Tobias Gerstenberg

November 2024

Address: Department of Psychology,
Stanford University, USA.
gerstenberg@stanford.edu
http://cicl.stanford.edu/member/tobias_gerstenberg/
tobigerstenberg
tobiasgerstenberg
google scholar
youtube

Professional experience and affiliations

Sep 2018 - present	**Assistant Professor in Cognitive Psychology** PI of the Causality in Cognition Lab (CICL)	Stanford, USA
Apr 2014 - Jul 2018	**Postdoctoral Associate** PI: Prof. Joshua Tenenbaum	MIT, USA
Apr 2013 - Apr 2014	**Postdoctoral Fellow** PI: Prof. Joshua Tenenbaum	MIT, USA

Education

Aug 2009 - Feb 2013	**PhD Candidate** Advisor: Prof. David Lagnado	University College London, England
Oct 2008 - Aug 2009	**Course Attendance** Psychology, Computer Science, Philosophy, and Theology	Humboldt University, Berlin, Germany
Sep 2007 - Sep 2008	**MSc Cognitive and Decision Sciences** Supervisor: Dr. David Lagnado	University College London, England
Sep 2005 - July 2007	**Vordiplom (eqvl. BSc)**	Humboldt University, Berlin, Germany
Sep 1995 - July 2004	**Abitur (eqvl. A-levels)**	Scheffel Gymnasium, Bad Säckingen, Germany

Awards, Grants & Scholarships

Nov 2024	Stanford Human-Centered Artifical Intelligence (HAI) seed grant: "A User-Centered Task-Driven Characterization of the Influence of AI Assistance on Clinical Decision-Making" ($75,000; co-PI with Prof. Shriti Raj, Prof. Jonathan Chen and Prof. Behnam Rahdari)
Nov 2024	Stanford Human-Centered Artifical Intelligence (HAI) seed grant: "Generalizable Physical Intuition for Robots" ($75,000; co-PI with Prof. Jeannette Bohg and Prof. Kayvon Fatahalian)
July 2024	Stanford Human-Centered Artifical Intelligence (HAI) Hoffman-Yee grant: "Integrating intelligence: Building shared conceptual grounding for interacting with Generative AI" ($500,000; co-PI)
July 2024	HAI-IBM funding "Teach Yourself Preference Optimization (TYPO): Preference Optimization without Preference Labels" ($85,000; PI)

June 2024	Stanton Prize from the Society for Philosophy and Psychology
Oct 2023	HAI-Google Funding ($80,000; co-PI with Prof. Noah Goodman)
Sep 2023	Microsoft research grant on "Accelerating Foundation Models Research" ($20,000 in Azure cloud credits; PI)
Sep 2023	Cooperative AI grant "ACES: ACtion Explanation through counterfactual Simulation" ($500,000; PI)
Nov 2022	Stanford Human-Centered Artifical Intelligence (HAI) seed grant: "Creative physical problem solving in humans and robots" ($75,000; co-PI with Prof. Jeannette Bohg)
Aug 2022	Stanford Human-Centered Artifical Intelligence (HAI) Hoffman-Yee grant: "MARPLE: Explaining what happened through multi-modal simulation" ($500,000; PI)
Sep 2021	Stanford Human-Centered Artifical Intelligence (HAI) seed grant: "In touch with causation" ($75,000; co-PI with Prof. Sean Follmer)
July 2021	Templeton Foundation: Attributions of Purpose and the Philosophy of Religion ($250,000; co-PI with Prof. Joshua Knobe)
June 2021	John Philip Coghlan Fellowship from Stanford University ($4,000)
Dec 2020	US-Israel Binational Science Foundation (BSF) "Judging responsibility under uncertainty" ($102,600; PI)
Sep 2020	Stanford Human-Centered Artifical Intelligence (HAI) seed grant: "The Science and Engineering of Explanations (SEE)" ($75,000; PI)
Jun 2019	Stanford Hellman Faculty Scholar Award: "Multi-modal inference through mental simulation" ($39,960; PI)
May 2019	Stanford Human-Centered Artifical Intelligence (HAI) seed grant: "Multi-modal inference in brains, minds, and machines" ($75,000; PI)
2015 - 2016	Templeton Foundation: "The Experience Project: Computational models of the intuitive theory of transformative experiences" ($90,000; co-PI)
Apr 2013 - 2014	MIT Intelligence Initiative PostDoc Scholarship
May 2004 - Dec 2012	e-fellows.net Scholarship
Sep 2009 - Sep 2012	AXA PhD Scholarship (€120,000)
Apr 2012	PPG Conference Best Presentation Award, London, England *Award for best oral presentation at the annual conference of UCL's postgraduate peer group*
Mar 2011	UCL Sully Scholarship (£1,466), London, England *Award for best PhD upgrade talk in the Department of Cognitive, Perceptual, and Brain Sciences, 2010*
Jan 2011	UCL Bogue Scholarship (£3,450), London, England *Grant to cover 3-month lab visits to Stanford and MIT*
Oct 2010	KogWis, Berlin, Germany *Conference travel grant*
Feb 2006 - Sep 2009	Scholar of the German National Academic Foundation
July 2009	Max Planck Institute Berlin, Germany *2nd prize in poster competition of the Summer Institute on Bounded Rationality*

Nov 2008	University College London, England *Best performing student of the academic year for the MSc in Cognitive and Decision Sciences*
July 2006	Humboldt University, Berlin, Germany *Winner of the competition for the first psychological experiment*

Publications

Note: * indicates joint first authorship.

Submitted papers

1. Morris, A., J. Phillips, T. Icard, J. Knobe, **T. Gerstenberg**, and F. A. Cushman (submitted). Looking back to plan ahead: Causal judgments as a sampling approximation for action effectiveness.
2. Beller, A. and **T. Gerstenberg** (2024). Causation, meaning, and communication. *PsyArXiv*.
3. Du, M., A. Khazatsky, **T. Gerstenberg**, and C. Finn (2024). To Err is Robotic: Rapid Value-Based Trial-and-Error during Deployment. *arXiv*. http://arxiv.org/abs/2406.15917.
4. Gandhi, K., Z. Lynch, J.-P. Fränken, K. Patterson, S. Wambu, **T. Gerstenberg**, D. C. Ong, and N. D. Goodman (2024). Human-like Affective Cognition in Foundation Models. *arXiv*. https://arxiv.org/abs/2409.11733.
5. Johnson, S. G. B., A.-H. Karimi, Y. Bengio, N. Chater, **T. Gerstenberg**, K. Larson, S. Levine, M. Mitchell, B. Schölkopf, and I. Grossmann (2024). Imagining and building wise machines: The centrality of AI metacognition. *arXiv*. https://arxiv.org/abs/2411.02478.

Journal articles

6. Amemiya, J., G. D. Heyman, and **T. Gerstenberg** (2024). Children use disagreement to infer what happened. *Cognition*.
7. **Gerstenberg**, **T.** (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*.
8. Prinzing, M., D. Rose, S. Zhang, E. Tu, A. Concha, J. Schaffer, M. Rea, **T. Gerstenberg**, and J. Knobe (2024). From artifacts to human lives: Investigating the domain-generality of judgments about purposes. en. *Journal of Experimental Psychology: General*. https://osf.io/7enkr.
9. **Gerstenberg**, **T.**, D. A. Lagnado, and R. Zultan (2023). Making a positive difference: Criticality in groups. *Cognition*.
10. Gong, T., **T. Gerstenberg**, R. Mayrhofer, and N. R. Bramley (2023). Active causal structure learning in continuous time. *Cognitive Psychology* **140**, 101542.
11. Wu, S. A. and **T. Gerstenberg** (2023). If not me, then who? Responsibility and replacement. *Cognition*.
12. Zhou, L., K. A. Smith, J. B. Tenenbaum, and **T. Gerstenberg** (2023). Mental Jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*.
13. **Gerstenberg**, **T.** (2022). What would have happened? Counterfactuals, hypotheticals, and causal judgments. *Philosophical Transactions of the Royal Society B: Biological Sciences*.
14. Kirfel, L., T. F. Icard, and **T. Gerstenberg** (2022). Inference from explanation. *Journal of Experimental Psychology: General*.
15. **Gerstenberg**, **T.**, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review* **128**(6), 936–975.
16. **Gerstenberg**, **T.** and S. Stephan (2021). A counterfactual simulation model of causation by omission. *Cognition* **216**, 104842.

