POST-AGI SCIENCE AND SOCIETY WORKSHOP

ABSTRACT

Artificial General Intelligence (AGI) has long seemed distant, but rapid advances in large-scale learning, autonomous reasoning, and open-ended discovery make its emergence increasingly plausible. The Post-AGI Science and Society Workshop asks what comes next. If AGI becomes ubiquitous, reliable, and affordable, how will it reshape scientific inquiry, the economy of knowledge, and human society? Will humans remain central to discovery or become curators and interpreters of machine-generated insights? The workshop brings together researchers from machine learning, philosophy of science, and policy to explore human-AI scientific coexistence. Topics include automated hypothesis generation, causal reasoning in AGI, collaborative discovery, epistemic alignment between humans and machines, and socio-economic shifts driven by pervasive intelligence. Through keynotes, talks, and a panel, we will examine how science and our understanding of knowledge might evolve in a post-AGI world.

https://p-agi.netlify.app/

Workshop Summary Never before has there been such a growing sentiment among researchers that artificial general intelligence (AGI), an AI system capable of performing most tasks at or beyond human level (Morris et al., 2023), may be within reach in the foreseeable future. Although uncertainty remains about precise timelines, an important shift has taken place: debates are no longer centered on whether AGI is possible, but increasingly on how to define it and what its implications might be for science, society, and culture.

This view is no longer limited to speculative debate; it is increasingly reflected in the expectations of the scientific community. A significant fraction of AI researchers estimate a 50% chance that unaided machines will surpass humans in every task by 2047 (Grace et al., 2024). The central disagreements are shifting away from whether AGI is possible to when it will emerge and what its implications will be for science and society.

Given this accelerating progress, one of the most critical and unresolved questions is how AGI will reshape the scientific process itself. When machines can autonomously formulate hypotheses, design experiments, and generate discoveries, what research remains for humans to do?

This workshop builds on that question. Rather than debating feasibility, we invite participants to explore how to make today's scientific research directions resilient and meaningful in the event that AGI becomes a pervasive tool. Our purpose is not to endorse specific timelines but to ask: **Assuming AGI becomes widely available, what is the future of science?** As scientists and innovators, it is our responsibility to address this question proactively, ensuring that as discovery becomes automated, it remains trustworthy, safe, and beneficial for humanity.

To that end, we bring together leading voices from AI, robotics, and safety to explore the following questions:

- What are the true limits of reasoning-capable systems, and could their capabilities scale beyond our control? We will hear from Noam Brown, co-creator of ChatGPT o1, one of the most advanced models to date.
- As these systems grow more capable, how will the scientific process evolve? Will machines generate discoveries for humans to validate, or will humans generate questions for machines to resolve? Jeff Clune and Francesco Locatello will examine this shift from complementary angles: the former through the lens of open-ended learning and automated discovery, and the latter through the role of causality in shaping how scientists will formulate and interpret causal questions in the age of AGI.
- As robotics and AI become more pervasive, so do challenges in adaptability, efficiency, and real-world integration. Daniela Rus will discuss how autonomous, reconfigurable, and

scalable robotic systems can advance energy-efficient, intelligent technologies that enhance daily life.

- Finally, what role will humans play? Been Kim, a leader in interpretability and human-machine understanding, will explore how to build systems that remain cooperative and enable meaningful human-AI collaboration.

We invite submissions of original research that explore the future of science, technology, and society in an age where Artificial General Intelligence is pervasive. Our goal is to foster a discussion that looks beyond immediate technical hurdles to identify and address the fundamental questions that will remain relevant as general-purpose AI systems become ubiquitous.

Submissions are encouraged across two primary tracks: **Track1 - Technical Foundations for a Post-AGI World** and **Track2 - Socio-Economical and Future Visions**. All the submissions will be in the "Tiny Papers" format.

Track 1: Technical Foundations for a Post-AGI World. This track focuses on the core technical challenges required to build, understand, and control highly capable AI systems. We are interested in work that addresses the safety, robustness, and scalability of models that approach and surpass human-level intelligence.

