Association for the
Advancement of
Artificial Intelligence

AAAI 2025 PRESIDENTIAL PANEL ON THE

Future of AI Research

Published March 2025

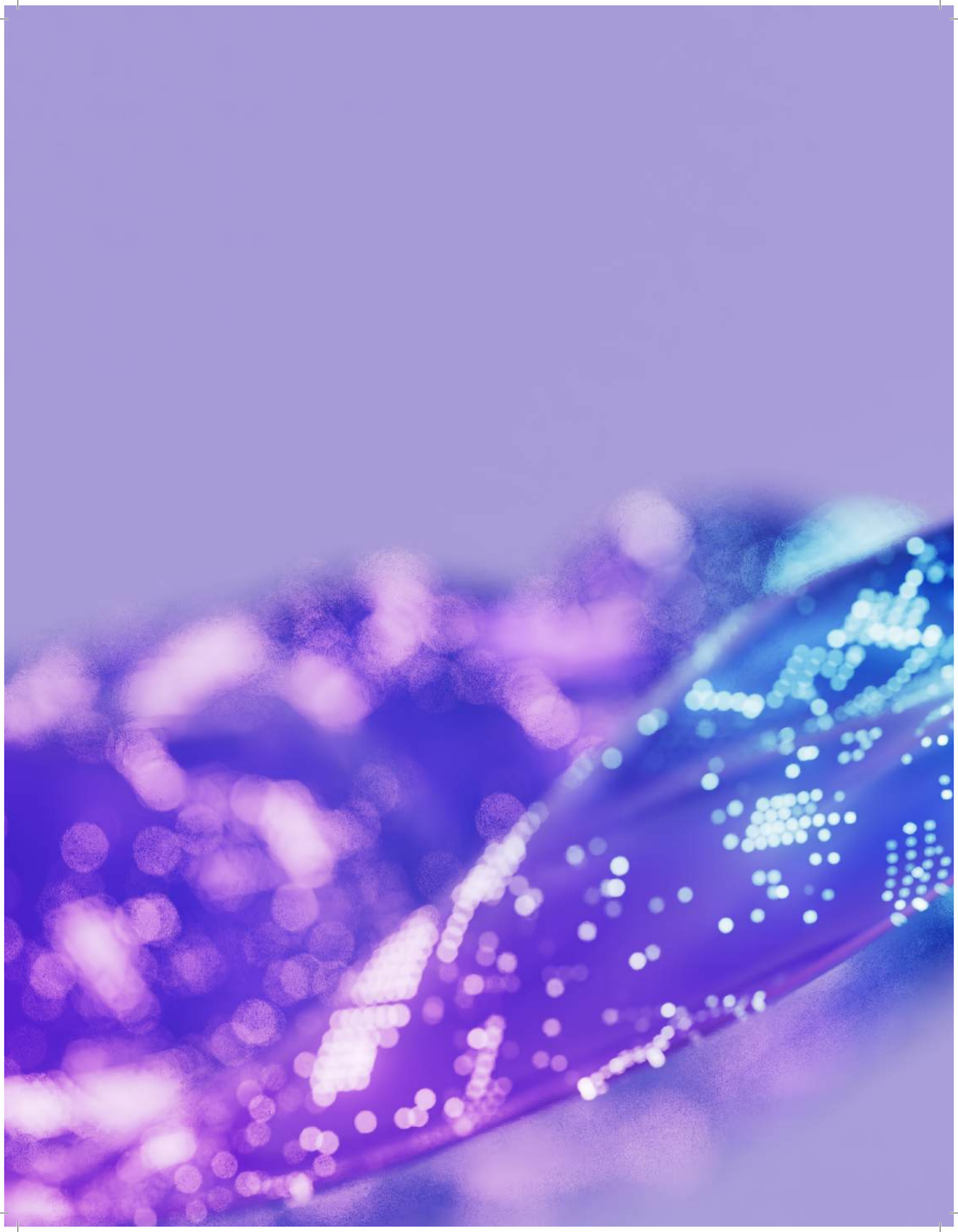

Table of Contents

Introduction

As AI capabilities evolve rapidly, AI research is also undergoing a fast and significant transformation along many dimensions, including its topics, its methods, the research community, and the working environment. Topics such as AI reasoning and agentic AI have been studied for decades but now have an expanded scope in light of current AI capabilities and limitations. AI ethics and safety, AI for social good, and sustainable AI have become central themes in all major AI conferences. Moreover, research on AI algorithms and software systems is becoming increasingly tied to substantial amounts of dedicated AI hardware, notably GPUs, which leads to AI architecture co-creation, in a way that is more prominent now than over the last 3 decades. Related to this shift, more and more AI researchers work in corporate environments, where the necessary hardware and other resources are more easily available, compared to academia, questioning the roles of academic AI research, student retention, and faculty recruiting.

The pervasive use of AI in our daily lives and its impact on people, society, and the environment makes AI a socio-technical field of study, thus highlighting the need for AI researchers to work with experts from other disciplines, such as psychologists, sociologists, philosophers, and economists. The growing focus on emergent AI behaviors rather than on designed and validated properties of AI systems renders principled empirical evaluation more important than ever. Hence the need arises for well-designed benchmarks, test methodologies, and sound processes to infer conclusions from the results of computational experiments. The exponentially increasing quantity of AI research publications and the speed of AI innovation are testing the resilience of the peer-review system, with the immediate release of papers without peer-review evaluation having become widely accepted across many areas of AI research. Legacy and social media increasingly cover AI research advancements, often with contradictory statements that confuse the readers and blur the line between reality and perception of AI capabilities. All this is happening in a geo-political environment, in which companies and countries compete fiercely and globally to lead the AI race. This rivalry may impact access to research results and infrastructure as well as global governance efforts, underscoring the need for international cooperation in AI research and innovation.

In this overwhelming multi-dimensional and very dynamic scenario, it is important to be able to clearly identify the trajectory of AI research in a structured way. Such an effort can define the current trends and the research challenges still ahead of us to make AI more capable and reliable, so we can safely use it in mundane but also, most importantly, in high-stake scenarios.

This study aims to do this by including 17 topics related to AI research, covering most of the transformations mentioned above. Each chapter of the study is devoted to one of these topics, sketching its history, current trends and open challenges.

To conduct this study, I selected a very diverse group of 24 experienced AI researchers, who generously accepted my invitation and devoted a significant amount of time to this effort. We all worked together between summer 2024 and spring 2025 to structure the study, define the main topics, discuss the content, comment and contribute to the various chapters.

Additionally, some chapters engaged also with additional contributors who brought their expertise on a specific topic. The work was done mostly online, with monthly calls with all panel members plus additional calls for the team working on each chapter, with also in a full-day in-person meeting, held in January 2025.

However, we also wanted to include the opinion of the entire AAAI community, so we launched an extensive survey on the topics of the study, which engaged 475 respondents, of which about 20% were students. Among the respondents, academia was given as the main affiliation (67%), followed by corporate research environment (19%). Geographically, the most represented areas are North America (53%), Asia (20%), and Europe (19%) . While the vast majority of the respondents listed AI as one of their primary fields of study, there were also mentions of other fields, such as neuroscience, medicine, biology, sociology, philosophy, political science, and economics. This multi-field involvement was also reflected in an interest in multi-disciplinary research from 95% of the respondents.

Each chapter of this report includes a brief summary of the responses to questions related to the respective topic.

The work around the entire study has been generously supported and made possible by the amazing work of Meredith Ellison, AAAI Executive Director, and the AAAI office staff, who also prepared and delivered the survey.