17. Kominsky, J. F., **T. Gerstenberg**, M. Pelz, M. Sheskin, H. Singmann, L. Schulz, and F. C. Keil (2021). The trajectory of counterfactual simulation in development. *Developmental Psychology* **57**(2), 253–268.
18. Langenhoff, A. F., A. Wiegmann, J. Y. Halpern, J. B. Tenenbaum, and **T. Gerstenberg** (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology* **129**, 101412.
19. Sosa, F. A., T. Ullman, J. B. Tenenbaum, S. J. Gershman, and **T. Gerstenberg** (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition* **217**, 104890.
20. **Gerstenberg**, **T.** and T. F. Icard (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General* **149**(3), 599–607.
21. Grinfeld, G., D. Lagnado, **T. Gerstenberg**, J. F. Woodward, and M. Usher (2020). Causal Responsibility and Robust Causation. *Frontiers in Psychology* **11**, 1069.
22. Niemi, L., J. Hartshorne, **T. Gerstenberg**, M. Stanley, and L. Young (2020). Moral Values Reveal the Causality Implicit in Verb Meaning. *Cognitive Science* **44**(6).
23. Morris, A., J. Phillips, **T. Gerstenberg**, and F. Cushman (2019). Quantitative causal selection patterns in token causation. *PLoS ONE* **14**(8), e0219704.
24. Bramley, N. R., **T. Gerstenberg**, R. Mayrhofer, and D. A. Lagnado (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **44**(12), 1880–1910.
25. Koskuba, K., **T. Gerstenberg**, H. Gordon, D. A. Lagnado, and A. Schlottmann (2018). What's fair? How children assign reward to members of teams with differing causal structures. *Cognition* **177**, 234–248.
26. **Gerstenberg**, **T.**, M. F. Peterson, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum (2017). Eye-Tracking causality. *Psychological Science* **28**(12), 1731–1744.
27. Gershman*, S. J., T. Gerstenberg*, C. L. Baker, and F. Cushman (2016). Plans, habits, and theory of mind. *PLoS ONE* **11**(9), e0162246.
28. Alicke, M. D., D. R. Mandel, D. Hilton, **T. Gerstenberg**, and D. A. Lagnado (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives on Psychological Science* **10**(6), 790–812.
29. Kominsky, J. F., J. Phillips, **T. Gerstenberg**, D. A. Lagnado, and J. Knobe (2015). Causal superseding. *Cognition* **137**, 196–209.
30. Lagnado, D. A., **T. Gerstenberg**, and R. Zultan (2013). Causal responsibility and counterfactuals. *Cognitive Science* **47**, 1036–1073.
31. **Gerstenberg**, **T.** and D. A. Lagnado (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review* **19**(4), 729–736.
32. Zultan, R., **T. Gerstenberg**, and D. A. Lagnado (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition* **125**(3), 429–440.
33. **Gerstenberg**, **T.** and D. A. Lagnado (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition* **115**(1), 166–171.
34. Meder, B., **T. Gerstenberg**, Y. Hagmayer, and M. R. Waldmann (2010). Observing and Intervening: Rational and Heuristic Models of Causal Decision Making. *Open Psychology Journal* **3**, 119–135.

Peer-reviewed conference proceedings articles

35. Andukuri, C., J.-P. Fränken, **T. Gerstenberg**, and N. D. Goodman (2024). STaR-GATE: Teaching Language Models to Ask Clarifying Questions. *Conference on Language Modeling*. http://arxiv.org/abs/2403.19154.

36. Brockbank, E., J. Yang, M. Govil, J. E. Fan, and **T. Gerstenberg** (2024). Without his cookies, he's just a monster: A counterfactual simulation model of social explanation. In: *Proceedings of the 46th Annual Conference of the Cognitive Science Society*. Ed. by L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, and E. Hazeltine.
37. Fränken, J.-P., K. Gandhi, T. Qiu, A. Khawaja, N. D. Goodman, and **T. Gerstenberg** (2024). Procedural dilemma generation for evaluating moral reasoning in humans and language models. In: *Proceedings of the 46th Annual Conference of the Cognitive Science Society*. Ed. by L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, and E. Hazeltine.
38. Fränken, J.-P., E. Zelikman, R. Rafailov, K. Gandhi, **T. Gerstenberg**, and N. D. Goodman (2024). Self-supervised alignment with mutual information: Learning to follow principles without preference labels. In: *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/2404.14313.
39. Jin*, E., Z. Huang*, J.-P. Fränken, W. Liu, H. Cha, E. Brockbank, S. Wu, R. Zhang, J. Wu, and **T. Gerstenberg** (2024). MARPLE: A Benchmark for Long-Horizon Inference. In: *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/2410.01926.
40. Keshmirian, A., M. Willig, B. Hemmatian, U. Hahn, K. Kersting, and **T. Gerstenberg** (2024). Chain versus common cause: Biased causal strength judgments in humans and large language models. In: *Proceedings of the 46th Annual Conference of the Cognitive Science Society*. Ed. by L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, and E. Hazeltine.
41. Tsirtsis, S., M. Gomez-Rodriguez, and **T. Gerstenberg** (2024). Towards a computational model of responsibility judgments in sequential human-AI collaboration. In: *Proceedings of the 46th Annual Conference of the Cognitive Science Society*. Ed. by L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, and E. Hazeltine.
42. Wu, S. A., E. Brockbank, H. Cha, J.-P. Fränken, E. Jin, Z. Huang, W. Liu, R. Zhang, J. Wu, and **T. Gerstenberg** (2024). Whodunnit? Inferring what happened from multimodal evidence. In: *Proceedings of the 46th Annual Conference of the Cognitive Science Society*. Ed. by L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, and E. Hazeltine.
43. Wu, S. A., X. Ren, **T. Gerstenberg**, Y. Choi, and S. Levine (2024). Resource-rational moral judgment. In: *Proceedings of the 46th Annual Conference of the Cognitive Science Society*. Ed. by L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, and E. Hazeltine.
44. Cao, A., A. Geiger, E. Kreiss, T. Icard, and **T. Gerstenberg** (2023). A Semantics for Causing, Enabling, and Preventing Verbs Using Structural Causal Models. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Ed. by M. B. Goldwater, F. Anggoro, B. Hayes, and D. C. Ong.
45. Chase, E., **T. Gerstenberg**, and S. Follmer (2023). Realism of Visual, Auditory, and Haptic Cues in Phenomenal Causality. In: *IEEE World Haptics Conference (WHC)*.
46. Fränken, J.-P., A. Khawaja, K. Gandhi, J. Moore, N. D. Goodman, and **T. Gerstenberg** (2023). Off The Rails: Procedural Dilemma Generation for Moral Reasoning. In: *AI Meets Moral Philosophy and Moral Psychology Workshop (NeurIPS 2023)*.
47. Fränken, J.-P., S. Kwok, P. Ye, K. Gandhi, D. Arumugam, J. Moore, A. Tamkin, **T. Gerstenberg**, and N. D. Goodman (2023). Social Contract AI: Aligning AI Assistants with Implicit Group Norms. In: *Socially Responsible Language Modelling Research Workshop (NeurIPS 2023)*.
48. Gandhi*, K., J.-P. Fränken*, **T. Gerstenberg**, and N. D. Goodman (2023). Understanding Social Reasoning in Language Models with Language Models. In: *Advances in Neural Information Processing Systems*.
49. Gonzalez, B., **T. Gerstenberg**, and J. Phillips (2023). Causal Reasoning Across Agents and Objects. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Ed. by M. B. Goldwater, F. Anggoro, B. Hayes, and D. C. Ong.