Suggested topics include, but are not limited to:

- **Automated Scientific Discovery:**
  - AI systems and algorithms designed to automate research, generate novel hypotheses, and accelerate scientific discovery. This includes areas like symbolic-neural theorem proving (Selsam et al., 2019; Bansal et al., 2021) and AI-guided program search (Shi et al., 2023; Rule et al., 2024).
  - Explorations into the future of the scientific process when machines can independently design, execute, and interpret experiments (Hu et al., 2024; Mankowitz et al., 2023; Zhang et al., 2025).
  - Research on "Deep Research" paradigms where AI drives discovery, including program synthesis and library learning (Bowers et al., 2023; Ellis et al., 2021).
- **Scalable and Efficient Intelligence:**
  - Hardware-AI co-design to drive scalable, energy-efficient intelligence (Mirhoseini et al., 2021; 2020; Fawzi et al., 2022).
  - Research into resource-aware scaling and the fundamental limits of reasoning-capable systems (Costello et al., 2025; Shao et al., 2024).
  - Methods for building and deploying powerful AI systems with reduced computational, financial, and environmental costs (Schwartz et al., 2020).
- **Safety, Robustness, and Alignment:**
  - Novel techniques for aligning powerful AI systems with human values and intentions, including next-generation preference learning (Lee et al., 2024; Rafailov et al., 2023; Meng et al., 2024) and broader human-AI alignment strategies (Ellis, 2023; Li et al., 2025).
  - Methods for ensuring safety and robustness, including superalignment and scalable oversight (Burns et al., 2024; Bowman et al., 2022; Belrose et al., 2023).
  - Frameworks and benchmarks for rigorously evaluating AGI-level reasoning (Chollet et al., 2025a;b; Li et al., 2024).
  - Advanced methods for transparency and human-AI collaboration, enabling meaningful human oversight (Tang et al., 2024; Yang & Deng, 2019).

Track 2: Socio-Economical and Future Visions This track invites contributions that analyze the broader societal, ethical, and economic implications of AGI. We encourage speculative works, position papers, and in-depth studies that grapple with the profound transitions society will face.

Suggested Topics include, but are not limited to:

- **Economic and Societal Impact:**
  - Studies analyzing AGI's impact on key domains such as economics, law, education, and healthcare (Lee et al., 2025; Dunlop et al., 2024).
  - Analyses of labor-market disruptions, the future of work, and the role of humans in a world with ubiquitous AGI (Shao et al., 2025; Tomlinson et al., 2025).
  - Frameworks for evaluating the societal effects of AI, such as "GDPval" and other systemic evaluations (Patwardhan et al., 2025).
- **Governance, Regulation, and Risk:**
  - Proposals for the regulation and governance of powerful, general-purpose AI systems (Raji et al., 2022; Nolte et al., 2025).
  - Comprehensive risk assessments, including the study of existential risks and pathways from AGI to Artificial Superintelligence (ASI) (Anderljung et al., 2023; Green, 2020).
  - Position papers on how society can best govern and benefit from systems that reshape every field (Bengio et al., 2023).
- **Foundational Questions and Position Papers:**
  - Analyses and positions of long-term AI trajectories, focusing on potential further developments and their technical and societal implications (Morris et al., 2023; Gabriel, 2020; Wen et al., 2025).
  - Studies and position papers examining the theoretical and computational aspects of machine consciousness, models of computational creativity, and the evolving role of AI in creative practices and art production (Kamb & Ganguli, 2025; Blili-Hamelin et al., 2025).
  - Ethical frameworks for the deployment and use of AGI for social good (Prabhakaran et al., 2022; Gabriel et al., 2024).

P-AGI invites the community to step beyond immediate technical challenges and to define the next set of fundamental questions that will matter in a world where general intelligence is no longer rare. We believe that by hosting this discussion now, ICLR will reaffirm its leadership not only in advancing the field of machine learning, but in preparing it for its most profound transitions.

Workshop Format P-AGI is structured to foster deep discussion and cross-disciplinary collaboration. The program will integrate three main elements: *invited talks*, a *panel discussion*, and a *poster session*.

