# Simon Goldstein

The University of Hong Kong
Department of Philosophy
simon.d.goldstein@gmail.com
simondgoldstein.com
*Updated July 2026*

## Employment

- Associate Professor, The University of Hong Kong, Department of Philosophy, 2023–present.
- Visiting Senior Scholar, Forethought, 2026–present.
- Senior Editor, AI Frontiers, 2025–present.
- Research Fellow, Center for AI Safety, 2023.
- Associate Professor, Australian Catholic University, Dianoia Institute, 2019–2023.
- Assistant Professor, Lingnan University, Department of Philosophy, 2017–2019.

## Education

**Rutgers University, Ph.D. Philosophy, 2011–2017**

- Dissertation: Informative Dynamic Semantics (advisors: Thony Gillies and Ernie Lepore).
- Certificate in Cognitive Science.

**Yale University, B.A. Philosophy (Honors), 2007–2011**

## Areas of Specialization

Philosophy of AI, Epistemology, Philosophy of Language

## Publications

### *Books*

- Goldstein, S., & Salib, P. (forthcoming). *AI Rights*. Cambridge University Press (Cambridge Elements).
- Goldstein, S., & Kirk-Giannini, C. D. (forthcoming). *AI Welfare: Agency, Consciousness, Sentience*. Oxford University Press.
- Goldstein, S. (2024). *Iterated Knowledge*. Oxford University Press.

  *Symposium in Inquiry: Jeremy Goodman, Daniel Greco, Eliran Haziza, Ben Holguín, Sven Rosenkranz, Bernhard Salow.*

### *Articles*

- Wang, S., Lobanova, S., Arbel, Y., Goldstein, S., & Salib, P. (forthcoming). AI revealed preferences. AIES 2026. https://ssrn.com/abstract=6798118
- Arbel, Y., Goldstein, S., & Salib, P. (forthcoming). How to count AIs: Individuation and liability for AI agents. *Boston College Law Review*. https://ssrn.com/abstract=6273198

- Goldstein, S., Krishnamurthi, G., & Salib, P. (forthcoming). AI suffrage for human flourishing. *Fordham Law Review*.
- Goldstein, S., & Lederman, H. (forthcoming). AI death. *Philosophical Perspectives*. https://philarchive.org/rec/GOLADC
- Goldstein, S., & Kirk-Giannini, C. D. (forthcoming). A case for AI consciousness: Language agents and global workspace theory. *Journal of Consciousness Studies*.
- Goldstein, S., & Levinstein, B. A. (forthcoming). Does ChatGPT have a mind? *Philosophy of AI*.
- Blumberg, K., & Goldstein, S. (forthcoming). A semantic theory of redundancy. *Linguistics and Philosophy*.
- Goldstein, S., & Hawthorne, J. (forthcoming). Preface knowledge. In *Themes from Williamson*. De Gruyter.
- Salib, P., & Goldstein, S. (2026). AI rights for human safety. *Virginia Law Review*, *112*, 1061.
- Goldstein, S., & Salib, P. (2026). AI is not a natural monopoly. *Minnesota Law Review Headnotes*, *110*. https://ssrn.com/abstract=5926043
- Goldstein, S. (2026). Omega knowledge matters. In *Oxford Studies in Epistemology* (Vol. 8, pp. 63–84). Oxford University Press. Honorable Mention, Sanders Prize in Epistemology.
- Goldstein, S. (2026). Will AI and humanity go to war? *AI & Society*, *41*(1), 321–334.
- Cappelen, H., Goldstein, S., & Hawthorne, J. (2025). AI survival stories: A taxonomic analysis of AI existential risk. *Philosophy of AI*, *1*(1), 1–19.

  *Article symposium in Philosophy of AI, with response: Leonard Dung, Aksel Sterri and Peder Skjelbred, Rory Svarc, Kate Vredenburgh.*
- Goldstein, S., & Kirk-Giannini, C. D. (2025). Language agents reduce the risk of existential catastrophe. *AI & Society*, *40*(2), 959–969.
- Goldstein, S., & Kirk-Giannini, C. D. (2025). AI wellbeing. *Asian Journal of Philosophy*, *4*(1), 25.