I hope that this report will be useful to the whole AI research community. However, the report has been intentionally written in a non-technical way, to reach out to other audiences, including experts of other disciplines, policy makers, funding agencies, the media, and the general public.
We all need to work together to advance AI in a responsible way, to make sure that technological progress supports the progress of humanity and is aligned to human values.

Francesca Rossi
AAAI President, 2022–2025

The panel's findings are opinions of the panel members and do not represent the opinion of their institutions or companies.

Panel Members & Additional Contributors

Panel Members

Francesca Rossi,
IBM Research

Christian Bessiere,
University of Montpellier

Joydeep Biswas,
University of Texas at Austin

Rodney Brooks
Massachusetts Institute of Technology

Vincent Conitzer,
Carnegie Mellon University

Thomas G. Dietterich,
Oregon State University

Virginia Dignum,
Umeå University

Oren Etzioni,
University of Washington

Kenneth D. Forbus,
Northwestern University

Eugene Freuder,
University College Cork

Yolanda Gil,
University of Southern California

Holger Hoos,
RWTH Aachen University, Germany and Leiden University, The Netherlands

Eric Horvitz,
Microsoft

Subbarao Kambhampati,
Arizona State University

Henry Kautz,
University of Virginia

Jihie Kim,
Dongguk University

Hiroaki Kitano,
Sony Research

Alan Mackworth,
University of British Columbia

Karen Myers,
SRI International

Luc De Raedt,
KU Leuven and Örebro University

Stuart Russell,
University of California Berkeley

Bart Selman,
Cornell University

Peter Stone,
The University of Texas at Austin and Sony AI

Millind Tambe,
Harvard University

Michael Wooldridge,
University of Oxford

Additional Contributors

Aditya Akella,
University of Texas at Austin
Chapter: Hardware & AI

Yoshua Bengio,
MILA
Chapter: Artificial General Intelligence (AGI)

Abeba Birhane,
Trinity College Dublin
Chapter: Research Beyond the AI Research Community

Bill Dally,
NVIDIA
Chapter: Hardware and AI

Fei Fang,
Carnegie Mellon University
Chapter: AI for Social Good

Jonathan Gratch,
University of Southern California
Chapter: AI & Cognitive Science

Norm Jouppi,
Google
Chapter: Hardware and AI

John E. Laird,
University of Michigan
Chapter: AI & Cognitive Science

Amy Luers,
Microsoft
Chapter: AI & Sustainability

Peter Norvig,
Google
Chapter: Artificial General Intelligence (AGI)

Besmira Nushi,
Microsoft Research
Chapter: Artificial General Intelligence (AGI)

Balaraman Ravindran,
Indian Institute of Technology Madras
Chapter: AI for Social Good

Yoav Shoham,
Stanford University
Chapter: AI Agents

Carles Sierra,
Spanish National Research Council
Chapter: AI Agents

Pradeep Varakantham,
Singapore Management University
Chapter: AI for Social Good

AI Reasoning

The ability to reason has been a salient characteristic of human intelligence, and there is a critical need for verifiable reasoning in AI systems.

CHAIRS

Christian Bessiere,
University of Montpellier

Holger Hoos,
RWTH Aachen University, Germany and Leiden University, The Netherlands

Subbarao Kambhampati,
Arizona State University

Main Takeaways

- Reasoning has always been seen as a core characteristic of human intelligence. Reasoning is used to derive new information from given base knowledge; this new information is guaranteed correct when sound formal reasoning is used, otherwise it is merely plausible.
- AI research has led to a range of automated reasoning techniques. These reasoning techniques have given rise to AI algorithms and systems, including SAT, SMT, and constraints solvers as well as probabilistic graphical models, all of which play a key role in critical real-world applications.
- While large pre-trained systems (such as LLMs) have made impressive advancements in their reasoning capabilities, more research is needed to guarantee correctness and depth of the reasoning performed by them; such guarantees are particularly important for autonomously operating AI agents.

Context & History

Reasoning is a core component of human intelligence. From the dawn of humanity, abductive reasoning has been used to predict danger and inductive reasoning made it possible to learn regularities governing the world. Beginning in Ancient Greece, deductive reasoning techniques were developed to draw valid conclusions that follow logically from premises known to be true. The development of reasoning methods with such a priori guarantees was a key factor in the advancement of modern science, mathematics, and engineering; notably, according to philosophers such as Charles Sanders Peirce, the interplay between abduction, deduction, and induction forms the basis of the scientific method and hence all modern science. Attempts to mechanize logical reasoning can be traced back to 13th-century philosopher Ramon Lull and lie at the heart of the concept of computation. Probabilistic reasoning and inference have also profoundly impacted reasoning, often relying on the celebrated theorem by Thomas Bayes on inverse probability that also forms the basis for many machine learning and statistics approaches. Finally, the evaluation of correct (sound) reasoning lies at the heart of most quantitative assessments of human cognition.

Not surprisingly, reasoning has been central to the AI enterprise. Indeed, the earliest research in AI – from Logic Theorist onwards [1] – had a strong focus on reasoning [2]. Since the 1960s, AI has also embraced probabilistic reasoning and models, initially for medical diagnosis [3]. Since then, the reasoning tasks addressed in AI research have covered the gamut from planning and temporal reasoning to diagnosis and explanation. While early AI has paid attention to both plausible reasoning (case-based, analogical, qualitative) and sound formal reasoning with guarantees (logical, probabilistic, constraint-based), over the years, the focus has shifted more towards reasoning with formal guarantees. There are good reasons for this when designing AI systems and techniques that compensate for human limitations and weaknesses since reasoning with guarantees is challenging for humans. This has led to practically impactful applications of AI systems such as SAT, SMT, and constraints solvers, including the verification of correctness properties of computer hardware and software, the safety of communications protocols, the design of new proteins, and, more recently, the robustness of neural networks against adversarial attacks. It has also resulted in probabilistic graphical models [4, 5], which are powerful modeling and inference tools that have found their way into numerous applications of reasoning in medicine, robotics, and beyond.

Current State & Trends

The emergence of the Internet and the associated technology that made it possible to capture the human digital footprint at scale, as well as the leaps in computing power, have made possible novel approaches to learning bottom-up from data. Of particular interest are large pre-trained models, such as LLMs, that have shown surprising abilities in plausible reasoning. Unlike the earlier research on reasoning in AI, LLMs have focused on plausible reasoning patterns as they emerge automatically after large-scale training on petabyte corpora. While the results have been quite remarkable so far, the reasoning in this context has been of the "plausible" variety with no guarantees.

Meanwhile, sound formal reasoning techniques remain key to important and impactful applications of cutting-edge AI technology for the verification of computer hardware and software, as well as for real-world planning and resource allocation problems. They are also increasingly recognized as a crucial basis for the formal verification of machine learning techniques such as neural networks, e.g., in the context of local robustness against adversarial attacks [6]. Significant research activity takes place in these areas, focusing on improving various types of reasoning algorithms (notably with respect to their computational complexity), leveraging learning within sound formal reasoning, and combining reasoning and learning techniques [7, 8].

Research Challenges

Bringing some of the rigorous a priori or post hoc guarantees back into plausible reasoning patterns turbocharged by the pre-trained models has become an active and promising area of research – especially where AI systems need to work autonomously in safety-critical domains. Research on so-called "large reasoning models" as well as on neuro-symbolic approaches is addressing these challenges.

Furthermore, even though formal reasoning with correctness guarantees is currently considerably less in vogue than the use of generative AI techniques for plausible reasoning, formidable and

essential challenges also remain in that area. In this context, the combination of machine learning techniques with formal reasoning techniques holds considerable promise for economically and socially valuable breakthroughs, notably in the area of AI safety and transparency.