50. Kirfel, L., X. Bunk, R. Zultan, and **T. Gerstenberg** (2023). Father, don't forgive them, for they could have known what they're doing. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Ed. by M. B. Goldwater, F. Anggoro, B. Hayes, and D. C. Ong.
51. Kirfel, L., R. J. MacCoun, T. Icard, and **T. Gerstenberg** (2023). Anticipating the risks and benefits of counterfactual world simulation models. In: *AI Meets Moral Philosophy and Moral Psychology Workshop (NeurIPS 2023)*.
52. Nam, A., C. Hughes, T. Icard, and **T. Gerstenberg** (2023). Show and tell: Learning causal structures from observations and explanations. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Ed. by M. B. Goldwater, F. Anggoro, B. Hayes, and D. C. Ong.
53. Nie, A., Y. Zhang, A. Amdekar, C. J. Piech, T. Hashimoto, and **T. Gerstenberg** (2023). MoCa: Measuring human-language model alignment on causal and moral judgment tasks. In: *Advances in Neural Information Processing Systems*.
54. Rose, D., S. Zhang, Q. Han, and **T. Gerstenberg** (2023). Teleology and generics. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Ed. by M. B. Goldwater, F. Anggoro, B. Hayes, and D. C. Ong.
55. Shin, S. M. and **T. Gerstenberg** (2023). Learning what matters: Causal abstraction in human inference. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Ed. by M. B. Goldwater, F. Anggoro, B. Hayes, and D. C. Ong.
56. Vasconcelos, H., M. Jörke, M. Grunde-McLaughlin, **T. Gerstenberg**, M. S. Bernstein, and R. Krishna (2023). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* **7**(CSCW1), 1–38.
57. Wu, S. A., S. Sridhar, and **T. Gerstenberg** (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Ed. by M. B. Goldwater, F. Anggoro, B. Hayes, and D. C. Ong.
58. Zhang*, S., J. S. She*, **T. Gerstenberg**, and D. Rose (2023). You are what you're for: Essentialist categorization in large language models. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Ed. by M. B. Goldwater, F. Anggoro, B. Hayes, and D. C. Ong.
59. Beller, A., Y. Xu, S. Linderman, and **T. Gerstenberg** (2022). Looking into the past: Eye-tracking mental simulation in physical inference. *Cognitive Science Proceedings*.
60. Outa*, J., X. J. Zhou*, H. Gweon, and **T. Gerstenberg** (2022). Stop, children what's that sound? Multi-modal inference through mental simulation. *Cognitive Science Proceedings*.
61. Srivastava, A. et al. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv*.
62. Vodrahalli, K., R. Daneshjou, **T. Gerstenberg**, and J. Zou (2022). Do humans trust advice more if it comes from AI? an analysis of human-AI interactions. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp.763–777.
63. Vodrahalli, K., **T. Gerstenberg**, and J. Y. Zou (2022). Uncalibrated models can improve human-AI collaboration. *Advances in Neural Information Processing Systems* **35**, 4004–4016.
64. Wu, S., S. Sridhar, and **T. Gerstenberg** (2022). That was close! A counterfactual simulation model of causal judgments about decisions. *Cognitive Science Proceedings*.
65. Davis, Z. J., K. R. Allen, and **T. Gerstenberg** (2021). Who went fishing? Inferences from social evaluations. In: *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.
66. Beller, A., E. Bennett, and **T. Gerstenberg** (2020). The language of causation. In: *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Ed. by S. Denison, M. Mack, Y. Xu, and B. C. Armstrong. Cognitive Science Society, pp.3133–3139.

67. Bridgers, S., C. Yang, **T. Gerstenberg**, and H. Gweon (2020). Whom will Granny thank? Thinking about what could have been informs children's inferences about relative helpfulness. In: *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.
68. Kirfel, L., T. F. Icard, and **T. Gerstenberg** (2020). Learning from explanations. In: *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.
69. Kominsky, J. F., **T. Gerstenberg**, M. Pelz, H. Singmann, M. Sheskin, and F. Keil (2019). The trajectory of counterfactual simulation in development. In: *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. Ed. by A. Goel, C. Seifert, and C. Freksa. Montreal, QB: Cognitive Science Society, pp.2044–2050.
70. Yildirim, I., B. Saeed, G. Bennett-Pierre, **T. Gerstenberg**, J. B. Tenenbaum, and H. Gweon (2019). Explaining intuitive difficulty judgments by modeling physical effort and risk. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.
71. Gates, M. A., T. L. Veuthey, M. H. Tessler, K. A. Smith, **T. Gerstenberg**, L. Bayet, and J. B. Tenenbaum (2018). Tiptoeing around it: Inference from absence in potentially offensive speech. In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.
72. Bramley, N. R., R. Mayrhofer, **T. Gerstenberg**, and D. A. Lagnado (2017). Causal learning from interventions and dynamics in continuous time. In: *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Ed. by G. Gunzelmann, A. Howes, T. Tenbrink, and E. Davelaar. Austin, TX: Cognitive Science Society, pp.150–155.
73. **Gerstenberg**, **T.**, L. Zhou, K. A. Smith, and J. B. Tenenbaum (2017). Faulty towers: A hypothetical simulation model of physical support. In: *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Ed. by G. Gunzelmann, A. Howes, T. Tenbrink, and E. Davelaar. Austin, TX: Cognitive Science Society, pp.409–414.
74. Stephan, S., P. Willemsen, and **T. Gerstenberg** (2017). Marbles in Inaction: Counterfactual Simulation and Causation by Omission. In: *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Ed. by G. Gunzelmann, A. Howes, T. Tenbrink, and E. Davelaar. Austin, TX: Cognitive Science Society, pp.1132–1137.
75. Yildirim, I., **T. Gerstenberg**, B. Saeed, M. Toussant, and J. B. Tenenbaum (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. In: *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Ed. by G. Gunzelmann, A. Howes, T. Tenbrink, and E. Davelaar. Austin, TX: Cognitive Science Society, pp.3584–3589.
76. Bramley, N., **T. Gerstenberg**, and J. B. Tenenbaum (2016). Natural science: Active learning in dynamic physical microworlds. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Ed. by A. Papafragou, D. Grodner, D. Mirman, and J. C. Trueswell. Austin, TX: Cognitive Science Society, pp.2567–2572.
77. **Gerstenberg**, **T.** and J. B. Tenenbaum (2016). Understanding "almost": Empirical and computational studies of near misses. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Ed. by A. Papafragou, D. Grodner, D. Mirman, and J. C. Trueswell. Austin, TX: Cognitive Science Society, pp.2777–2782.
78. Niemi, L., J. Hartshorne, **T. Gerstenberg**, and L. Young (2016). Implicit measurement of motivated causal attribution. In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Ed. by A. Papafragou, D. Grodner, D. Mirman, and J. C. Trueswell. Austin, TX: Cognitive Science Society, pp.1745–1750.
79. Allen, K., J. Jara-Ettinger, **T. Gerstenberg**, M. Kleiman-Weiner, and J. B. Tenenbaum (2015). Go fishing! Responsibility judgments when cooperation breaks down. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Ed. by D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, and P. P. Maglio. Austin, TX: Cognitive Science Society, pp.84–89.

80. **Gerstenberg**, **T.**, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum (2015). How, whether, why: Causal judgments as counterfactual contrasts. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Ed. by D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, J. Matlock T., C. D., and P. P. Maglio. Cognitive Science Society, pp.782–787.
81. **Gerstenberg**, **T.**, J. Y. Halpern, and J. B. Tenenbaum (2015). Responsibility judgments in voting scenarios. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Ed. by D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, J. Matlock T., C. D., and P. P. Maglio. Austin, TX: Cognitive Science Society, pp.788–793.
82. Kleiman-Weiner, M., **T. Gerstenberg**, S. Levine, and J. B. Tenenbaum (2015). Inference of intention and permissibility in moral decision making. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Ed. by D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, J. Matlock T., C. D., and P. P. Maglio. Austin, TX: Cognitive Science Society, pp.1123–1128.
83. Bramley, N., **T. Gerstenberg**, and D. A. Lagnado (2014). The order of things: Inferring causal structure from temporal patterns. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Ed. by P. Bello, M. Guarini, M. McShane, and B. Scassellati. Austin, TX: Cognitive Science Society, pp.236–241.
84. **Gerstenberg**, **T.**, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum (2014). From counterfactual simulation to causal judgment. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Ed. by P. Bello, M. Guarini, M. McShane, and B. Scassellati. Austin, TX: Cognitive Science Society, pp.523–528.
85. **Gerstenberg**, **T.**, T. D. Ullman, M. Kleiman-Weiner, D. A. Lagnado, and J. B. Tenenbaum (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Ed. by P. Bello, M. Guarini, M. McShane, and B. Scassellati. Austin, TX: Cognitive Science Society, pp.2263–2268.
86. Kominsky, J. F., J. Phillips, J. Knobe, **T. Gerstenberg**, and D. A. Lagnado (2014). Causal supersession. In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Ed. by P. Bello, M. Guarini, M. McShane, and B. Scassellati. Austin, TX: Cognitive Science Society, pp.761–766.
87. **Gerstenberg**, **T.**, C. Bechlivanidis, and D. A. Lagnado (2013). Back on track: Backtracking in counterfactual reasoning. In: *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Ed. by M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth. Austin, TX: Cognitive Science Society, pp.2386–2391.
88. **Gerstenberg**, **T.** and N. D. Goodman (2012). Ping Pong in Church: Productive use of concepts in human probabilistic inference. In: *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Ed. by N. Miyake, D. Peebles, and R. P. Cooper. Austin, TX: Cognitive Science Society, pp.1590–1595.
89. **Gerstenberg**, **T.**, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In: *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Ed. by N. Miyake, D. Peebles, and R. P. Cooper. Austin, TX: Cognitive Science Society, pp.378–383.
90. McCoy, J., T. Ullman, A. Stuhlmüller, **T. Gerstenberg**, and J. B. Tenenbaum (2012). Why blame Bob? Probabilistic generative models, counterfactual reasoning, and blame attribution. In: *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Ed. by N. Miyake, D. Peebles, and R. P. Cooper. Austin, TX: Cognitive Science Society, pp.1996–2001.
91. **Gerstenberg**, **T.**, A. Ejova, and D. A. Lagnado (2011). Blame the skilled. In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Ed. by C. Carlson, C. Hölscher, and T. Shipley. Austin, TX: Cognitive Science Society, pp.720–725.
92. **Gerstenberg**, **T.**, D. A. Lagnado, M. Speekenbrink, and C. Cheung (2011). Rational order effects in responsibility attributions. In: *Proceedings of the 33rd Annual Conference of the Cognitive Sci-*