The *invited talks* will bring together leaders from frontier AI, robotics, AI safety, philosophy, and social sciences. Each speaker will present a vision or provocation addressing the scientific, societal, or creative challenges in a world where AGI is widespread.

The *panel discussion* will gather experts from across disciplines to debate key questions raised during the workshop: Which research problems will endure? How will society adapt? What forms of creativity and governance will be most important? The panel will be guided by live audience questions to keep the discussion responsive and grounded.

The *poster session* will showcase work submitted to the two workshop tracks. We will organize the poster session to maximize interaction, including assigning experienced attendees to engage with selected posters. This will ensure that junior researchers receive substantive feedback and encourage rich discussions around forward-looking ideas. Extended coffee breaks and lunch will be designed to encourage informal conversations and collaborations across fields. Our goal is to create not just a series of presentations, but an environment where bold ideas about the future of research can emerge and take root.

The organizers' prior experience running successful workshops at top venues will be instrumental in coordinating a smooth, high-impact event.

Points of difference Unlike workshops focused on advancing the current state of AI, P-AGI is designed to prepare the research community for the downstream effects of general intelligence. Our unique contribution is to create a dedicated space where researchers can shape a scientific agenda that is resilient, relevant, and responsible in the face of this transformative change. We will achieve this through three core objectives:

Tentative Schedule			
09:15	Opening Remarks	13:00	Lunch Break
09:30	Invited Talk 1	14:00	Invited Talk 4
10:00	Invited Talk 2	14:30	Invited Talk 5
10:30	Invited Talk 3	15:00	Invited Talk 6
11:00	Contributed Talks	15:30	Panel Discussion
11:30	Poster Session	16:30	Closing Remarks

- **Fostering Cross-Disciplinary Discussion:** The workshop is structured to bring together leading voices from diverse fields like AI, robotics, safety, and social sciences. Through invited talks, a cross-disciplinary panel, and interactive poster sessions, we will facilitate the deep discussions needed to tackle the multifaceted challenges of a post-AGI world.
- **Future-Proofing the Research Agenda:** Rather than pushing the immediate frontier, we seek to **future-proof the questions we ask**. By considering the implications of widely available AGI, our goal is to help the community identify and prioritize research directions that will hold fundamental value under a broad range of futures, ensuring today's work remains meaningful.
- **Defining the Next Set of Questions:** We aim to challenge participants to look beyond immediate technical hurdles and collectively **define the next set of fundamental questions** that will matter in a world where general intelligence is no longer rare. By hosting this forward-looking dialogue, P-AGI will help ensure the field is proactively and thoughtfully preparing for its most profound transitions.

To our knowledge, no other workshop takes this long-term perspective, placing future relevance at the core of its mission.

Diversity and Inclusivity We believe that diversity and inclusivity are essential to fostering a forward-thinking research community, especially when confronting the broad implications of AGI. P-AGI is committed to welcoming participants from all backgrounds and identities, and this commitment is reflected across every aspect of the event. We have prioritized diversity in seniority, gender, nationality, and institutional affiliation. Our organizing team and invited speakers span PhD students, early-career researchers, and senior academics, with an intentional 50/50 gender balance in both groups. The committee also represents a wide geographic spread, including Europe, North America, and Asia. To further broaden participation, we will actively engage affinity groups such as Black in AI, Women in Machine Learning (WiML), Queer in AI, and Latinx in AI, distributing calls for program committee roles and general participation among these corresponding networks. Finally, pending sponsorship, we intend to offer travel support for members of underrepresented groups to help reduce financial barriers to attendance. Through these efforts, we aim to cultivate an inclusive, interdisciplinary environment where all voices are heard, supported, and empowered to shape the future of AGI research.

Attendance We expect a vibrant and diverse audience, with over 150 in-person attendees and an additional 200 joining virtually. This strong interest is driven by the workshop's themes, which are widely felt across the AI community yet often confined to social media threads and informal conversations. P-AGI offers a space to bring these discussions into the open, with the rigor, depth, and interdisciplinary engagement they deserve. Participants will include students, academic researchers, and industry professionals from a wide range of backgrounds and cultures. To encourage interaction and community building, all attendees will have access to an open-source communication workspace. This shared space allows newcomers to introduce themselves, share their work, and stay informed about upcoming opportunities, events, and research highlights. A dedicated social channel will support informal meetups and casual conversations, helping participants connect beyond the formal sessions. Our goal is to create a dynamic, inclusive environment where everyone can contribute, learn, and remain engaged long after the workshop concludes.

