each component in relation to the network or system in aggregate. Furthermore, network science is applicable to a wide variety of scientific disciplines and is adaptable to new technologies and fields of study; for example, the growing use of network analysis to study the proliferation of information through the Internet represents the use of network methods to study a technology that has only emerged in the last half century. Finally, the increasing awareness of the public of network concepts, especially through the lenses of social media platforms and viral marketing, has spurred interest in network science and provided network scientists with additional opportunities to obtain both funding opportunities and support for research in domains with commercial- or consumer-oriented applications.

This dissertation highlighted three studies that utilized network data and network techniques to answer questions from a diverse set of scientific domains: the evolution of human behaviors, the planning of successful public health interventions, and the use of artificial intelligence tools in scientific research. Our main findings regarding these questions were as follows:

1. In our series of online network games exploring cooperation and punishment using the lens of decision times, we found that punishment was slower than both cooperation and defection and that experimentally imposed time pressure did not reduce the occurrence of

punishment. We also evaluated instances of punishment using several theorized mechanisms from the literature and identified a tendency for punishment decisions to slow down as the perceived complexity of the punishment decision increased. However, we were unable to definitively identify mechanisms of punishment that were affected by time pressure. Future studies in this space should consider experimentally isolating punishment decisions to be based on specific mechanisms the researchers intend to study.

2. Our development of an agent-based simulation framework for a network intervention to increase the lung cancer screening rate successfully identified an intervention setup that would increase the lung cancer screening rate in a community of eligible individuals past the currently low rate of adoption. We also tested two modifications of this setup, seeding individuals with screening behavior and increasing the network size, that could potentially boost the screening rate further; we identified that increasing the network size until the intervention population was sufficiently large resulted in additional increases in screening rate beyond the rate achieved by the intervention alone. This study highlights potential for such a network intervention to be deployed in the real world; future simulation studies should highlight specific subpopulations that could benefit from the intervention and test our simulation framework using those subpopulations' characteristics.
3. Our investigation of potential biases in ChatGPT's responses to a research task using a citation network of articles from epidemiology and public health determined that the model reinforces geographic and linguistic biases that are recognized in the landscape of scientific publishing. Specifically, we determined that ChatGPT gave higher ratings to citation-reference pairs with cited articles from English-speaking countries and countries with higher income compared to pairs with cited articles from non-English speaking countries

and low income countries. We also found that our sample of articles from epidemiology and public health reflected global trends found across scientific publishing: most references were to papers from English-speaking, high-income nations like the United States and United Kingdom. We were also encouraged to find increased rates of citation between papers from non-English speaking countries. These results suggest that researchers should remain cognizant of the issues with using ChatGPT in the research process, especially in the context of evaluating citations and references.