ence Society. Ed. by C. Carlson, C. Hölscher, and T. Shipley. Austin, TX: Cognitive Science Society, pp.1715–1720.

93. Schächtele, S., **T. Gerstenberg**, and D. A. Lagnado (2011). Beyond outcomes: The influence of intentions and deception. In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Ed. by L. Carlson, C. Hölscher, and T. Shipley. Austin, TX: Cognitive Science Society, pp.1860–1865.
94. **Gerstenberg**, **T.**, D. A. Lagnado, and Y. Kareev (2010). The dice are cast: The role of intended versus actual contributions in responsibility attribution. In: *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Ed. by S. Ohlsson and R. Catrambone. Austin, TX: Cognitive Science Society, pp.1697–1702.

Book chapters

95. Goodman, N. D., **T. Gerstenberg**, and J. B. Tenenbaum (2024). "Probabilistic programs as a unifying language of thought". In: *Reverse-engineering the mind: The Bayesian approach to Cognitive Science*. Ed. by T. L. Griffiths, N. Chater, and J. B. Tenenbaum.
96. Smith, K. A., J. B. Hamrick, A. N. Sanborn, P. W. Battaglia, **T. Gerstenberg**, T. D. Ullman, and J. B. Tenenbaum (2024). "Probabilistic models of physical reasoning". In: *Reverse-engineering the mind: The Bayesian approach to Cognitive Science*. Ed. by T. L. Griffiths, N. Chater, and J. B. Tenenbaum.
97. Bramley, N. R., **T. Gerstenberg**, R. Mayrhofer, and D. A. Lagnado (2019). "Intervening in time". In: *Time and Causality Across the Sciences*. Ed. by S. Kleinberg. Cambridge University Press, pp.86–115.
98. **Gerstenberg**, **T.** and J. B. Tenenbaum (2017). "Intuitive Theories". In: *Oxford Handbook of Causal Reasoning*. Ed. by M. Waldmannn. Oxford University Press, pp.515–548.
99. Lagnado, D. A. and **T. Gerstenberg** (2017). "Causation in legal and moral reasoning". In: *Oxford Handbook of Causal Reasoning*. Ed. by M. Waldmann. Oxford University Press, pp.565–602.
100. Goodman, N. D., J. B. Tenenbaum, and **T. Gerstenberg** (2015). "Concepts in a probabilistic language of thought". In: *The Conceptual Mind: New Directions in the Study of Concepts*. Ed. by E. Margolis and S. Lawrence. MIT Press, pp.623–653.
101. Lagnado, D. A. and **T. Gerstenberg** (2015). "A difference-making framework for intuitive judgments of responsibility". In: *Oxford Studies in Agency and Responsibility*. Ed. by D. Shoemaker. Vol. 3. Oxford University Press, pp.213–241.
102. **Gerstenberg**, **T.** and D. A. Lagnado (2014). "Attributing responsibility: Actual and counterfactual worlds". In: *Oxford Studies in Experimental Philosophy*. Ed. by J. Knobe, T. Lombrozo, and S. Nichols. Vol. 1. Oxford University Press, pp.91–130.

ArXiv

103. Morris, A., J. Phillips, T. Icard, J. Knobe, **T. Gerstenberg**, and F. Cushman (2018). Judgments of actual causation approximate the effectiveness of interventions. *PsyArXiv*. https://psyarxiv.com/nq53z.

PhD thesis & MSc thesis

104. **Gerstenberg**, **T.** (2013). Making a difference: Responsibility, causality, and counterfactuals. *Unpublished PhD thesis*.
105. **Gerstenberg**, **T.** (2009). The allocation of responsibility amongst multiple causes. *Unpublished MSc thesis*.

Conferences & Workshops

March 2025	Invited Symposium: Moral Cognition and Competence in LLMs, San Francisco, USA ***Causal and moral reasoning in humans and LLMs***
February 2025	Bellairs Workshop on Causality: Causality in the era of foundation models, Barbados ***Invited talk***
September 2024	Workshop on "Analyzing High-dimensional Traces of Intelligent Behavior", Institute for Pure & Applied Mathematics (IPAM), Los Angeles, USA ***Invited talk**: Counterfactual simulation in causal cognition*
July 2024	Cognitive Science Conference (46th), Rotterdam, Holland ***Invited talk in symposium**: Holding others responsible: The role of counterfactual contrasts* ***Talk**: Procedural dilemma generation for evaluating moral reasoning in humans and language models* ***Talk**: Do as I explain: Explanations communicate optimal interventions (presented by Lara Kirfel)* ***Talk**: Towards a computational model of responsibility judgments in sequential human-AI collaboration (presented by Stratis Tsirtsis)* ***Poster**: Whodunnit? Inferring what happened from multimodal evidence (presented by Sarah Wu)* ***Poster**: Resource-rational moral judgment (presented by Sarah Wu)* ***Poster**: Without his cookies, he's just a monster: A counterfactual simulation model of social explanation (presented Erik Brockbank)* ***Poster**: Chain versus common cause: Biased causal strength judgments in humans and large language models (presented Anita Keshmirian)*
June 2024	Society for Philosophy and Psychology (SPP) Conference, Purdue University, USA ***Stanton Prize Address*** ***Talk**: Do as I explain: Explanations communicate optimal interventions* ***Talk**: Without his cookies, he's just a monster: A counterfactual simulation model of social explanation (presented by Erik Brockbank)* ***Talk**: Resource-rational moral judgment (presented by Sarah Wu)* ***Poster**: Off The Rails: Procedural Dilemma Generation for Evaluating Moral Reasoning*
June 2024	Cooperative AI Workshop, Santa Cruz, USA
June 2024	Simons Center Workshop, Berkeley, USA ***Invited Speaker:** Beyond the here and now: Counterfactual simulation in causal cognition*
May 2024	Workshop on Mental Models and Learning from Observations, New York, USA ***Invited Speaker:** Causal abstraction*
March 2024	Symposium on Causation, Complexity, Cognition, and Representation for Responsible AI (virtual event organized by Google and the Santa Fe Institute) ***Invited Speaker:** Counterfactual simulation in causal cognition*
Dec 2023	NeurIPS 2023 Workshop, New Orleans, USA ***Talk**: Social Contract AI: Aligning AI Assistants with Implicit Group Norms (presented by Philipp Fränken)* ***Poster**: Off The Rails: Procedural Dilemma Generation for Moral Reasoning (presented by Philipp Fränken)* ***Poster**: Anticipating the risks and benefits of counterfactual world simulation models (presented by Lara Kirfel)* ***Poster**: MoCa: Measuring human-language model alignment on causal and moral judgment tasks (presented by Allen Nie)*

Nov 2023 Causality in Minds and Machines Workshop, San Francisco, USA
***Talk**: Counterfactual simulation in causal cognition*