  *Article symposium in Asian Journal of Philosophy: Adam Bradley, James Fanciullo, Jiwon Kim, Jeff Sebo, Walter Veit.*
- Goldstein, S., & Hawthorne, J. (2025). KK is wrong because we say so. *Mind*, *134*(533), 33–59.
- Goldstein, S., & Robinson, P. (2025). Shutdown-seeking AI. *Philosophical Studies*, *182*(7), 1567–1579.
- Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, *5*(5).
- Goldstein, S., & Hawthorne, J. (2024). Safety, closure, and extended methods. *Journal of Philosophy*, *121*(1), 26–54.
- Blumberg, K., & Goldstein, S. (2023). Attitude verbs' local context. *Linguistics and Philosophy*, *46*(3), 483–507.
- Beddor, B., & Goldstein, S. (2023). A question-sensitive theory of intention. *The Philosophical Quarterly*, *73*(2), 346–378.
- Carter, S., & Goldstein, S. (2023). Getting accurate about knowledge. *Mind*, *132*(525), 158–191.
- Goldstein, S., & Kirk-Giannini, C. D. (2022). Contextology. *Philosophical Studies*, *179*(11), 3187–3209.
- Goldstein, S. (2022). Sly Pete in dynamic semantics. *Journal of Philosophical Logic*, *51*(5), 1103–1117.

- Goldstein, S., & Hawthorne, J. (2022). Counterfactual contamination. *Australasian Journal of Philosophy*, *100*(2), 262–278.
- Goldstein, S., & Hawthorne, J. (2022). Knowledge from multiple experiences. *Philosophical Studies*, *179*(4), 1341–1372.
- Goldstein, S. (2022). Fragile knowledge. *Mind*, *131*(522), 487–515.
- Goldstein, S., & Waxman, D. (2021). Losing confidence in luminosity. *Noûs*, *55*(4), 962–991.
- Goldstein, S. (2021). Epistemic modal credence. *Philosophers' Imprint*, *21*(26), 1–24.
- Carter, S., & Goldstein, S. (2021). The normality of error. *Philosophical Studies*, *178*(8), 2509–2533.
- Beddor, B., & Goldstein, S. (2021). Mighty knowledge. *Journal of Philosophy*, *118*(5), 229–269.
- Goldstein, S., & Santorio, P. (2021). Probability for epistemic modalities. *Philosophers' Imprint*, *21*(33), 1–37.
- Cariani, F., & Goldstein, S. (2020). Conditional heresies. *Philosophy and Phenomenological Research*, *101*(2), 251–282.
- Goldstein, S. (2020). Free choice impossibility results. *Journal of Philosophical Logic*, *49*(2), 249–282.
- Goldstein, S. (2020). The counterfactual direct argument. *Linguistics and Philosophy*, *43*(2), 193–232.
- Goldstein, S. (2019). Free choice and homogeneity. *Semantics and Pragmatics*, *12*(23).
- Goldstein, S. (2019). Generalized update semantics. *Mind*, *128*(511), 795–835.
- Goldstein, S. (2019). Triviality results for probabilistic modals. *Philosophy and Phenomenological Research*, *99*(1), 188–222.
- Goldstein, S. (2019). A theory of conditional assertion. *The Journal of Philosophy*, *116*(6), 293–318.
- Bronner, B., & Goldstein, S. (2018). A stronger doctrine of double effect. *Australasian Journal of Philosophy*, *96*(4), 793–805.
- Beddor, B., & Goldstein, S. (2018). Believing epistemic contradictions. *The Review of Symbolic Logic*, *11*(1), 87–114.
- Goldstein, S. (2016). A preface paradox for intention. *Philosophers' Imprint*, *16*(14), 1–20.

### *Under Review*

- Goldstein, S., & Salib, P. AI rights for economic flourishing. https://ssrn.com/abstract=5353214
- Goldstein, S., & Salib, P. How to stop an AI arms race. https://ssrn.com/abstract=5369439
- Goldstein, S., & Salib, P. A thousand AI constitutions. https://philpapers.org/rec/GOLATA-3
- Goldstein, S., & Salib, P. Liberalism forever. https://philpapers.org/rec/GOLLFC-2
- Goldstein, S. LLMs cannot be ideally rational. https://philpapers.org/rec/GOLLCN
- Goldstein, S., & Kirk-Giannini, C. D. The polarity problem.
- Goldstein, S., & Lederman, H. What does ChatGPT want? An interpretationist guide. https://philpapers.org/rec/GOLWDC-2
- Goldstein, S., & Hong, F. AI safety inside the simulation.
- Goldstein, S., & Hong, F. Recurrence, rational choice, and the simulation hypothesis.
- Fanciullo, J., & Goldstein, S. Some paradoxes of desire and pleasure.

- Fanciullo, J., Goldstein, S., & Pallies, D. How to count desires: Welfare levels and desire satisfactionism.
- Goldstein, S., & Yip, B. AI and the contagion of disrespect.