To maximize the workshop's reach and impact for a global audience, all talks and presentations will be recorded (pending speaker permission) and made publicly available on the official workshop

website. This digital archive will also feature a complete collection of the accepted papers and posters, ensuring that the content remains accessible to the entire research community long after the event.

Invited Speakers and Panelists

Noam Brown *Research Scientist, OpenAI* [Confirmed]
Noam Brown leads the development of advanced reasoning models at OpenAI, including o1, designed for general problem-solving in math, science, and code. He was named one of MIT's 35 Innovators Under 35 and received the Marvin Minsky Medal for his contributions to AI. Having led work on one of the most capable reasoning AIs to date, Noam offers a rare perspective on the capabilities, limits, and implications of general-purpose systems.

Jeff Clune *Full Professor, University of British Columbia* [Confirmed]
Jeff Clune is a Full Professor at the University of British Columbia, a Senior Research Advisor at DeepMind, and a Canada CIFAR AI Chair at the Vector Institute. His research focuses on open-ended learning, AI-generating algorithms, and deep reinforcement learning – core ideas for scaling intelligence. As a pioneer in evolving increasingly general and autonomous AI systems, Jeff brings a visionary yet technically grounded perspective, making him an ideal contributor to the workshop themes.

Daniela Rus *Andrew & Erna Viterbi Professor, MIT | Director, MIT CSAIL* [Confirmed]
Daniela Rus leads MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), advancing robotics and AI for real-world autonomy, from soft robots to modular and swarm systems. A MacArthur Fellow and elected member of the National Academy of Engineering and National Academy of Sciences, she has received several awards, including the IEEE Edison Medal and the John Scott Medal. Her work focuses on scalable, energy-efficient AI–robotics integration that enhances daily life.

Francesco Locatello *Assistant Professor, ISTA* [Confirmed]
Francesco Locatello is is an assistant professor at ISTA and an AI Resident at the Chan Zuckerberg Initiative. His research focuses on causal representation learning, developing models that understand cause and effect, adapt to interventions, and generalize beyond fixed conditions. At P-AGI, he will discuss how AGI may transform the way scientists approach causal questions, exploring how advances in causal reasoning could redefine discovery and explanation in a post-AGI scientific landscape.

Been Kim *Senior Staff Research Scientist, Google DeepMind* [Confirmed]
Been Kim is a leading voice in interpretable AI and creator of TCAV, a concept-based explanation method that won the UNESCO Netexplo Award. Her work focuses on enabling human–machine collaboration, detecting model failures, and using AI knowledge for human benefit. As humans and AI systems grow increasingly entangled, her perspective is not only advisable, but essential to ensuring reliable, understandable, and aligned intelligence.

Organizers

Organizational experience The P-AGI organizing committee brings deep and proven experience in curating high-impact scientific events. Three of its members were among the founding organizers of UniReps, one of the most attended workshops at NeurIPS in both 2023 and 2024. Collectively, the committee has organized a total of more than 20 workshops across major conferences in machine learning, natural language processing, robotics, and computer vision. This track record reflects not only logistical capability but also a strong ability to attract diverse, interdisciplinary communities and sustain meaningful discussion around frontier research questions, precisely the kind of engagement P-AGI aims to foster.