July 2023 Cognitive Science Conference (45th), Sydney, Australia
***Talk**: Father, don't forgive them, for they could have known what they're doing (presented by Lara Kirfel)*
***Talk**: Learning what matters: Causal abstraction in human inference (presented by Steven Shin)*
***Poster**: Looking into the past: Eye-tracking mental simulation in physical inference (presented by Ari Beller)*
***Poster**: A computational model of responsibility judgments from counterfactual (presented by Sarah Wu)*
***Poster**: Causal reasoning across agents and objects (presented by Bryan Gonzalez)*
***Poster**: You are what you're for: Essentialist categorization in large language models (presented by Siying Zhang)*
***Poster**: Teleology and generics (presented by Siying Zhang)*
***Poster**: Show and tell: Learning causal structures from observations and explanations (presented by Andrew Nam)*
***Poster**: Children use disagreement to infer what happened (presented by Jamie Amemiya)*
***Poster**: A semantics for causing, enabling, and preventing verbs using structural causal models (presented by Angela Cao)*

July 2023 CogSci 2023 Abstractions Workshop (virtual)
Invited Speaker: Causal abstraction

May 2023 Heuristics and Causality in the Social Sciences, London, UK
Invited Speaker: Inferring what happened through multi-modal mental simulation

Feb 2023 AAAI Conference on Artificial Intelligence, Washington DC, USA
Invited Speaker: *Bridge Program on Continual Causality*

Dec 2022 Neural Information Processing Systems (NeurIPS), New Orleans, USA
Invited Keynote Speaker: *Workshop on Neuro Causal and Symbolic AI (nCSI)*
Invited Keynote Speaker: *Gaze meets ML*
Invited Keynote Speaker: *Workshop on All Things Attention: Bridging Different Perspectives on Attention*
Invited Keynote Speaker: *Workshop on Deep Reinforcement Learning*

Sep 2022 XAI workshop, Birkbeck University, London, UK
***Invited Keynote Speaker**: Varieties of explanation*

Aug 2022 Computational Cognitive Neuroscience (CCN), San Francisco, USA
***Poster**: Looking into the past: Eye-tracking mental simulation in physical inference*

July 2022 Cognitive Science Conference (44th), Toronto, Canada
***Talk**: Stop, children what's that sound? Multi-modal inference through mental simulation (presented by Joseph Outa)*
***Poster**: Looking into the past: Eye-tracking mental simulation in physical inference (presented by Ari Beller)*
***Poster**: That was close! A counterfactual simulation model of causal judgments about decisions (presented by Sarah Wu)*
***Poster**: Inferences from disagreement (presented by Jamie Amemiya)*

July 2022 ICML Workshop "Beyond Bayes": Paths Towards Universal Reasoning Systems
***Talk:** MoCa: Cognitive Scaffolding for Language Models in Causal and Moral Judgment Tasks (presented by Allen Nie)*

July 2022 Society for Philosophy and Psychology Conference (SPP), Milan, Italy
Talk: *Looking into the past: Eye-tracking mental simulation in physical inference* (presented by Ari Beller)
Talk: *That was close! A counterfactual simulation model of causal judgments about decision*

June 2022 Robotics Science and Systems (RSS), "Social Intelligence in Humans and Robots" workshop
Talk: *That was close! A counterfactual simulation model of causal judgments about decisions (presented by Sarah Wu)*

May 2022 Annual Meeting of the Association for Computational Linguistics (ACL)
Invited Keynote Speaker: *Workshop on Commonsense Representation and Reasoning (CSRR).*

Dec 2021 NeurIPS (virtual)
Invited talk: *Workshop on Causal Inference and Machine Learning.*

July 2021 Cognitive Science Conference (43rd) (virtual)
Talk*: Who went fishing? Inferences from social evaluations (presented by Zach Davis).*
Poster*: In touch with causation: Understanding the impact of kinesthetic haptics on causality (presented by Elyse Chase).*
Poster*: The role of counterfactual reasoning in responsibility judgments (presented by Sarah Wu).*
Poster*: Eye-tracking multi-modal inference (presented by Ari Beller & Yingchen Xu).*

July 2021 Invited Panelist at ELLIS Workshop on Causethical ML.

July 2021 International Conference on Machine Learning (ICML)
Invited Talk: *Workshop on Algorithmic Recourse*.

July 2021 IEEE World Haptics Conference
Poster*: A causal feeling: How kinesthetic haptics affects causal perception (presented by Elyse D. Z. Chase).* — **Winner of the Best Work-in-Progress Paper Award**.

June 2021 Society for Philosophy and Psychology Conference (virtual)
Talk*: Inferences from explanation (presented by Lara Kirfel).*
Talk*: The role of counterfactual reasoning in responsibility judgments (presented by Sarah Wu).*
Poster*: The language of causation (presented by Aaron Beller).*
Poster*: Inference from social evaluations (presented by Zach Davis).*
Poster*: What explains causal judgments? Counterfactual versus hypothetical simulations*.

July 2020 Cognitive Science Conference (42nd), Toronto, Canada (virtual)
Poster*: Whom will Granny thank? Thinking about what could have been informs children's inferences about relative helpfulness (presented by Sophie Bridgers).*
Poster*: The language of causation (presented by Aaron Beller).*
Poster*: Learning from explanations (presented by Lara Kirfel).*

Apr 2020 ICLR Eighth International Conference on Learning Representations, Virtual Conference, Formerly Addis Ababa, Ethiopia
Invited Talk: *Workshop on Causal Learning for Decision Making*.

Sep 2019 Psychology and Economics of Causal Reasoning Conference, London, UK
Invited Talk*: "Understanding 'why': From Counterfactual Simulations to Explanations"*.

July 2019 Cognitive Science Conference (41st), Montreal, Canada
Talk*: Explaining intuitive difficulty judgments by modeling physical effort and risk (presented by Ilker Yildirim).*
Poster*: The trajectory of counterfactual simulation in development (presented by Jonathan Kominsky).*

July 2019 Society for Philosophy and Psychology (SPP), San Diego, California, USA
***Poster**: What happened? Reconstructing the past through vision and sound.*

Mar 2019 AAAI symposium: Beyond curve fitting – Causation, counterfactuals, and imagination-based AI, Stanford, USA
***Invited Talk**: Understanding "why" – Causation, Counterfactuals, and Imagination.*

Oct 2018 Mental Simulation Workshop, UC Merced, USA
***Invited Talk**: What happened? Reconstructing the past from vision and sound.*

July 2018 Cognitive Science Conference (40th), Madison, Wisconsin, USA
***Talk**: What happened? Reconstructing the past from vision and sound.*
***Poster**: Tiptoeing around it: Inference from absence in potentially offensive speech (presented by Monica Gates).*
***Poster**: Moral dynamics (presented by Felix Sosa).*

July 2018 Society for Philosophy and Psychology (SPP), Ann Arbor, Michigan, USA
Talk**: Judgments of actual causation approximate the effectiveness of interventions (presented by Adam Morris) — **William James Prize for best contributed paper by a graduate student

May 2018 MIT, Boston, USA
***Flash Talk**: What Happened? Reconstructing the Past Through Vision and Sound*

Jan 2018 Causal pluralism: a multi-disciplinary investigation of causality in philosophy and the sciences, Cambridge, USA
***Invited Talk**: A counterfactual simulation model of causal judgment.*

July 2017 Cognitive Science Conference (39th), London, UK
***Talk**: Faulty Towers: A hypothetical simulation model of physical support.*
***Talk**: Causal learning from interventions and dynamics in continuous time (presented by Neil Bramley).*
***Talk**: Marbles in inaction: Counterfactual simulation and causation by omission (presented by Simon Stephan).*
***Poster**: Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints.*
***Poster**: Eye movement-based probabilistic models for physical scene understanding (presented by Eghbal Hosseini).*

July 2017 Morality, Language, and Thought Workshop, Paris, France
***Talk**: A counterfactual simulation model of causal judgment.*

June 2017 Society for Philosophy and Psychology (SPP), Baltimore, USA
***Talk**: A counterfactual simulation model of causation by omission.*

June 2017 Time and causality in the sciences, Hoboken, USA
***Talk**: A counterfactual simulation model of causal judgment.*

May 2017 Brains, Minds, and Machines Workshop, San Juan, Puerto Rico
***Talk**: Understanding why: From counterfactual simulations to responsibility judgments.*

Nov 2016 Society for Judgment and Decision Making, Boston, USA
***Talk**: Lucky or clever? From changed expectations to responsibility judgments.*