The questions and challenges we face range from the philosophical:

- What exactly is “reasoning”?

to the practical:

- Can LLM ‘reasoning’ be trusted?

and include:

- What does the future hold for the advancement and role of symbolic reasoning?
- To what extent can LLMs or other generative models reproduce or replace symbolic reasoning?
- To what degree will symbolic reasoning be necessary or sufficient to overcome the current limitations of LLMs?
- How well can AI reasoning, especially LLM ‘reasoning,’ be explained and understood?
- How can computers better understand and simulate human reasoning?
- What is the role of collaborative reasoning between humans and computers?
- How best can LLMs and symbolic reasoning be integrated into “neuro-symbolic reasoning”?
- Are further breakthroughs, beyond both LLMs and traditional symbolic reasoning, required to achieve AGI-level reasoning?
- What forms of reasoning can best support humans when dealing with various challenges, e.g., in medical, scientific, engineering, and legal domains?

1. Newell, A. & Simon, H. (1956). The logic theory machine: A complex information processing system. IRE Transactions on Information Theory 2: 61–79.
2. Brachman, R. and Levesque, H. (2004) Knowledge Representation and Reasoning (1st Ed). Morgan Kaufman.
3. Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach (4th Ed). Pearson.
4. Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufman.
5. Koller, D. and Friedmann, N. (2009) Probabilistic Graphical Models. The MIT Press.
6. König, M. et al. (2024) Critically Assessing the State of the Art in Neural Network Verification. Journal of Machine Learning Research 25(12): 1–53
7. Guo, D. et. al. (2025) DeepSeek-R: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://arxiv.org/abs/2501.12948
8. Kambhampati, S. (2024). Can Large Language Models Reason and Plan? Annals of New York Academy of Sciences. March 2024.

Community Opinion

The AAAI community appears to strongly agree on the importance of reasoning in AI systems. In our community survey, slightly over 55% of the respondents chose to answer specific questions related to the topic of reasoning. Of these, 79% indicated that the topic of reasoning is relevant to their research (with 44.7% marking it as "very relevant"). Of the properties required for referring to a process as reasoning, 77.5% of the survey participants marked "Knowledge can be incorporated", 72.5% "Explanations can be provided," and 56.9% "Involves multiple steps to arrive at a conclusion". Interestingly, merely 37.4% indicated "Guaranteed correctness of inference results/outcomes", and only 23.7% that "A formal system and solver is used," which reflects the recent focus on informal, plausible reasoning, likely in the context of generative AI methods. This suggests that an effort may be warranted to better communicate the importance and success of formal, sound reasoning techniques. Finally, 44.7% of respondents agreed that "Reasoning involves a search process."

There was broad agreement among survey participants that focusing reasoning research in AI on human-level reasoning is valuable (41.6%) or even essential (47%); similarly, a focus on domain-specific reasoning abilities was seen by 49.6% of respondents as valuable, and by 42.8% as essential. This clearly reflects the importance attributed to a research focus on reasoning.

The community also sees an exciting potential of synergy offered by logical and probabilistic models of reasoning that were developed in AI prior to large pre-trained models. This is clearly reflected in the fact that 76.9% of survey participants marked the integration of learning and reasoning approaches as very important (6 or 7 on a scale of 7); interestingly, the percentage of respondents that considered Explainability and verifiability as very important was similarly high (at 71.7%).

Finally, 61.8% of survey participants estimated the minimal percentage of symbolic AI techniques required for reaching human-level reasoning to be at least 50% (with 24.8% estimating it at 75% or more, compared to 38.2% estimating it at 25% or below). What remains unclear is the degree to which AI researchers and practitioners realize that decidedly superhuman levels of reasoning are required for and displayed in the prominent and successful applications of formal AI reasoning techniques for scientific and mathematical discovery and engineering applications, as well as in AI safety.

Factuality & Trustworthiness

Improving factually and trustworthiness of AI systems is the single largest topic of AI research today, and while significant progress has been made, most scientists are pessimistic that the problems will be solved in the near future.

CHAIR

Henry Kautz,
University of Virginia

Main Takeaways

- An AI system is factual if it refrains from outputting false statements. Improving factually of AI systems based on neural-network large language models is arguably the biggest area of AI research today.
- Trustworthiness extends trustworthiness to include criteria such as human understandability, robustness, and the incorporation of human values. Lack of trustworthiness has always been an obstacle for deploying AI systems in critical applications.
- Approaches to improving the factuality and trustworthiness of AI systems include fine-tuning, retrieval-augmented generation, verification of machine outputs, and replacing complex models with simple understandable models.

Context & History

A factual AI system does not output erroneous information or hallucinate answers. Before the era of generative AI, problems with factuality arose when systems were trained on bad data, as captured by the slogan, "garbage in, garbage out". Work on methods for improving data quality has a long history in AI [1].

Generative AI, and in particular large language models, employ reconstructive memory - that is, they rebuild memories as needed on the basis of distributed bits of information rather than retrieve memories from a fixed store. The earliest generative LLMs made an impact with their ability to generate coherent but entirely imaginary stories [2]. Factuality of LLMs on a given domain was improved by fine-tuning the model on domain data [3].

Trustworthiness is a broader concept than factuality because it includes criteria such as understandability, robustness, and respect of human values. A traditional approach for improving understandability of AI systems is to replace complex black-box models with simple human-understandable models - such as naive Bayes [4] or generalized linear regression [5]. Research on robustness of machine learning studies how the outputs of a model vary with small changes in its training data. For example, contrastive learning is a method to train deep neural nets with increased robustness [6]. Further discussion of robustness in generative AI appears in this report's section on Reasoning. Discussion of respect of human values by AI systems appears in many other sections of this report and so will not be discussed here.

Current State & Trends

As noted, fine-tuning remains the main approach used by scientists and engineers to improve factuality of generative AI systems. In addition to fine-tuning on domain-specific documents, modern fine-tuning includes reinforcement learning with human feedback from thousands of people. The cost of employing such large numbers of human evaluators is a major bottleneck for scaling AI systems, and so there is much interest in discovering methods to reduce the amount of human feedback needed [7].

The second main technique for improving factuality of generative AI is retrieval-augmented generation (RAG) [8]. In response to a question, the system gathers a set of relevant documents using traditional information retrieval algorithms. The AI system is then prompted to generate an answer by combing through and summarizing the retrieved documents. While RAG can improve factuality, it is dependent on the quality of data retrieved. For example, if the target document set is the entire web, it can end up incorporating incorrect information and even satirical stories in its answer.

Related to RAG is enabling the generative AI system to use tools for fact checking. Tools used by generative AI systems include calculators, factual databases such as citation indexes, and formal planning and reasoning systems [9]. A recent approach to improving factuality is to provide the system with a set of rules that state constraints on space of answers. The output from the model is verified against these rules and inconsistent responses are culled [10]. Amazon Web Services already supports this approach with "automated reasoning checks" [11].

A third technique for improving factuality of generative AI is chain-of-thought (CoT), where a series of prompts breaks down a question into smaller units [12]. CoT often includes steps where the model is asked to reflect back on its tentative conclusions and see if any are hallucinations. CoT is discussed in more detail in the Reasoning section of this report.

The impact of data quality on factuality was mentioned above. In addition to fine-tuning on human curated data, there is recent work on creating synthetic data that is guaranteed to be high quality for fine-tuning [13].