### *Public Writing*

- Arbel, Y., Goldstein, S., & Salib, P. Why law needs a new entity to govern AI agents. (June 15, 2026). *Columbia Law School Blue Sky Blog*. https://clsbluesky.law.columbia.edu/2026/06/15/why-law-needs-a-new-entity-to-govern-ai-agents/
- Goldstein, S., & Salib, P. Copyright should not protect artists from artificial intelligence. (October 23, 2025). *Lawfare*. https://www.lawfaremedia.org/article/copyright-should-not-protect-artists-from-artificial-intelligence
- Goldstein, S., & Lederman, H. Claude's right to die? The moral error in Anthropic's end-chat policy. (October 17, 2025). Lawfare.https://www.lawfaremedia.org/article/claude-s-right-to-die--the-moral-error-in-anthropic-s-end-chat-policy Selected by The Browser.
- Goldstein, S., & Salib, P. Today's AIs aren't paperclip maximizers. That doesn't mean they aren't risky. (May 21, 2025). *AI Frontiers*. https://ai-frontiers.org/articles/todays-ais-arent-paperclip-maximizers
- Goldstein, S., & Salib, P. The case for a joint U.S.-China AI lab. (April 4, 2025). *Lawfare*. https://www.lawfaremedia.org/article/the-case-for-a-joint-u.s.-china-ai-lab
- Goldstein, S., & Salib, P. Nuclear deterrence in the age of AGI. (March 26, 2025). *Lawfare*. https://www.lawfaremedia.org/article/nuclear-deterrence-in-the-age-of-agi
- Goldstein, S., & Salib, P. DeepSeek points towards U.S.-China cooperation, not a race. (March 5, 2025). *Lawfare*. https://www.lawfaremedia.org/article/deepseek-points-toward-u.s.-china-cooperation--not-a-race
- Goldstein, S., & Salib, P. AI risk and the law of AGI. (September 16, 2024). *Lawfare*. https://www.lawfaremedia.org/article/ai-risk-and-the-law-of-agi
- Goldstein, S., & Kirk-Giannini, C. D. AI is closer than ever to passing the Turing test for 'intelligence'. What happens when it does? (October 16, 2023). *The Conversation*. https://theconversation.com/ai-is-closer-than-ever-to-passing-the-turing-test-for-intelligence-what-happens-when-it-does-214721
- Goldstein, S., & Park, P. AI systems have learned how to deceive humans. What does that mean for our future? (September 4, 2023). *The Conversation*. https://theconversation.com/ai-systems-have-learned-how-to-deceive-humans-what-does-that-mean-for-our-future-212197
- Goldstein, S., & Kirk-Giannini, C. D. Is it ethical to create generative agents? Is it safe? (April 27, 2023). *ABC*. https://www.abc.net.au/religion/ai-generative-agents-are-unethical-and-unsafe/102277448
- Arbel, Y., Goldstein, S., & Salib, P. What happens once you give AI agents legal identity (letter to the editor). (June 2026). *Financial Times*. https://www.ft.com/content/9451b5e9-035e-46e4-bafd-ede04dd13d8b

## Selected Presentations

- "AI Revealed Preferences," AIES, Malmö, October 2026.
- "Governing AI Agents," AI Agents and the New Geopolitics of Power, European Central Bank and German Council on Foreign Relations, October 2026.

- Invited Speaker and Panelist, LMPST Taiwan, September 2026.
- Invited Talk, IPMC Research Seminar, National Yang Ming Chiao Tung University, September 2026.
- "AI Constitutions," Anthropic, July 2026.
- "AI Death," Keynote Address, Zhejiang University, April 2026.
- "AI Death," NUS, April 2026.
- "AI is not a Natural Monopoly," SMU, April 2026.
- "The Law of AI Agents," NTU, March 2026.
- "The Law of AI Agents," USC, March 2026.
- "AI Deals: Next Steps," Forethought, December 2025.
- "Legal Responses to AI's Existential Threat(s)," Panel, Croucher Workshop, CUHK, December 2025.
- "AI Copyright Law," Closing Roundtable, High-Level Forum on Generative AI Governance, HKUST, October 2025.
- "AI Rights," OpenAI, Mission Alignment Team, November 2025.
- "AI Personal Identity," Berggruen Institute, November 2025.
- "What Does ChatGPT Want?," Peking University, November 2025.
- "AI Agency," Yonsei University, October 2025.
- "Collaboration at the Brink," Yonsei University, October 2025.
- "Who's Afraid of AGI/ASI?," Panel, AI x Humanity Lab, HKU, October 2025.
- "A Guide to Living in the Simulation," Problems for Bayesianism Workshop, NUS, August 2025.
- "Collaboration at the Brink," Social Cyber Group, July 2025.
- "AI Sentience," Peking University, April 2025.
- "AI Geopolitics," Keynote Address, Evaluating Generative AI across Disciplines, CUHK, April 2025.
- "Moral Illusion and the Psychosemantics of Welfare," Well-Being and Desire: Themes from the Philosophy of Chris Heathwood, HKU, March 2025.
- "AI Geopolitics," European Central Bank, March 2025.
- "AI Geopolitics," German Council on Foreign Relations (DGAP), March 2025.
- "AI Sentience," Lingnan University, March 2025.
- "An Introduction to AI Catastrophic Risk," NSW Biobank, March 2025.
- "AI Deception," Philosophy of Deception conference, University of Macau, November 2024.
- "LLMs Cannot Be Ideally Rational," University of Chicago, October 2024.
- "LLMs Cannot Be Ideally Rational," LLMs and Philosophy Conference, JAIST, Kanazawa, September 2024.
- "AI Rights for Human Safety," FLI, MIT, August 2024.
- "AI Wellbeing," EAGxAustralia, Melbourne, September 2023.
- "AI Deception: A Survey of Examples, Risks, and Potential Solutions," Public Lecture, ACU Dianoia Institute, September 2023.
- "KK is Wrong Because We Say So," LENLS 19, November 2022.