Aug 2016 Cognitive Science Conference (38th), Philadelphia, USA
***Talk in workshop**: The role of space, time, and causality in evaluating counterfactual closeness.*
***Talk**: Understanding "almost": Empirical and computational studies of near misses.*
***Talk**: Natural science: Active learning in dynamic physical microworlds (presented by Neil Bramley).*
***Talk**: Implicit measurement of motivated causal attribution (presented by Laura Niemi).*

Aug 2016 International Conference on Thinking, Providence, USA
***Organized symposium**: Holding others responsible.*
***Talk in contributed symposium**: From changed expectations to attributions of responsibility.*

June 2016 Society for Philosophy and Psychology (SPP), Austin, USA
***Talk in invited symposium**: A counterfactual simulation model of causal judgments.*

Nov 2015 Society for Judgment and Decision Making (SJDM), Chicago, USA
***Poster**: A counterfactual simulation model of causal judgments.*

Nov 2015 Psychonomics, Chicago, USA
***Poster**: Inference of Intention and Permissibility in Moral Judgment (presented by Max Kleiman-Weiner). Nominated for best student poster award.*

Aug 2015 MIT, Boston, USA
Flash Talk: *Eye-tracking Causality: A Counterfactual Simulation Model of Causal Judgments*

July 2015 Multidisciplinary University Research Initiative (MURI) meeting, Lake Arrowhead, USA
***Talk**: A counterfactual simulation model of causal judgments.*

July 2015 Cognitive Science Conference (37th), Pasadena, USA
***Talk in workshop**: A counterfactual simulation model of causal judgments.*
***Talk**: How, whether, why: Causal judgments as counterfactual contrasts.*
***Talk**: Responsibility judgments in voting scenarios.*
***Talk**: Inference of intention and permissibility in moral decision making (presented by Max Kleiman-Weiner).*
***Poster**: Go fishing! Responsibility judgments when cooperation breaks down (presented by Kelsey Allen).*

Jan 2015 Ranch Metaphysics Workshop, Arizona, USA
***Invited Talk**: Intuitive theories.*

Aug 2014 Modality Workshop Yale, New Haven, USA
***Invited Talk**: From counterfactual simulation to causal judgment.*

July 2014 Cognitive Science Conference (36th), Quebec City, Canada
***Talk**: From counterfactual simulation to causal judgment.*
***Poster**: Responsibility attributions as counterfactual replacements.*
***Talk**: The order of things (co-author).*
***Talk**: Causal supersession (co-author).*

June 2014 Society for Philosophy and Psychology, Vancouver, Canada
***Talk**: From counterfactual simulation to causal judgment.*

Nov 2013 Roundtable on Causal Overdetermination, Urbana-Champaign, USA
***Invited Talk**: Causal responsibility and counterfactuals.*

Nov 2013 New Orleans Workshop on Agency and Responsibility, USA
***Talk (joint with David Lagnado)**: A difference-making framework for intuitive judgments of responsibility.*

Aug 2013 SPUDM conference (24th), Barcelona, Spain
***Talk in symposium**: How intentions and deception influence responsibility attributions in groups.*
***Talk in symposium**: A counterfactual model of responsibility attributions in groups.*

Aug 2013 Cognitive Science Conference (35th), Berlin, Germany
***Poster**: Back on track – Backtracking in counterfactual reasoning.*

July 2013 London Reasoning Workshop, London, UK
***Talk**: Back on track – Backtracking in counterfactual reasoning.*

June 2013 PostDocs Share Their Science Event, Boston, USA
***Poster**: General principles of causal attribution.*

June 2013 Society of Philosophy and Psychology Conference, Providence, USA

Aug 2012 Cognitive Science Conference (34th), Sapporo, Japan
***Talk**: Noisy Netwons: Unifying process and dependency accounts of causal attribution.*
***Poster**: Ping Pong in Church: Productive use of concepts in human probabilistic inference.*
***Poster**: Why blame Bob? Probabilistic generative models for counterfactual reasoning and blame attribution (presented by John McCoy and Tomer Ullman).*

July 2012 International Conference on Thinking, University College London, UK
***Invited talk in symposium**: Finding fault: Causality and Counterfactuals in Group Attributions.*
***Talk**: Noisy Netwons: Unifying process and dependency accounts of causal attribution.*

May 2012 'Towards Consilience' Workshop, London Business School, UK

May 2012 'Personal and Shared Intentions' Workshop, MPI Berlin, Germany
***Invited talk**: How intentions and deception influence attributions of responsibility in groups.*

Apr 2012 Postgraduate Peer Group Conference, Cumberland Lodge, UK
***Talk**: Noisy Newtons - People's intuitive understanding of physics explains their causal attributions; **1st prize** in the oral presentation competition.*

Apr 2012 Tagung experimentell arbeitender Psychologen (TEAP), Mannheim, Germany
***Talk**: Noisy Newtons: Unifying process and dependency accounts of causal attribution.*

Jan 2012 CEU Conference on Cognitive Development, Budapest, Hungary
***Invited talk in symposium**: Noisy Newtons: People's intuitive understanding of physics explains their cause and prevention judgments.*

Sep 2011 Experimental Philosophy Workshop, Sheffield, England

Aug 2011 SPUDM conference (23rd), Kingston, England
***Talk**: On perceived criticality and expected effort.*
***Talk**: Rational order effects in responsibility attributions.*

July 2011 Cognitive Science Conference (33rd), Boston, USA
***Talk**: Rational order effects in responsibility attributions.*
***Poster**: Beyond outcomes: The influence of intentions and deception.*
***Poster**: Blame the skilled.*

July 2011 European Association of Social Psychology Conference, Stockholm, Sweden
***Talk in symposium**: Rational order effects in responsibility attributions.*

Mar 2011	International Conference on Behavioral Decision Making, Israel ***Poster**: Responsibility Attributions in Teams; **2nd prize** in the conference poster competition (presented by Ro'i Zultan).*
Oct 2010	KogWis, Potsdam, Germany ***Poster**: The Allocation of Responsibility amongst Multiple Agents.*
Sep 2010	Actual Causation Workshop, Konstanz, Germany
Aug 2010	Cognitive Science Conference (32nd), Portland, Oregon ***Poster**: The Dice are Cast: The Influence of Intended versus Actual Contributions on Responsibility Attribution.*
Aug 2010	Mathematical Psychology Conference, Portland, Oregon
June 2010	AXA Talent Day, Paris, France ***Poster**: The Dice are Cast: The Influence of Intended versus Actual Contributions on Responsibility Attribution.*
May 2010	Cognitive Science and Machine Learning Summer School, Sardinia, Italy
Apr 2010	EPS/SEPEX Joint Conference, Granada, Spain ***Talk**: The Dice are Cast.*
Aug 2009	SPUDM Conference (22nd), Rovereto, Italy ***Poster**: Allocation of Responsibility amongst Multiple Agents.*
July 2009	Summer Institute, Max Planck Institute Berlin, Germany ***Poster**: Allocation of Responsibility amongst Multiple Agents; **2nd prize** in the **poster** competition.*
June 2009	MICRAC Workshop, Toulouse, France ***Talk**: Allocation of Responsibility amongst Multiple Agents.*
July 2008	PsyPag Conference, Manchester, England ***Talk**: Presentation of MSc thesis (honorable mention).*

Invited Presentations

2025	Psychology & Neuroscience Colloquium Series	Boston College, USA
2024	Psychology Colloquium	Göttingen University, Germany
	Symbolic Systems Forum	Stanford University, USA
2023	Cognitive Salon	San Francisco, USA
	Social seminar	Department of Psychological and Brain Sciences, University of Delaware, USA
	Deepmind	London, UK
	Psychology Colloquium	University of California Santa Cruz, USA
2022	Cognitive Forum	University of California, Los Angeles, USA
	Institute of Cognitive and Brain Sciences	University of California, Berkeley, USA