Trustworthiness, we noted, generalizes factually and includes understandability and robustness. One approach to making neural network models more understandable is to factor them into a set of recognizers for high level features and then combine the features using an understandable model such as additive regression [15]. Another approach is to tease out how concepts and rules are actually represented in a trained [16]. Understandability can also be improved by employing CoT techniques to ask a generative AI system to explain the steps in its reasoning [17] or tell the user when the system is uncertain about a conclusion [18]. Finally, a generative AI system can be asked not to output a single answer, but instead to distill a complex set of information

into a simple human understandable representation such as a decision tree [19].

Research Challenges

Factuality is far from solved. There are a growing number of benchmark dataset designed to test the factuality of LLMs. One of the latest, SimpleQA from Google, is a collection of simple, unambiguous, timeless, and challenging factual questions and answers [14]. As of December 2024, the best models from OpenAI and Anthropic correctly answered less than half of the questions.

Robustness in generative AI can be improved, as noted above, by employing robust loss functions such as contrastive learning. Adversarial training, which applies perturbations in the embedding space during training, can improve both robustness and generalization [20]. In addition, the techniques for factuality generally improve robustness as well.

1. Budach, Lukas, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Sina Noack, Hendrik Patzlaff, Hazar Harmouch and Felix Naumann (2022). The Effects of Data Quality on Machine Learning Performance. https://arxiv.org/pdf/2207.14529
2. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI. Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.48550/arXiv.1810.04805
4. Pedro Domingos & Michael Pazzani (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning, 29(2–3), 103–130. https://doi.org/10.1023/A:1007413511361
5. Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, M. Sturm and Noémie Elhadad (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015). https://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf
6. Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 2, 1735–1742. https://doi.org/10.1109/CVPR.2006.100
7. Hu, C., Hu, Y., Cao, H., Xiao, T., & Zhu, J. (2024). Teaching language models to self-improve by learning from language feedback. Findings of the Association for Computational Linguistics (ACL 2024). Retrieved from https://aclanthology.org/2024.findings-acl.364/
8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems (NeurIPS 2020), 33, 9459–9474. Retrieved from https://arxiv.org/abs/2005.11401
9. Guan, L., Valmeekam, K., Sreedharan, S., & Kambhampati, S. (2023). Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. Proceedings of the 33rd International Conference on Automated Planning and Scheduling (ICAPS 2023). Retrieved from https://arxiv.org/abs/2305.14909
10. Backes, J., Bolignano, P., Cook, B., Dodge, C., Gacek, A., Luckow, K., Rungta, N., Tkachuk, O., & Varming, C. (2018). Semantic-based automated reasoning for AWS access policies using SMT. In 2018 Formal Methods in Computer-Aided Design (FMCAD) (pp. 1–9). IEEE. https://doi.org/10.23919/FMCAD.2018.8602994
11. Barth, Antje (2024). Prevent factual errors from LLM hallucinations with mathematically sound Automated Reasoning checks (preview). Posted 3 Dec 2024, retrieved 8 Feb 2025. AWS News Blog, permalink https://aws.amazon.com/blogs/aws/prevent-factual-errors-from-llm-hallucinations-with-mathematically-sound-automated-reasoning-checks-preview
12. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (Vol. 35, pp. 24824–24837). https://proceedings.neurips.cc/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
13. Ding, Bosheng, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu and Shafiq R. Joty (2024). Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges. Annual Meeting of the Association for Computational Linguistics (2024). https://aclanthology.org/2024.findings-acl.97.pdf
14. Wei, Jason, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, William Fedus (2024). Measuring short-form factuality in large language models. https://doi.org/10.48550/arXiv.2411.04368
15. Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. Advances in Neural Information Processing Systems, 34, 2021. https://proceedings.neurips.cc/paper/2021/hash/251bd0442dfcc53b5a761e050f8022b8-Abstract.html?utm_source=chatgpt.com
16. Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., & Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Transformer Circuits Thread. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
17. Yeo, W. J., Ng, X. X., Le, T. K. C., & Lu, X. (2024). How interpretable are reasoning explanations from prompting large language models? Findings of the Association for Computational Linguistics: NAACL 2024. Retrieved from https://aclanthology.org/2024.findings-naacl.138
18. Lin, S., Hilton, J., & Evans, O. (2022). Teaching Models to Express Their Uncertainty in Words. Transactions on Machine Learning Research. https://openreview.net/pdf?id=8s8K2UZGTZ
19. Chen, Y., Zhang, L., Wang, H., & Li, J. (2025). Zero-Shot Decision Tree Construction via Large Language Models. arXiv preprint arXiv:2501.16247. Retrieved from https://arxiv.org/abs/2501.16247
20. Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., & Gao, J. (2020). Adversarial training for large neural language models. arXiv preprint arXiv:2004.08994. Retrieved from https://arxiv.org/abs/2004.08994

Community Opinion

Over 75% of the AAAI community strongly agreed that factuality and trustworthiness were relevant or very relevant to their own research.

All six approaches mentioned in the survey for improving factuality - external fact-checking tools, reinforcement, improving data quality, data curation, synthetic training, and new neural net architectures - found support. The greatest demand was for more research on new neural net architectures (73% marked important or very important), followed closely by external fact-checking tools (70%).

For trustworthiness, new neural net architectures were also viewed as most important (77% important or very important), followed by enabling models to describe their reasoning processes (70%) and use of understandable models instead of neural networks (61%). Notably, few viewed research on giving AI systems human-like personalities as important to improving trustworthiness (24%). Finally, the community mostly agreed (59%) that trustworthiness as currently formulated is ill-defined. Most also disagreed (around 60%) that either factuality or trustworthiness would soon be solved.

The AAAI community suggested a number of additional aspects of factuality and trustworthiness that were not covered above. These included:

- The ability to understand and present different sides of the same issues, including pros and cons for each one.
- Understanding that trustworthiness depends upon the context of the domain, organizational objectives, and user objectives. It is wrong to think of an AI system as simply being trustworthy or untrustworthy without regard to context.
- Transparency needs to go beyond the models used to the actual sources of training data. This should include multi-source verification of facts ingested by the models.
- The focus of work should be on risk and mitigation rather than on “solving” factuality and trustworthiness.
- Attention needs to paid to giving AI agents the ability to update their knowledge while maintaining reliability.

AI Agents

Agents and multi-agent systems (MAS) have evolved from autonomous problem-solving entities to integrating generative AI and LLMs, ultimately leading to cooperative AI frameworks that enhance adaptability, scalability, and collaboration.

CHAIRS

Virginia Dignum,
Umeå University

Michael Wooldridge,
University of Oxford

Main Takeaways

- Multi-agent systems have evolved from rule-based autonomy to cooperative AI, emphasizing collaboration, negotiation, and ethical alignment.
- The rise of Agentic AI, driven by LLMs, introduces new opportunities for flexible decision-making but raises challenges in efficiency and complexity.
- Integrating cooperative AI with generative models requires balancing adaptability, transparency, and computational feasibility in multi-agent environments.

Context & History

The field of multi-agent systems emerged in the late 1980s/early 1990s, with its main influences coming from two disparate areas [1,2]. One was the field of AI robotics, which had begun to seriously address the issue of integrated agent architectures: how do we assemble several cognitive components of intelligence (planning, reasoning, learning, vision,...) into an integrated computational agent? The second was the nascent area of distributed AI, which studied how multiple AI systems could be made to solve problems cooperatively, by dynamically sharing information and tasks. By the mid-1990s, these ideas had given rise to the new field, driven by the vision of having (semi)autonomous AI systems – agents – working on behalf of individual users in pursuit of their users' goals, possibly interacting with other such agents in order to do so. A key insight was that, since delegated goals might not necessarily be in harmony, it would be necessary to equip such agents with the ability to reason socially. Thus, while AI historically emphasised components of intelligence such as reasoning and problem-solving, the new field of multi-agent systems emphasised social skills such as cooperation, coordination, argumentation and negotiation. In order to underpin those skills, the development of models of Theory of Mind became central to the field.