Emanuele Rodolà *Sapienza University of Rome*

Emanuele is Full Professor of Computer Science at Sapienza University of Rome, where he leads the GLADIA group of learning and applied AI, funded by an ERC Grant, a FIS Grant, and a Google Research Award. Previously, he was Assistant and then Associate Professor at Sapienza (2017-2020), a postdoc at USI Lugano (2016-2017), an Alexander von Humboldt Fellow at TU Munich (2013-2016), and a JSPS Research Fellow at The University of Tokyo (2013). He is a fellow of ELLIS and the Young Academy of Europe, has received a number of research prizes, has been serving in the program and organizing committees of the top rated conferences in computer vision and machine learning, founded and chaired several successful workshops including UniReps at NeurIPS 2023 and 2024. His research interests lie at the intersection of representation learning, model merging, graph / geometric deep learning, language and learning for audio, and has published more than 170 papers in these areas. Previously, he has organized and lectured at 15 tutorials, and has co-organized and chaired more than 10 workshops co-located with ECCV, ICCV, and NeurIPS.

Pratyusha Sharma *Microsoft & NYU*

Pratyusha Sharma is a Senior Research Scientist at Microsoft Research and an incoming Assistant Professor at the Courant Institute and Center for Data Science at New York University. She did her PhD from MIT in the Computer Science and Artificial Intelligence Lab. She studies the interplay between language, sequential decision making and intelligence in natural and AI systems. Her research has also been featured in articles in the New York Times, National Geographic Magazine, BBC, etc. She was recently a speaker at TED AI and was selected as a Rising Star in EECS, Data Science, and GenAI. She has previously organized the Language and Reinforcement Learning Workshop (LaReL) NeurIPS 2022, the workshop on Language and Robotics at Conference on Robot Learning in 2022 and Social Intelligence in Humans and Robots Workshop, Robotics Science and Systems 2023.

Andrea Santilli *Independent Researcher*

Andrea Santilli is a Research Scientist at Nous Research, where he focuses on post-training large language models (LLMs) to improve their robustness, reliability, and alignment. He holds a PhD in Computer Science (2025) from Sapienza University of Rome, with a thesis on building "Effective, Efficient, and Reliable Large Language Models." Previously, he was a Research Scientist at Apple MLR (2024) and collaborated with leading institutions like Hugging Face (2021). Despite being early in his career, Andrea's work has already had a significant impact: his research has been published at top-tier conferences (ICLR, ICML, ACL, EMNLP, CVPR, SIGIR, AAAI) and has accumulated over 7,000 citations. In addition to his research contributions, he actively serves on program committees for major conferences and has received multiple awards, including the prestigious Imminent Research Grant.

Valentina Pyatkin *Allen Institute for AI & University of Washington*

Valentina Pyatkin is a Postdoctoral Researcher at the Allen Institute for AI and the University of Washington, working with Prof. Yejin Choi and Prof. Hannaneh Hajishirzi. She completed her PhD in Computer Science at the NLP lab of Bar Ilan University. Her work has been awarded an ACL Outstanding Paper Award and the ACL Best Theme Paper Award. Her research focuses on language model alignment, with a focus on steerability, values, pragmatics and context. Valentina was a co-organizer of the 2nd and 3rd UnImplicit workshop at NAACL 2022 and EACL 2024, and a co-organizer of the 2nd and 3rd SoLaR workshops collocated with NeurIPS 2024 and CoLM 2025.

Donato Crisostomi *Sapienza University of Rome*

Donato Crisostomi is an ELLIS PhD student at Sapienza University of Rome and University of Cambridge, focusing on model merging and representational alignment. He currently leads the "Model Reuse" work package for the 1.5M€ project "*NEXUS: Interoperable Machine Learning with Universal Representations*". He previously held roles as a visiting researcher at the University of Cambridge, a Research Scientist at Amazon Alexa, and an Applied Scientist at Amazon Search. His research has been featured in top-tier AI conferences and journals, including CVPR, NeurIPS, ACM, ACL, and LoG. In addition to his scientific contributions, he has played an active role in the research community as the organizer of several workshops, including the UniReps workshop at NeurIPS (one of the most attended in NeurIPS 2023 and 2024) and as a program committee member for leading conferences such as CVPR, ICML, NeurIPS, ICLR, etc.