	Social lunch	Stanford University, USA
	Stanford Vision and Learning Lab	Stanford University, USA
	COCOA: Converging On Causal Ontology Analyses	Virtual seminar
	CEU Department of Cognitive Science colloquium	Vienna, Austria (virtual)
	Presentation at author-meets-critics symposium for Jim Woodward's new book "Causation with a human face"	Center for Philosophy of Science, Pittsburgh University, USA
	Construction of meaning workshop	Stanford University, USA
2021	Human-Centered Artificial Intelligence Seminar	Stanford University, USA
	Computation, Cognition, and Development Lab (Prof. Tomer Ullman)	Harvard University, USA
	Biological and Behavioral, Computational and Critical Data Initiative (B2C2)	Arizona State University, USA
	Computational Cognitive Science Colloquium	TU Darmstadt, Germany
	LINGUAE Seminar (Prof. Philippe Schlenker)	Institut Nicod Paris, France
	Schuck Lab (Dr. Nicolas Schuck)	Max Planck Institute Berlin, Germany
	Computation, Cognition, and Development Lab (Prof. Tomer Ullman)	Harvard University, USA
2020	Decision Making and Economic Psychology Seminar	Ben-Gurion University, Israel
	Developmental Brownbag	UC San Diego, USA
	Computational Approaches to Social Cognition Talk Series	Harvard University, USA
	Symbolic Systems Forum	Stanford University, USA
	Computational Cognitive Science Group (Prof. Neil Bramley)	Edinburgh University, UK
2019	Robotics Lunch	Stanford University, USA
	Seminar on History and Philosophy of Science	California Institute of Technology, USA
	Cognition & Language seminar	Stanford University, USA
	Mind, Brain, Computation, and Technology seminar	Stanford University, USA
2018	FriSem (Cognitive & Neuroscience seminar)	Stanford, USA
	MIT Brain and Cognitive Sciences Retreat	Newport, USA
	Visual Attention Lab	Brigham & Women's Hospital, USA
	Northeastern Cognition Area Speaker Series	Northeastern University, USA
	The Imagination and Modal Cognition Lab (Prof. Felipe De Brigard)	Duke University, USA
2017	Cognitive Lunch	Yale, USA
	Computational Social Cognition boot camp	Harvard, USA

	ConCats seminar series	NYU, USA
	Language Learning Lab (Prof. Joshua Hartshorne)	Boston College, USA
	Psychology Department Colloquium	Berkeley, USA
	Psychology Department Colloquium	Stanford University, USA
	Cushman & Greene lab meeting	Harvard University, USA
	Behavioral Science Seminar	The University of Chicago Booth School of Business, USA
2016	Computational Cognitive Neuroscience Lab (Prof. Sam Gershman)	Harvard University, USA
	Applied Statistics Workshop	Harvard University, USA
	Program for Evolutionary Dynamics	Harvard University, USA
	Center for Adaptive Rationality	Max Planck Institute Berlin, Germany
	Language and Cognition seminar series	Harvard University, USA
	Cognition seminar series	Brown University, USA
2015	Crockett lab (Prof. Molly Crockett)	Oxford University, UK
	London Judgment and Decision Making Seminar	University College London, UK
	Office of Naval Research	Washington, DC, USA
2014	Boston Moral Reasoning Group	Boston University, USA
	Social Lunch	Harvard University, Cambridge, USA
	Adaptive Behavior and Cognition Research Group (Prof. Gerd Gigerenzer)	Max Planck Institute, Berlin, Germany
	Computation and Cognition Lab (Prof. Noah Goodman)	Stanford University, USA
	Lab meeting (Prof. Joshua Knobe)	Yale University, USA
2013	Moral Cognition Lab (Prof. Joshua Greene)	Harvard University, USA
	Morality Lab (Prof. Liane Young)	Boston College, USA
	Sloman Lab (Prof. Steven Sloman)	Brown University, USA
	Early Childhood Cognition Lab (Prof. Laura Schulz)	Massachusetts Institute of Technology, USA
	Computational Cognitive Science Lab (Prof. Josh Tenenbaum)	Massachusetts Institute of Technology, USA
2012	Invited Speakers Series	University of Leicester, England
	Special Seminar (Dr. Christopher Summerfield)	Oxford University, England
	Lab meeting (Dr. Ralf Mayrhofer)	University Göttingen, Germany
2011	Computational Cognitive Science Group (Prof. Josh Tenenbaum)	Massachusetts Institute of Technology, Boston, USA

	CogLunch of the Cognitive Science Department	Massachusetts Institute of Technology, Boston, USA
	Lab meeting (Prof. Noah Goodman)	Stanford University, Palo Alto, USA
2010	Research seminar series	Decision Technology, London, England
	London Judgement and Decision Making Group	University College London, England
	Strategic Interaction group seminar (Prof. Werner Güth)	Max Planck Institute for Economics, Jena, Germany
2009	Cognitive and Decision Sciences seminar (Prof. Michael Waldmann)	University Göttingen, Germany
	Philosophy of Mind seminar (Prof. Michael Pauen)	Humboldt University, Berlin, Germany

Public outreach

We share my lab's work via twitter and youtube.

1) Press

2024 Humans Use Counterfactuals to Reason About Causality. Can AI? – Press release on the paper "Counterfactual simulation in causal cognition".

Children's ability to detect ambiguity in disagreements sharpens between ages 7 and 11 – News article about the paper "Children use disagreement to infer what happened".

2022 'Worse' AI Counterintuitively Enhances Human Decision Making and Performance – Press release about the paper "Uncalibrated models can improve Human-AI collaboration".

2021 Modeling How People Make Causal Judgments – Press release on the paper "A counterfactual simulation model of causal judgments for physical events".

2018 Quantifying How We Intuit the Physical World – Medium blog post about the *Intuitive experimentation in the physical world* paper.

2017 How we determine blame – Article on MIT News covering our *Eye-tracking causality* paper.

Wo das essen nicht kalt wird – Article in Deutsche Universitätszeitung (DUZ) on what working at MIT is like.

2012 Are you responsible for the outcome of the election? – Blog entry written by Tania Lombrozo which covers the "When contributions make a difference: Explaining order effects in responsibility attributions" paper.

2010 AXA Talent Day – Video created in the context of a talent day organized by AXA in Paris.

Who is to blame when groups succeed or fail? – Blog entry written by Art Markman in *Psychology Today* summarizing the "Spreading the blame: The allocation of responsibility amongst multiple agents" paper.

2) Blog Posts & Podcasts

2022 The Brains Blog: "Anything goes! From a pluralism of methods towards a unified theory of causal cognition"

2021 Stanford Psychology Podcast on "Whose fault is it? Causal judgments in everyday life"

3) Diversity and Inclusion

You can find a statement of our lab's research values here.

2018–2024 Participation in annual "Paths to PhD" workshop at Stanford: Orientation program for underrepresented students

2019–2022 Participation in Stanford's CSLI summer internship program.

2021, 2023, 2024 Speaker at Annual Future Advancers of Science and Technology (FAST) Symposium at Stanford.

2017 Tutorials at Brains, Minds, and Machines Workshop in San Juan, Puerto Rico

Teaching

2025 PSYCH 198: Advanced Research (Psychology Undergraduate Honors Class)

PSYCH 252: Statistical Methods for Behavioral and Social Sciences (Graduate statistics course)

2024 PSYCH 198: Advanced Research (Psychology Undergraduate Honors Class)

PSYCH 252: Statistical Methods for Behavioral and Social Sciences (Graduate statistics course)

2023 No teaching (Junior Faculty Leave)

2022 PSYCH 198: Advanced Research (Psychology Undergraduate Honors Class)

PSYCH 252: Statistical Methods for Behavioral and Social Sciences (Graduate statistics course)

2021 PSYCH 252: Statistical Methods for Behavioral and Social Sciences (Graduate statistics course)

Guest lecture on "Causal judgments and inference" at Dartmouth College in Prof. Jonathan Phillip's class

Guest lecture on "Intuitive theories" at University College London in Dr. Constantin Rezlescu's class

2020 PSYCH 293: What makes a good explanation? Psychological and philosophical perspectives (co-taught with Prof. Thomas Icard) (Graduate seminar)

PSYCH 251: Experimental methods (co-taught with Prof. Michael Frank) (Graduate methods course)

PSYCH 187: Research Methods in Cognition and Development (co-taught with Prof. Hyo Gweon) (Advanced undergraduate seminar)

PSYCH 252: Statistical Methods for Behavioral and Social Sciences (Graduate statistics course)

2019 Guest lecture in Symbolic Systems 1 on "Intuitive Theories".

PSYCH 291: Causal Cognition (Graduate seminar)

PSYCH 252: Statistical Methods for Behavioral and Social Sciences (Graduate statistics course)

2017 Tutorial on "Mental models as probabilistic programs" in class 9.66 (Computational Cognitive Science taught by Prof. Josh Tenenbaum).

Tutorial on "Mental models as probabilistic programs" at the Brains, Minds, and Machines summer school in Woods Hole. Teaching assistant for three student projects.

Tutorial on "Mental models as probabilistic programs" at the Brains, Minds, and Machines workshop in San Juan, Puerto Rico.