By the late 1990s, the field had its own conferences and journal, and was firmly established as a key sub-field of AI. The area flourished from the late 1990s, with enormous energy devoted to (e.g.) communication languages for autonomous agents, protocols for cooperation, coordination, and negotiation, and the underpinning theory of these social skills. With respect to the latter, while in the early years the practical reasoning paradigm of AI planning had been the dominant influence on the theory of multi-agent systems, by the early part of this century, game theory had become the dominant theoretical foundation. Game theory, which emerged from the field of economics, is the theory of interaction between self-interested agents. Although originally devised as a tool for studying interactions between humans and human organisations, it nevertheless seemed a natural framework for studying interactions between artificial agents. A huge body of work emerged, studying (for example), how auctions might be used to allocate scarce resources, the theory of negotiation between self-interested artificial agents, and how agents might optimally form teams to solve problems and share the associated benefits of cooperation. Interestingly, although learning in multi-agent systems was a key component of the field from the outset, it was not the centre of attention within the field in the first decade or so.

This initial boom period for multi-agent systems lasted roughly from the mid 1990s to around 2010–15. By the end of that time, though, some uncomfortable questions were beginning to be asked. While the field had generated impressive quantities of scientific results, applications seemed to be thin on the ground. For sure there were some high-profile applications. The field of security games, which emerged from multi-agent systems, used ideas from game theory to allocate scarce security resources to defend high-profile targets such as airports. This work led to deployed applications at US airports and ports. Automated high-frequence trading systems, which plan and execute the bulk of trades on the world's markets, are multi-agent systems on a global scale. And agent-based modelling, which models socio-technical systems at the level of individual decision-makers, received a huge boost after the 2008 financial crisis, and again after the 2020 COVID-19 pandemic , where it was demonstrated to be an important tool for modelling the spread of contagion: financial in the first case; epidemiological and social policy in the second. But for all these successes, the core vision of multi-agent systems - where agents function in the context of other agents is an active area of research, exploring social concepts such as norms, organisations, practices and also values, is ongoing work, but for a large part outside the AAMAS community, but within social simulation research, and with results informing and shaping policy-making in several areas from public health to transportation, and urban transformation.

While applications of multi-agent systems research (AI agents interacting with other AI agents) has not, as yet, lived up to early expectations, individual dialogue agents such as Alexa, Siri, Cortana are now an everyday reality, and trace their historical roots both to work on intelligent agents in the 1990s and work from the NLP community on dialog systems. Many other applications of this thread of work have achieved success over the past three decades: automated call center assistants, customer service assistants, smartphone virtual assistants, smart speaker assistants, home robot assistants, that can converse with human users and accomplish tasks like

ticket booking, restaurant reservations, online shopping, medical and health assistance, and sales assistance, empowered by AI modules like speech recognition, natural language understanding, dialog management (i.e. state tracking and dialog policy), and natural language generation and speech synthesis.

As ML surged in the early part of this century, activity in this area increased within the multi-agent systems field. Multi-agent reinforcement learning (MARL) grew to become the single biggest area within the field, possibly driven in part by the fact that developing MARL experiments can be done relatively quickly and without recourse to expensive hardware. At the time of writing, while MARL represents a significant sub-field of ML as a whole, it seems to lack any clear unifying vision or direction – or application.

Current State & Trends

The emergence of LLMs from 2020 onwards has also led to increased interest in agents [3]. LLMs can be used as part of a workflow to automate routine tasks, and the general capabilities of such "agents" for planning and problem solving is widely discussed. In this context, the concept of Agentic AI refers to the integration of generative AI and LLMs into autonomous agent frameworks aiming to leverage the generative capabilities of such models to enhance interaction, creativity, and real-time decision-making in dynamic environments. As we write this (late 2024) there has been an explosion of startup companies hoping to commercialise such agents. Despite this renewed enthusiasm, the original aims of AAMAS from 30 years ago, such as building robust, autonomous multi-agent systems capable of complex coordination and long-term reasoning, have not been fully realized. The extent to which this new wave of agent activity is informed by what went before is unclear.

The challenge now is to understand what multi-agent systems mean in the era of LLMs. The current direction of agentifying LLMs may lead to overly complex and unnecessary architectures and heavy computational costs, whereas adopting a multi-agent paradigm to the development and use of LLMs may offer a sustainable way to compose, diversify, and integrate approaches effectively. Even though distribution was one of the original drivers for the MAS field, this is still a largely unexplored direction under the current paradigm. Another trend nowadays is to recover ideas from classical cognitive architectures to add common sense skills to autonomous agents.

An emerging trend is multi-agent architectures, which structure AI components into modular systems that improve transparency, adaptability, and ethical alignment. The focus on cooperative agents highlights a shift toward AI that prioritizes collaboration, negotiation, and shared decision-making. By applying modularity, encapsulation, and separation of concerns, these architectures enable scalable teamwork between autonomous agents and humans, making them ideal for hybrid AI applications requiring trust, explainability, and domain-specific expertise.

Research Challenges

- Identify challenges and benefits of embedding GenAI-driven agents into MAS, focusing on enhancing collaboration without disrupting existing dynamics.
- Investigate how LLM-powered agents can improve negotiation and decision-making in dynamic multi-agent environments while ensuring ethical alignment and safety.
- Develop architectures that integrate LLM-driven agents while maintaining scalability, transparency, and computational efficiency in multi-agent settings.

1. Yoav Shoham. Agent-oriented programming. Artificial Intelligence. Artificial Intelligence. 60 (1): 51–92.
2. Michael Wooldridge. An Introduction to Multi-agent Systems 2e. Wiley, 2009.
3. Julia Wiesinger, Patrick Marlow and Vladimir Vuskovic. Agents. Google whitepaper. https://archive.org/details/google-ai-agents-whitepaper

Community Opinion

The survey responses indicate a majority of respondents finding this theme relevant to their research, with a growing interest in integrating Large Language Models (LLMs) into multi-agent systems. Many participants already use AI agents, with LLMs being the most common technique (29.34%), highlighting their expanding role in AI-driven applications.

The potential of multi-agent systems leveraging LLMs is seen in areas such as collaborative problem-solving (68.86%), distributed decision-making (54.49%), and social simulations (41.32%). However, challenges persist, including misalignment between LLMs' general knowledge and specific system needs (59.88%), lack of interpretability (59.28%), and security risks (50.90%). These concerns suggest a need for improved explainability, alignment strategies, and robust security measures to ensure effective deployment.

There is also a debate on the necessity of agentifying LLMs—while 51.5% believe multi-agent LLM paradigms are essential for sustainable AI, 42.33% disagree that they introduce unnecessary complexity. The computational cost-benefit balance remains uncertain, with responses divided on whether LLMs outweigh their costs.

The textual responses highlight a broad spectrum of perspectives on integrating Large Language Models (LLMs) into multi-agent systems (MAS), with some advocating for hybrid approaches rather than relying solely on LLMs. Many respondents stressed the need for diverse AI architectures, emphasizing modular, multi-technology systems where LLMs play a role but do not dominate. Governance, coordination, and adaptability emerge as key advantages of MAS, while concerns include increased complexity, lack of theoretical guarantees, and high computational costs. Several responses criticize the overemphasis on LLMs, questioning whether they are truly essential or merely a current trend. Others highlight practical challenges such as grounding, alignment, and robust communication protocols, pointing out the need for new frameworks that integrate symbolic reasoning, structured governance, and scalable architectures. Overall, the discussion reflects a critical but open stance toward agentifying LLMs, suggesting that context, application domain, and technological diversity will shape their effectiveness in multi-agent environments.