Zorah Lähner *University of Bonn and Lamarr Institute*

Zorah Lähner is assistant professor and head of the Geometry in Machine Learning group at the University of Bonn and the Lamarr Institute for Machine Learning and Artificial Intelligence in Germany. She previously worked at the University of Siegen, Technical University of Munich, Meta Reality Labs and Toshiba Research Europe. Her research focuses on geometric deep learning and 3D vision, and has been published in major machine learning and computer vision conferences, including NeurIPS, ICLR, CVPR, and ICCV. Beyond that she has been involved in the organization of the UniReps workshop at NeurIPS, a section of Women in Vision, been program chair of the German Conference on Pattern Recognition, and received several outstanding reviewer awards.

Program Committee

The following is a preliminary list of 50 *confirmed* Program Committee members. Based on an anticipated 100 submissions and a target load of no more than four short papers per reviewer, we will need roughly 75–100 reviewers. We are therefore actively recruiting additional members, especially from diverse and historically under-represented communities in AI, and expect to reach the required committee size within the next few weeks.

1. Tommaso Mencattini (EPFL) [Confirmed]
2. Andrea Caciolai (Meta) [Confirmed]
3. Simone Antonelli (CISPA) [Confirmed]
4. Stefano Esposito (U Tübingen) [Confirmed]
5. Irene Tallini (Area Science Park, Trieste) [Confirmed]
6. Pietro Barbiero (USI) [Confirmed]
7. Lorenzo Giusti (CERN) [Confirmed]
8. Eleonora Gualdoni (Apple) [Confirmed]
9. Irene Cannistraci (ETH) [Confirmed]
10. Marco Fumero (ISTA) [Confirmed]
11. Daniele Baieri (U Milano-Bicocca) [Confirmed]
12. Luca Moschella (Apple) [Confirmed]
13. Luca Cosmo (Ca' Foscari U Venice) [Confirmed]
14. Robert Adrian Minut (Sapienza) [Confirmed]
15. Adam Goliński (Apple) [Confirmed]
16. Daniele Solombrino (Sapienza) [Confirmed]
17. Miao Xiong (NUS) [Confirmed]
18. Maks Ovsjanikov (Ecole polytechnique) [Pending]
19. Federico Danieli (Apple) [Confirmed]
20. Gabriele Sarti (U Groningen) [Confirmed]
21. Arianna Rampini (Autodesk) [Confirmed]
22. Davide Marincione (Sapienza) [Confirmed]
23. Michele Miranda (Translated) [Confirmed]
24. Emanuele Rossi (VantAI) [Confirmed]
25. Michele Mancusi (Moises) [Confirmed]
26. Emily Cheng (UPF) [Confirmed]
27. Roberto Dessì (Samaya AI) [Confirmed]
28. Tal Remez (Meta) [Pending]
29. Giorgio Mariani (U Milano-Bicocca) [Confirmed]
30. Antonio Norelli (MIT) [Confirmed]
31. Giorgio Strano (Sapienza) [Confirmed]
32. Florian Bernard (U Bonn) [Pending]
33. Alessio Devoto (NVIDIA) [Confirmed]
34. Lucas Weber (Fraunhofer IIS) [Confirmed]
35. Or Litany (Technion) [Confirmed]
36. Riccardo Marin (TUM) [Confirmed]
37. Giovanni Trappolini (Sapienza) [Confirmed]
38. Emtiyaz Khan (Riken) [Pending]
39. Emilian Postolache (IRIS Audio) [Confirmed]
40. Simone Melzi (U Milano-Bicocca) [Confirmed]
41. Asako Kanezaki (Tokyo Tech) [Pending]
42. Zorah Lähner (U Bonn) [Confirmed]
43. Federico Bomba (Sineglossa) [Confirmed]
44. Silvia Zuffi (CNR) [Confirmed]
45. Alex Bronstein (ISTA) [Confirmed]
46. Fabrizio Frasca (Technion) [Confirmed]
47. Thomas Möllenhoff (Riken) [Confirmed]
48. Michael Möller (U Siegen) [Confirmed]
49. Laura Leal-Taixé (NVIDIA) [Pending]
50. Antonio Gargiulo (Sapienza) [Confirmed]
51. Haggai Maron (Technion) [Confirmed]
52. Silvio Severino (Amazon) [Confirmed]
53. Jacob Morrison (Ai2) [Confirmed]
54. Ethan Shen (U Washington) [Confirmed]
55. Osamu Hirose (U Kanazawa) [Confirmed]
56. Clementine Domine (UCL) [Confirmed]
57. Pritish Chakraborty (IIT Bombay) [Pending]
58. Viktor Stenby Johansson (DTU) [Pending]
59. Akshit Achara (King's College London) [Pending]
60. Hugo Daniel Monzón Maldonado (Riken) [Pending]