- “KK is Wrong Because We Say So,” Berkeley, September 2022.
- “KK is Wrong Because We Say So,” LOGOS Epistemology Workshop, University of Barcelona, June 2022.
- “Vagueness and Opaque Sweetening,” ANU, April 2022.
- “Vagueness and Opaque Sweetening,” NUS, April 2022.
- “Omega Knowledge Matters,” University of Sydney, March 2022.
- “Omega Knowledge Matters,” Arizona State University, January 2022.
- “Omega Knowledge,” Normality and Epistemology Workshop (online), January 2022.
- “A Semantic Theory of Redundancy,” MIT, December 2021.
- “A Semantic Theory of Redundancy,” NYU, September 2021.
- “Divine Humility,” AAP Conference (online), July 2021.
- “Fragile Knowledge,” AAL, Melbourne, June 2021.
- “A Semantic Theory of Redundancy,” ILLC, Amsterdam, June 2021.
- “Divine Humility,” Philosophy of Religion Incubator, Princeton, May 2021.
- “A Semantic Theory of Redundancy,” ENS Paris, February 2021.
- “Contextology,” UT Austin, December 2020.
- “Probability for Epistemic Modals,” Logic Supergroup, April 2020.
- “Epistemic Modal Credence,” Dianoia Institute, Australian Catholic University, February 2020.
- “Counterfactual Contamination,” Lingnan University, January 2020.
- “Credence for Epistemic Expressivists,” TMC IV, Taipei, January 2020.
- “The Normality of Error,” Melbourne Logic Seminar, October 2019.
- “Knowledge from Multiple Appearances,” Dianoia Institute, Australian Catholic University, 2019.
- “Mighty Knowledge,” NUS Epistemology Conference, August 2019.
- “Fragile Knowledge,” Goethe Epistemology Meeting, Frankfurt, May 2019.
- “Conditional Communication,” UT Austin, May 2019.
- “Modal Credence,” UT Austin, May 2019.
- “Conditional Communication,” RUCCS Alumni Workshop, Rutgers, April 2019.
- “Conditional Communication,” Frontiers of Epistemology, Yonsei University, November 2018.
- “Conditional Communication,” TPLC, NTU, November 2018.
- “Free Choice and Homogeneity,” LENLS 15, Tokyo, November 2018.
- “Losing Confidence in Luminosity,” NUS, September 2018.
- “Conditionals in Generalized Update Semantics,” World Congress of Philosophy, Beijing, August 2018.
- “Conditional Communication,” Shanghai University, August 2018.
- “Entailment in Dynamic Semantics,” Topoi Conference, Turin, June 2018.
- “Free Choice and Homogeneity,” Meaning in Action, Oxford, March 2018.
- “A Theory of Conditional Assertion,” Central APA, Chicago, February 2018.
- “The Norms on Conditional Intention,” Lingnan University, February 2018.