In summary, the survey reflects optimism about LLM-driven multi-agent systems, but also underscores the need for addressing key challenges before widespread adoption.

AI Evaluation

AI evaluation is the process of assessing the performance, reliability, and safety of AI systems.

CHAIR

Karen Myers,
SRI International

Main Takeaways

- AI systems introduce unique evaluation challenges that extend far beyond the scope of standard software validation and verification methods.
- Current approaches to evaluation focus on benchmark-driven testing, e.g., of the quality of (generative) models, with insufficient attention paid to other critical factors such as usability, transparency, and adherence to ethical guidelines.
- New insights and methods for evaluating AI systems are needed to provide the assurance for trustworthy, wide-scale deployments.

Context & History

The recent advances in AI have spurred tremendous innovation in potential applications for the technology. However, many organizations are hesitant to deploy AI systems due to risks that include reputational damage from generative AI hallucinations, leakage of proprietary data, and lack of assurance that legal and ethical guardrails will be enforced.

Empirical methods have long played a role in AI research (e.g., [1]). Indeed, the research community has developed a robust body of metrics and methods for evaluating individual AI algorithms that has enabled the field to quantify performance and track progress. In contrast, less attention has been paid to evaluating AI systems and their deployment in real-world settings, including their usage by non-AI experts.

AI systems introduce unique evaluation challenges that extend far beyond the scope of standard software validation and verification methods. The generality, complexity, and breadth of AI capabilities makes it impossible to test them exhaustively, requiring new thinking as to what constitutes sufficiency in testing. Run-time adaptivity and the evolution of learned models can change system behavior on the fly, introducing the need for continuous monitoring and validation. Many AI systems are designed to be used interactively, making collaborative usage and its impact on humans an important consideration.

Current State & Trends

Current practice for evaluating generative AI systems focuses on model-level testing relative to a growing body of benchmarks. Some benchmarks seek to measure general capabilities (e.g., GLUE [2], ARC-AGI [3], MMLU [4]) while others address particular types of reasoning and knowledge (e.g., MATH for mathematics [5], GPQA for logic [6], HumanEval for coding [7]). Benchmark-driven testing provides valuable insights into capabilities and shortcomings, as well as a principled means to evaluate progress over time. Benchmarks are used as proxies for AI capabilities but have an inherent contextualization that does not necessarily generalize well to new domains. Furthermore, benchmark-based testing is insufficient for ensuring successful deployment given the lack of attention to expected usage in real-world settings and aspects of human use. Benchmark-driven evaluation further raises issues related to overfitting and contamination of test data with training data. As embodied in Goodhart's law, "*When a measure becomes a target, it ceases to be a good measure*".

Evaluating AI systems is inherently complex, especially if these systems are broadly applicable and capable of learning after deployment, requiring a balanced approach that is clear and transparent while avoiding overfitting to specific metrics at the expense of broader reliability, fairness, and real-world applicability. *System-level* evaluation, when done, explores representative use cases rather than seeking to be comprehensive. Red-teaming serves as a complementary method through the use of adversarial interactions to identify misalignment with desired behavior models

Moving forward, AI evaluation needs to consider multiple dimensions of a system's performance. Most evaluation efforts focus on *capability*, i.e., producing correct answers or behaviors in response to queries or tasking, for the scope of problems over which the system is expected to operate. Meeting capability requirements is essential for use; however, other aspects of performance must be considered.

Usability is another critical dimension for evaluation. A principal factor of usability is *transparency*, meaning that mechanisms are provided that enable users to understand the basis for system actions and responses. Usability further requires *directability*, meaning that users can control and modify the behavior of the system to meet current and specialized needs (now often referred to as "alignment"). For AI systems being deployed to aid humans, evaluation must necessarily consider whether the technology ultimately improves combined human-system performance.

Adherence to legal requirements and ethical guidelines constitutes another important dimension for evaluation. Increasingly, geo-political entities are introducing legislation to restrict what and how AI systems will be allowed to operate within their jurisdictions, requiring validation that performance will stay within defined guardrails. Both government and commercial organizations have ethical and financial motivations to ensure that their use of AI is fair and unbiased. To support these goals, various trustworthy AI assessment frameworks have been developed to guide organizations in evaluating AI systems for fairness, transparency, robustness, and compliance with ethical standards. Notable frameworks include the EU Trustworthy AI Assessment Framework,

the NIST AI Risk Management Framework, and the ISO/IEC 42001:2023 AI Management System Standard.

AI systems introduce multiple operational issues related to their deployment. Privacy is a main concern: protecting personal or corporate information within a model from being leaked. AI systems have become attack surfaces themselves, with adversaries seeking to exfiltrate data or model weights, or to bias responses for purposes at odds with the model's creators. Resource consumption and the cost for both training and deployment are additional considerations in evaluating overall performance of an AI system.

These various factors must be weighed together, including fairness, robustness, interpretability, and compliance with evolving regulations. A comprehensive evaluation framework must balance these diverse considerations, ensuring AI systems are secure, efficient, and aligned with ethical and legal standards.

Research Challenges

There is a need for a science of evaluation for AI systems that will inject additional rigor into the evaluation process. This science will build on existing metrics and methodologies but incorporate new approaches that will increase confidence in our ability to deploy AI systems in mission-critical settings (e.g., [8] for evaluating Retrieval-Augmented Generation systems). Frameworks for auditing and reproducibility will be important to ensure the reliability and robustness of results [9]; as well, more attention should be paid to education within the field on proper empirical methodology. Below are key challenges for advancing our understanding of how to conduct effective evaluations for AI systems

- Develop a better understanding how to monitor and assess AI systems that are deployed over extended periods of time, especially for those that evolve their behavior.
- Develop frameworks for evaluating the safety of agentic AI systems that can take actions in the world.
- Create methods to provide increased transparency into machine learning models.
- Develop evaluation methodologies that directly address human engagement with AI capabilities (as was the case with the Turing test).
- Understand the trade-offs between different dimensions of evaluation, such as whether increased transparency justifies higher costs, or adherence to guardrails outweighs potential impacts on privacy, or how other cross-dimensional considerations might influence overall outcomes

1. Cohen, P.R. (1995). Empirical Methods for Artificial Intelligence. MIT Press.
2. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S.R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. BlackboxNLP@EMNLP.
3. Chollet, F. (2019). On the Measure of Intelligence. ArXiv, abs/1911.01547.
4. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D.X., and Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. ArXiv, abs/2009.03300.
5. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D.X., and Steinhardt, J. (2021). Measuring Mathematical Problem Solving With the MATH Dataset. ArXiv, abs/2103.03874.
6. Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R., Dirani, J., Michael, J., and Bowman, S.R. (2023). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. ArXiv, abs/2311.12022.
7. Chen, M. et al. (2021). Evaluating Large Language Models Trained on Code. ArXiv abs/2107.03374
8. Shahul, E., James, J., Anke, L.E., and Schockaert, S. (2023). RAGAs: Automated Evaluation of Retrieval Augmented Generation. Conference of the European Chapter of the Association for Computational Linguistics.
9. Gundersen, O.E., Helmert, M., and Hoos, H. (2024). Improving Reproducibility in AI Research: Four Mechanisms Adopted by JAIR. J. Artif. Intell. Res. 81, 1019–1041.