REFERENCES

Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. Frontier ai regulation: Managing emerging risks to public safety, 2023. URL https://arxiv.org/abs/2307.03718. 3

Kshitij Bansal, Christian Szegedy, Markus Norman Rabe, Sarah M. Loos, and Viktor Toman. Learning to reason in large theories without imitation, 2021. URL https://openreview.net/forum?id=qbRv1k2AcH. 2

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023. 2

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian K. Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atilim Gunes Baydin, Sheila A. McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Markus Brauner, and Sören Mindermann. Managing extreme ai risks amid rapid progress. *Science*, 384:842 – 845, 2023. URL https://api.semanticscholar.org/CorpusID:269929051. 3

Borhane Blili-Hamelin, Christopher Graziul, Leif Hancox-Li, Hananel Hazan, El-Mahdi El-Mhamdi, Avijit Ghosh, Katherine A Heller, Jacob Metcalf, Fabricio Murai, Eryk Salvaggio, Andrew J Smart, Todd Snider, Mariame Tighanimine, Talia Ringer, Margaret Mitchell, and Shiri Dori-Hacohen. Position: Stop treating 'AGI' as the north-star goal of AI research. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=1RlrtH6ydW. 3

Matthew Bowers, Theo X. Olausson, Lionel Wong, Gabriel Grand, Joshua B. Tenenbaum, Kevin Ellis, and Armando Solar-Lezama. Top-down synthesis for library learning. *Proc. ACM Program. Lang.*, 7(POPL), January 2023. doi: 10.1145/3571234. URL https://doi.org/10.1145/3571234. 2

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models, 2022. URL https://arxiv.org/abs/2211.03540. 2

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 4971–5012, 2024. 2

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc prize 2024: Technical report, 2025a. URL https://arxiv.org/abs/2412.04604. 2

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems, 2025b. URL https://arxiv.org/abs/2505.11831. 2

Caia Costello, Simon Guo, Anna Goldie, and Azalia Mirhoseini. Think, prune, train, improve: Scaling reasoning without scaling models, 2025. URL https://arxiv.org/abs/2504.18116. 2

Connor Dunlop, Weiwei Pan, Julia Smakman, Lisa Soder, Siddharth Swaroop, and Noam Kolt. Position: AI agents & liability – mapping insights from ML and HCI research to policy. In *Workshop on Socially Responsible Language Modelling Research*, 2024. URL https://openreview.net/forum?id=pa80BLEavx. 3

Kevin Ellis. Human-like few-shot learning via bayesian reasoning over natural language. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 2

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B. Tenenbaum. Dreamcoder: bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, PLDI 2021, pp. 835–850, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383912. doi: 10.1145/3453483.3454080. URL https://doi.org/10.1145/3453483.3454080. 2

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J. R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, David Silver, Demis Hassabis, and Pushmeet Kohli. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610:47 – 53, 2022. URL https://api.semanticscholar.org/CorpusID:252717185. 2

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020. 3

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. The ethics of advanced ai assistants, 2024. URL https://arxiv.org/abs/2404.16244. 3

Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. Thousands of ai authors on the future of ai. *arXiv preprint arXiv:2401.02843*, 2024. 1

Ben Green. The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 594–606, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372869. URL https://doi.org/10.1145/3351095.3372869. 3

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. *ArXiv*, abs/2408.08435, 2024. URL https://api.semanticscholar.org/CorpusID:271892234. 2

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ilpL2qACla. 3

Gyeonggeon Lee, Lehong Shi, Ehsan Latif, Yizhu Gao, Arne Bewersdorff, Matthew Nyaaba, Shuchen Guo, Zhengliang Liu, Gengchen Mai, Tianming Liu, et al. Multimodality of ai for education: Towards artificial general intelligence. *IEEE Transactions on Learning Technologies*, 2025. 3