2016 Lecture on "Counterfactual simulation" in class 9.66 (Computational Cognitive Science taught by Prof. Josh Tenenbaum).

2015 Kaufman Teaching Certificate Program.

2012 UCL MSc lecture: Theoretical frameworks of responsibility attribution.

UCL 1st year undergraduate psychology lab class: Causal attribution.

2011 UCL 1st year undergraduate psychology lab class: Responsibility attribution.

2010 UCL teaching demonstratorship.

Mentorship

2024 Yang Xiang (Visiting PhD student): "A signaling model of self-handicapping"

Alexa Tran (Psych summer intern): "Action abstraction"

Sunny Yu (Symbolic Systems Summer intern): "Children's books project"

Anujin Naranbaatar: (Summer fellows program) "Action abstraction"

Tuvana Soronzonbold (Summer fellows program): "Action abstraction"

Hannah Cha: "Whodunit"

Mishika Govil: "Action abstraction"

Chuqi Hu: "Causal abstraction"

Verona Teo (Pre-doc): "Whodunit"

2023 Sam Kwok (Psychology summer intern): "Social contract AI"

Patrick Ye (Symbolic Systems summer intern): "Social contract AI"

Shruti Sridhar (Symbolic Systems summer intern): "Simulated agents"

Cindy Xin: "Explanation valence"

Ayesha Khawaja: "Off the rails"

Haoran Zhao: "Pragmatic reasoning in LLMs"

Sunny Yu: "Children's books project"

Philip Miao: "Children's books project"

Emily Jin: "Marple: Figuring out what happened through multi-modal inference"

Hannah Cha: "Marple: Figuring out what happened through multi-modal inference"

Adam Chun: "Spot the ball"

2022 Xenia Bunk (3 month visit): "Willful ignorance"

Jeong Yeon Shin: "Explanation valence"

Selena She (CSLI summer intern): "Are large language models teleological essentialists?"

Siying Zhang: "The development of counterfactual reasoning"

Damini Kusum: "Mental rotation in people with Aphantasia"

Shruti Sridhar (CSLI summer intern): "Simulated agents"

2021 Adam Huang: "A causal conjunction fallacy in intuitive physical predictions"

Addison Jadwin: "The role of anticipated blame in action selection"

Joseph Outa: "The development of multi-modal integration"

Yingchen Xu: "Modeling multi-modal inference"

Yanal Ramzi Qushair (SymSys summer intern): "Mental simulation in people with aphantasia"

Mansi Verma (CSLI summer intern): "Using omissions as evidence for early counterfactual reasoning"

Shruti Sridhar (Psychology summer intern): "Counterfactual simulation with agents"

Ricky Ma: "Counterfactual inference"

Laila Johnston: "Productive use of concepts in human probabilistic inference"

Yixiu Zhao (Stanford Interdisciplinary Graduate Fellow): "Abstraction in human inference"

2020 Bryce Lynford: "Productive use of concepts in human probabilistic inference"

Liang Zhou: "A hypothetical simulation model of causal support"

Disha Dasgupta: "Resolving ambiguities in counterfactual reasoning"

Xi Jia (Laura) Zhou: "The development of multi-modal integration"

Tina Hua: "Eye-tacking motivated reasoning"

2019 Antonia Langenhoff (3 month visit): "Predicting responsibility judgments from dispositional inferences and causal attributions"

Lara Kirfel (3 month visit): "Learning from language: Inferences about norms and causal structure from explanations"

Ross Kempner (CSLI summer intern): "Does motivation influence causal judgments via biased counterfactual simulations?"

Alan Brown (Research assistant): "Eye-tracking causality: Reconstructing the past"

Jingren Wang (Summer intern): "Hypotheticals, counterfactuals, and causal judgments"

University service

2022 - present I help run a group in the psychology department that meets once a week to meditate and drink tea together afterwards.

2024 Chairperson at thesis defense (Gerry Wan, Computer Science)

Dissertation defense committe member (Marianna Zhang, Psychology)

Dissertation defense committe member (Kayla Good, Psychology)

	Thesis committe member (Andrew Nam, Psychology)
	Thesis committee member (Kate Petrova, Psychology)
	Thesis committee member (Peter Zhu, Psychology)
2023	Chairperson at thesis defense (Duligur Ibeling, Computer Science)
	Thesis committe member (Nicky Sullivan, Psychology)
2022	Undergraduate Advisor (Charles Hicks, Symbolic Systems)
	Undergraduate Advisor (Addison Reese Jadwin, Symbolic Systems)
	Thesis committee member (Effie Li, Psychology)
	Reader of first year project (Kate Petrova, Psychology)
	Chairperson at thesis defense (Negin Heravi, Computer Science)
	Thesis committee member (Bryan Gonzalez, Psychology, Dartmouth University)
2021	Chairperson at thesis defense (Omer Korat, Linguistics)
	Reader of masters thesis (Jack Beasley, Symbolic Systems)
	Thesis committe member (Marianna Zhang, Psychology)
	Thesis committe member (Elyse Chase, Mechanical Engineering)
2020	Chairperson at thesis defense (Evan Strasnick, Computer Science)
	Chairperson at thesis defense (Francesca Zaffora Blando, Philosophy)
	Chairperson at thesis defense (Sebastian Schuster, Linguistics)
	Chairperson at thesis defense (Reuben Cohn, Linguistics)
	Reader of PhD thesis (Erin Bennett, Psychology)
	Committee member in PhD thesis defense (Sophie Bridges, Psychology)
	Committee member in PhD thesis defense (Andrew Kyle Lampinen, Psychology)
	Committee member in qualifying project (Elisa Kreiss, Linguistics)
	Reader of first year project (Effie Li, Psychology)
2019	Chairperson at thesis defense (Prerna Nadathur, Linguistics)
	Reader of first year project (Andrew Nam, Psychology)

Reviews

Since	Journal / Conference / Grant	Reviews
Sep 2024	Psychological Bulletin	|
Apr 2023	Management Science	|
Sep 2022	Open Mind: Discoveries in Cognitive Science	|
June 2022	Psychological Review	|||
Apr 2022	Journal of Personality and Social Psychology	|
May 2021	Child Development	||
Oct 2020	Social Psychological and Personality Science	|

Aug 2020	Nature Computational Science	1
May 2019	Journal of Experimental Psychology: General	4
Apr 2019	Israel Science Foundation	1
Jan 2019	Panelist at NSF panel: Perception, Action, & Cognition	
Aug 2018	Trends in Cognitive Science	1
May 2018	Proceedings of the National Academy of Sciences	1
May 2018	KogWis	3
Mar 2018	Grant referee for National Science Foundation (NSF) proposal	2
Feb 2018	Psychological Science	2
Mar 2017	Cognitive Psychology	3
Mar 2017	Nature Human Behavior	1
Dec 2016	Grant referee for the European Research Council (ERC)	2
Sep 2016	Topics in Cognitive Science	1
Aug 2016	Perspectives on Psychological Science	1
Sep 2015	Social Psychology	1
July 2015	Learning and Motivation	1
May 2015	The Oxford Handbook of Causal Reasoning	1
Feb 2015	Society for Philosophy and Psychology Conference	7 2
Dec 2014	PLoS ONE	1
June 2014	Journal of Experimental Psychology: Learning, Memory & Cognition	3
Oct 2013	Scandinavian Journal of Psychology	1
Aug 2013	Thinking and Reasoning	1
Jan 2013	Journal of Cognitive Psychology	1
Jan 2013	Cognitive Science	7 7 1
Nov 2012	Topoi – An International Review of Philosophy	1
Nov 2012	Journal of Experimental Social Psychology	1
Oct 2012	Oxford Studies of Experimental Philosophy	1
Feb 2012	Quaterly Journal of Experimental Psychology	1
Oct 2011	Psychonomic Bulletin and Review	3
June 2011	Cognition	7 7 7 7 1
Jan 2011	Cognitive Science Conference Proceedings	7 7 7 7 7 7 7

Editorial positions

Since	Conference	Manuscripts
2022	Associate Editor at Open Mind: Discoveries in Cognitive Science	2
2019	Meta-reviewer at Cognitive Science Conference Proceedings	7 7 7 7

Skills & Interests

Languages German (native), English (fluent)

Software R, Python, Javascript, LaTeX

Interests Surfing, Beachvolleyball, Chess, Cycling, Yoga, Music, Piano, Guitar, Meditation