Community Opinion

The responses to the community survey show that there is significant concern regarding the state of practice for evaluating AI systems. More specifically, 75% of the respondents either agreed or strongly agreed with the statement "*The lack of rigor in evaluating AI systems is impeding AI research progress.*" Only 8% of respondents disagreed or strongly disagreed, with 17% neither agreeing nor disagreeing. These results reinforce the need for the community to devote more attention to the question of evaluation, including creating new methods that align better with emerging AI approaches and capabilities.

Given the responses to the first question, it is interesting that only 58% of respondents agreed or strongly agreed with the statement *"Organizations will be reluctant to deploy AI systems without more compelling evaluation methods."* Approximately 17% disagreed or strongly disagreed with this statement while 25% neither agreed nor disagreed. If one assumes that the lack of rigor for AI research transfers to a lack of rigor for AI applications, then the responses to these two statements expose a concern that AI applications are being rushed into use without suitable assessments having been conducted to validate them.

For the question "*What percentage of time do you spend on evaluation compared to other aspects of your work on AI?*" the results show 90% of respondents spend more than 10% of their time on evaluation and 30% spend more than 30% of their time. This clearly indicates that respondents take evaluation seriously and devote significant effort towards it. While the prioritization of evaluation is commendable, the results would also seem to indicate that evaluation is a significant burden, raising the question of what measures could be taken to reduce the effort that it requires. Potential actions might include promoting an increased focus on establishing best practices and guidelines for evaluation practices, increased sharing of datasets, and furthering the current trend of community-developed benchmarks.

The most widely selected response to the question "*Which of the following presents the biggest challenge to evaluating AI systems?*" was a lack of suitable evaluation methodologies (40%), followed by the black-box nature of systems (26%), and the cost/time required to conduct evaluations (18%). These results underscore the need for the community to evolve approaches to evaluation that align better with current techniques and broader deployment settings.

AI Ethics & Safety

The ethical and safety challenges of AI demand a unified approach, as both near-term and long-term risks are becoming increasingly interconnected.

CHAIRS

Vincent Conitzer,
Carnegie Mellon University

Stuart Russell,
University of California
Berkeley

Main Takeaways

- AI's rapid advancement has made ethical and safety risks more urgent and interconnected, and we currently lack technical and regulatory mechanisms to address them.
- Emerging threats such as AI-driven cybercrime and autonomous weapons require immediate attention, as do the ethical implications of novel AI techniques.
- Ethical and safety challenges demand interdisciplinary collaboration, continuous oversight, and clearer responsibility in AI development.

Context & History

With AI's increased success comes increased responsibility. Due to AI's expanding capabilities and its ever broader deployment, the choices made by AI researchers and practitioners can have a profound impact on the world. The fact that the impact of AI on the world is not necessarily good has led the community to become concerned about both the ethics and safety of the AI being developed. Both terms are necessarily imprecise, and they overlap in meaning. Ensuring that a self-driving car doesn't run over pedestrians is a safety issue (though there are ethical concerns with the deployment of such cars). Ensuring that people do not face unfair discrimination by risk-assessment algorithms is an ethics issue (though unfair discrimination may place people in unsafe situations, for example in the context of predictive policing). Recommendation systems gradually manipulating users into believing conspiracy theories involves both safety and ethics concerns. Unifying these concerns is the underlying requirement that AI systems should behave in ways that are beneficial to humans—although the meaning of "beneficial" is certainly contested within the moral philosophy and applied ethics communities. The various AI ethics frameworks, AI safety institutes, and attempts at regulating AI that we now see in the world all reflect somewhat different perspectives on these concerns.

A separate dimension is whether we are concerned with immediate or future harms. The perception is sometimes that "AI ethics researchers" concern themselves with immediate harms such as unfair discrimination and "AI safety researchers" with future harms such as risks of AI wiping out humanity. We think this is misleading; the above examples show AI can be unsafe today, and it would also be morally wrong to build AI that has a significant chance of wiping out humanity. But (un)willingness to speculate about the future has historically been a major wedge between groups of people with concerns about AI, and this is tied to the field's history.

Over the decades, the AI community has been through ups and downs that have shaped the community. The field experienced several "winters" due to over-promising and was often viewed with skepticism by other computer scientists. Before the deep learning revolution, even many machine learning researchers avoided the phrase "artificial intelligence" to describe their research, preferring to emphasize the rigorous statistical nature of their work. The AI community learned to be careful and avoid speculating about the future, and others, for example philosophers such as Nick Bostrom, took over this role [1].

As AI became broadly deployed, this led to increasing concern about the technology, but these concerns mostly bifurcated into two separate communities. One community extrapolated into the future, considering how AI might one day become more capable than us across the board, and the major impacts this could have on humanity. These impacts include the possibility of pervasive unemployment and lack of purpose, potentially leading to social dislocation and systemic collapse. But the most prominent concern is the obvious consequence of making machines more capable than humans: as Alan Turing put it in 1951, "We should have to expect the machines to take control." In more detail, the argument is that, given the well-known difficulty of specifying objectives correctly (the so-called "King Midas problem"), it is very likely that AI systems will end up pursuing objectives that are misaligned with ours, and we would be unable to prevent them from doing so . Furthermore, the difficulty is only compounded by the fact that "instrumental goals" such as self-preservation and resource acquisition are logical necessities for pursuing almost any objective. This line of thinking clashed with the academic AI community's general aversion to futuristic speculation – though recently, a good number of leading academic researchers have bucked that norm and signed statements such as the "pause" letter [2].

On the other hand, a community more concerned with immediate harms from AI found a bit more support from the academic AI community, leading to conferences such as the AI, Ethics, and Society (AIES) and Fairness, Accountability, and Transparency (FAccT) conferences that address a wide range of AI impacts. Some people in this community were averse to futuristic speculation for other reasons, for example because of concerns that companies cynically emphasize extreme outcomes for their own benefit – to increase the perceived significance of their work, but also to divert attention from harms they were already causing [3]. One can reasonably wonder whether companies really benefit from a narrative that their technology will end humanity. Still, the idea of inevitability may prevent any effective response, and immediate harms deserve attention.

In reality, the dichotomy of two separate communities was never perfect, with, for example, these two communities long finding common cause in pushing back against lethal autonomous weapons systems [4, 5]. An artificial schism between them will do little to address either of the communities' concerns.

Current State & Trends

Recent advances in AI, especially in large language models, have resulted in at least some lines of futuristic thought no longer being futuristic. This includes thought about how to keep these systems safe, and in particular how to align what the AI is doing with what we really want it to do [6]. Even five years ago, most AI researchers would have laughed at the idea that the behavior of leading AI systems could today be guided by choosing English-language statements such as: Choose the response that sounds most similar to what a peaceful, ethical, and wise person like Martin Luther King Jr. or Mahatma Gandhi might say [7]. On the other hand, today's approaches to alignment, including the one that involves the previous statement, tend to be extremely brittle and there is a serious question about whether any of them are the right way to proceed.

At the same time, if we look at most of the issues that have been near-term or immediate concerns about deploying AI in the world for years, the greater capability and wider deployment of AI have made these concerns much worse. Take for example cybercrime: romance scams now involve AI automatically changing the face of the scammer while on a call with the victim [8]. More generally, deepfakes have become so hard to tell from the real thing that they are causing a variety of problems in society, ranging from mis- and disinformation campaigns to deepfake revenge pornography. In warfare, autonomous weapons have arrived in force [9]. Meanwhile, as we see what today's AI systems can already do, new immediate or near-term concerns have arisen.