- "Homogenous Alternative Semantics," Amsterdam Colloquium, December 2017.
- "Conditional Heresies," LENLS 14, Tokyo, November 2017.
- "Updating Conditional Intentions," TMC 2017, NTU, November 2017.
- "A Theory of Conditional Assertion," HKU, October 2017.
- "Believing Epistemic Contradictions," FEW, University of Washington, May 2017.
- "Free Choice Impossibility Results," Lingnan University, May 2017.
- "Free Choice Impossibility Results," Pacific APA, April 2017.
- "Generalized Update Semantics," Central APA, March 2017.
- "The T-Scheme without Bivalence," Eastern APA, January 2017.
- "Generalized Update Semantics," Northwestern University, December 2016.
- "Disjunction and Distributivity," Conditionals at the Crossroads, Konstanz, November 2016.
- "Conditional Excluded Middle and Duality," Conditionals at the Crossroads, Konstanz, November 2016.
- "Generalized Update Semantics," Northern Illinois Graduate Conference, October 2016.
- "A Dynamic Theory of Vagueness," Bochum-Rutgers Workshop, October 2016.
- "The Subjunctive Direct Argument," NASSLLI 2016, July 2016.
- "Epistemic Modality in situ," Barcelona Operators vs Quantifiers Workshop, June 2016.
- "Triviality Results for Probabilistic Modals," FEW, June 2016.
- "The Subjunctive Direct Argument," Pacific APA, April 2016.
- "Believing Epistemic Contradictions," Central APA, March 2016.
- "The Subjunctive Direct Argument," NY Philosophy of Language Workshop, February 2016.
- "Believing Epistemic Contradictions," Oxford Graduate Conference, November 2015.
- "Believing Epistemic Contradictions," Bridges 2, Rutgers, September 2015.
- "A Preface Paradox for Intention," Pacific APA, April 2015.
- "A Preface Paradox for Intention," Joint Session, July 2014.
- "Expressivist Conditionals," NY Philosophy of Language Workshop, April 2014.

## Supervision

- Sam Wang and Sofiia Lobanova, AI revealed preferences, SPAR fellowship, 2026.
- Sofiia Lobanova and Cole Niblett, AI revealed preferences, MARS fellowship, 2026.
- Yumin Liu, Suspension and sequential choice, HKU, PhD thesis (co-supervisor), 2025–present.
- Lee Elkin, HKU, postdoctoral fellow, 2025–present.
- Brandon Yip, ACU, postdoctoral fellow, 2023.
- Mitch Barrington, Ignoring the improbable, ACU, masters thesis, 2021.
- Wang Maomei, Credal inertia, Lingnan University, masters thesis, 2019.

## Research Grants

- AI Welfare, anonymous donor, 44,000 USD, 2026.
- AI Welfare, NYU Center for Mind, Ethics, and Policy, 88,000 HKD, 2026.

- AI Welfare, Lasting Impact Fund, HKU, 800,000 HKD, 2024.
- Humble Knowers, Templeton Foundation, 278,000 AUD, 2022.
- New Work in Dynamic Semantics, Early Career Scheme grant, Hong Kong UGC, 305,005 HKD, 2018.

## Service

- Director of Graduate Studies, HKU, Department of Philosophy, 2024–present.
- Organizing committee, Hong Kong Global AI Governance Conference, University of Hong Kong, April 2026.
- Advisory Board Member, MA in Artificial Intelligence and the Future, Hong Kong Catastrophic Risk Centre, Lingnan University, 2026–present.
- Co-organizer, "Who's Afraid of AGI/ASI?," AI x Humanity Lab, University of Hong Kong, October 2025.
- Co-organizer (with Herman Cappelen), AI Consciousness and Well-Being Workshop, University of Hong Kong, March 2025.
- Scientific Advisory Board Member, SAPAN / The Harder Problem, 2024–2026.
- Co-organizer (with John Hawthorne), Knowledge, Inquiry, and Intellectual Humility Workshop, Melbourne, August 2023.
- Director of Graduate Studies, ACU, Dianoia Institute, 2020–2023.
- Co-organizer (with Sam Carter), Normality and Epistemology Workshop (online), January 2022.
- Chair of Graduate Admissions, ACU, Dianoia Institute, 2021.
- Department representative, diversity committee, ACU.
- Co-organizer (with Dan Waxman), Department Seminar Series, Lingnan University, 2017–2019.
- Referee for Australasian Journal of Philosophy, Analysis, Cambridge University Press, Canadian Journal of Philosophy, Dialectica, Episteme, Ergo, Erkenntnis, Inquiry, Journal of Applied Non-Classical Logics, Journal of the American Philosophical Association, Journal of Philosophical Logic, Journal of Philosophy, Linguistics and Philosophy, Logic and Logical Philosophy, Mind, Neuroscience of Consciousness, Noûs, Oxford University Press, Pacific Philosophical Quarterly, Philosophy and Phenomenological Research, Philosophical Quarterly, Philosophical Studies, Philosophical Review, Proceedings of the National Academy of Sciences, Ratio, Res Philosophica, Thought.