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*, pp. 26874–26901. PMLR, 2024. 2

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LvDwwAgMEW. 2

Wen-Ding Li, Keya Hu, Carter Larsen, Yuqing Wu, Simon Alford, Caleb Woo, Spencer M. Dunn, Hao Tang, Michelangelo Naim, Dat Nguyen, Wei-Long Zheng, Zenna Tavares, Yewen Pu, and Kevin Ellis. Combining induction and transduction for abstract reasoning, 2024. URL https://arxiv.org/abs/2411.02272. 2

Daniel J. Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, Thomas Köppe, Kevin Millikin, Stephen Gaffney, Sophie Elster, Jackson Broshear, Chris Gamble, Kieran Milan, Robert Tung, Minjae Hwang, Taylan Cemgil, Mohammadamin Barekatain, Yujia Li, Amol Mandhane, Thomas Hubert, Julian Schrittwieser, Demis Hassabis, Pushmeet Kohli, Martin Riedmiller, Oriol Vinyals, and David Silver. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263, Jun 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06004-9. URL https://doi.org/10.1038/s41586-023-06004-9. 2

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3Tzcot1LKb. 2

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Sungmin Bae, et al. Chip placement with deep reinforcement learning. *arXiv preprint arXiv:2004.10746*, 2020. 2

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nova, et al. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021. 2

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi for operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*, 2023. 1, 3

Henrik Nolte, Miriam Rateike, and Michèle Finck. Robustness and cybersecurity in the EU artificial intelligence act. In *NeurIPS 2024 Workshop on Regulatable ML*, 2025. URL https://openreview.net/forum?id=1m1XKCGerV. 3

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374. 3

Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A human rights-based approach to responsible ai, 2022. URL https://arxiv.org/abs/2210.02667. 3

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9. 2

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 557–571, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534181. URL https://doi.org/10.1145/3514094.3534181. 3

Joshua S. Rule, Steven T. Piantadosi, Andrew Cropper, Kevin Ellis, Maxwell Nye, and Joshua B. Tenenbaum. Symbolic metaprogram search improves learning efficiency and explains rule learning in humans. *Nature Communications*, 15(1):6847, 2024. doi: 10.1038/s41467-024-50966-x. URL https://doi.org/10.1038/s41467-024-50966-x. 2

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12): 54–63, November 2020. ISSN 0001-0782. doi: 10.1145/3381831. URL https://doi.org/10.1145/3381831. 2

Daniel Selsam, Matthew Lamm, Benedikt Bünz, Percy Liang, Leonardo de Moura, and David L. Dill. Learning a SAT solver from single-bit supervision. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJMC_iA5tm. 2

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. Scaling retrieval-based language models with a trillion-token datastore. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=iAkhPz7Qt3. 2

Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with ai agents: Auditing automation and augmentation potential across the u.s. workforce, 2025. URL https://arxiv.org/abs/2506.06576. 3

Kensen Shi, Hanjun Dai, Wen-Ding Li, Kevin Ellis, and Charles Sutton. Lambdabeam: neural program search with higher-order functions and lambdas. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. 2

Hao Tang, Darren Yan Key, and Kevin Ellis. Worldcoder, a model-based LLM agent: Building world models by writing code and interacting with the environment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=QGJSXMhVaL. 2

Kiran Tomlinson, Sonia Jaffe, Will Wang, Scott Counts, and Siddharth Suri. Working with ai: Measuring the applicability of generative ai to occupations, 2025. URL https://arxiv.org/abs/2507.07935. 3

Ying Wen, Ziyu Wan, and Shao Zhang. Language games as the pathway to artificial superhuman intelligence, 2025. URL https://arxiv.org/abs/2501.18924. 3

Kaiyu Yang and Jia Deng. Learning to prove theorems via interacting with proof assistants. In *International Conference on Machine Learning (ICML)*, 2019. 2

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents, 2025. URL https://arxiv.org/abs/2505.22954. 2