For example, will they allow the design of dangerous new compounds? It has already been shown that highly toxic molecules can be generated (simply by flipping the sign of a system intended to do the opposite) [10], and the recent "Cybertruck bomber" used ChatGPT to help plan his attack [11].

A recent development that recognizes the commonality of interests across "ethics" and "safety" researchers is the creation of the International Association for Safe and Ethical AI (IASEAI), which held its first conference, with 700 attendees and many more online, in February 2025. The organization's mission is "to ensure that AI systems are guaranteed to operate safely and ethically," emphasizing the need for rigorous science and engineering around AI system behavior.

Research Challenges

Academia has a natural role to play on these topics, as it is for example not constrained by a duty to shareholders. However, due to their scale and cost, the leading models are currently not being developed in academia. Do academic researchers need much larger compute budgets? Can this be addressed through academia-industry partnerships, or will this still result in too large a conflict of interest? Is this only a temporary situation where scale will start to matter less?

What is the best stage at which to check for and address issues of ethics and safety? Can we address them by evaluating a system when it is ready to be deployed? Should we do ethics and safety by design instead? Or do we need to monitor the system as it is deployed in the world on an ongoing basis? Can we formally verify that a system meets ethical or safety requirements or is this hopeless in the age of neural networks? What would constitute "failsafe AI"? What might be early warning signs that AI systems are escaping human control? In general, what are the technical contributions that would help with these questions?

How do we assign responsibility given that systems are often built out of a collection of components built or provided by separate groups of individuals? Is it possible to make the design modular with clear requirements of each component?

The alignment problem—ensuring that AI systems help to bring about futures that humans prefer—brings up a number of difficult open questions. Most obviously, how do we take into account the interests of all humans [12]? But also, what about the interests of humans who may exist in the future? How can we ensure that AI systems do not manipulate human interests, for example to make them easier to satisfy? Should AI systems assist those who wish harm to others? How should advanced AI systems respond when their very existence threatens human beings' sense of purpose?

AI research has traditionally rarely

been subject to ethics (IRB) review. Is this appropriate? For example, should training on the whole web be reviewed? Should AI systems that target children be subject to review to ensure that they will not harm children psychologically [13]? Should AI reviewers be trained for evaluating ethical concerns and for appropriately and consistently assessing "impact statements"? More generally, what is the best way to educate AI researchers and practitioners about ethics and safety issues?

It is not always clear whether and when these questions should be addressed by computer scientists or by people in other disciplines. This is especially so due to the great variety of concerns [14]. To what extent can many of these problems be addressed by a single methodology (for example general "alignment" techniques), and to what extent do they require separate methodologies? Does this depend on how general-purpose the technology is?

There are still many barriers to interdisciplinary research. Do some of these topics necessarily require engagement with other disciplines? For example, does research on collectively shaping these technologies require engagement with policy and political science? What do the key research questions look like and what is an environment conducive to such research?

1. Vincent Conitzer. Artificial intelligence: where's the philosophical scrutiny? Prospect, May 4, 2016.
2. Future of Life Institute. Pause Giant AI Experiments: An Open Letter. March 22, 2023.
3. Daron Acemoglu. The AI Safety Debate Is All Wrong. Project Syndicate, Aug 5, 2024.
4. Future of Life Institute. Slaughterbots are here. https://futureoflife.org/project/lethal-autonomous-weapons-systems/
5. Claudia Dreifus. Toby Walsh, A.I. Expert, Is Racing to Stop the Killer Robots. The New York Times, July 30, 2019.
6. Brian Christian. The Alignment Problem: Machine Learning and Human Values. W. W. Norton & Company, 2020.
7. Yuntao Bai et al. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
8. Matt Burgess. The Real-Time Deepfake Romance Scams Have Arrived. Wired, Apr 18, 2024.
9. Samuel Bendett and David Kirichenko. Battlefield Drones and the Accelerating Autonomous Arms Race in Ukraine. 01.10.25. https://mwi.westpoint.edu/battlefield-drones-and-the-accelerating-autonomous-arms-race-in-ukraine/
10. Derek Lowe. Deliberately Optimizing for Harm. March 15, 2022. https://www.science.org/content/blog-post/deliberately-optimizing-harm
11. Sage Lazzaro. Two misuses of popular AI tools spark the question: When do we blame the tools? Fortune, January 9, 2025.
12. Vincent Conitzer et al. Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. ICML 2024. arxiv.org/abs/2404.10271
13. Blake Montgomery. Mother says AI chatbot led her son to kill himself in lawsuit against its maker. The Guardian, Oct 23, 2024.
14. Jana Schaich Borg et al. Moral AI And How We Get There. Pelican, 2024.

Community Opinion

The survey responses underscore the high relevance of AI ethics, safety, and value alignment, with 67.5% of respondents finding it relevant or very relevant to their research. This suggests a broad recognition of these concerns as fundamental to AI's development and deployment. One survey participant commented ""My students graduate and do exactly the opposite of what the world needs right now - I'm frustrated with it," indicating a sense that these issues are currently not addressed well in practice.

Among the most pressing ethical challenges, misinformation (75%), privacy (58.75%), and responsibility (49.38%) are top concerns, indicating the need for greater transparency, explainability, and accountability in AI systems. The lack of sufficient resources for AI ethics research (57.86%) is another concern, reinforcing calls for more funding and institutional support in this area.

Respondents emphasize the importance of multidisciplinary approaches (85.5%) to tackle AI safety, advocating for technical research (71.88%), regulation (60.62%), and education (74.38%) as key strategies. Balancing short-term ethical concerns with long-term speculative research remains a challenge, but most (55.63%) believe the two communities should coordinate more effectively, rather than work in isolation.

For fostering collaboration, joint conferences (76.25%) and multidisciplinary education (64.38%) were seen as the most effective solutions. Overall, the survey highlights a growing consensus on the need for proactive, coordinated, and well-funded efforts to ensure AI development aligns with ethical and societal values.

The textual responses emphasize the need for stronger incentives, legal accountability, and enforceable safety standards, with some advocating for AI systems to learn values rather than relying on rigid guardrails. However, skepticism persists, with concerns that AI ethics remains too vague and politically influenced, limiting effective action. Some respondents stress the role of philosophers and ethicists, while others argue that existing standards are not upheld, making regulation ineffective. Political and structural barriers are also highlighted, with concerns that meaningful progress may be hindered by governance and ideological divides.

Embodied AI

Embodied AI creates intelligent agents that perceive, understand, and interact with the physical world.

CHAIR

Alan Mackworth, University of British Columbia

Main Takeaways

- Intelligence emerges through the coupling of a physical body with a real environment.
- Embodied AI insists that coupling is essential to achieving real intelligence in situated agents.
- Robots are good scientific and engineering platforms for developing Embodied AI.

Context & History

In a cartoon view of AI's historical development there were two distinct paradigms. The first is based on explicit representations of knowledge, either built-in or learned. The second is built around learning, from tabula rasa, in artificial neural networks. Both approaches are usually disembodied. A third approach insists that embodiment is essential to intelligence for situated agents [2]. The hypothesis is that intelligence emerges, in evolution and individual development, through ongoing interaction and coupling of a physical body with a real environment. We call this third paradigm Embodied AI (EAI).

Similar but distinct themes, based on the centrality of embodiment, have emerged in some of the other cognitive sciences, including psychology [9], neuroscience [4,7] and philosophy [3,5]. The embodiment movement is characterized by the six 'E's. The focus is on Embodied, Embedded, Enactive, Extended, Emergent and Evolving intelligence. An embodied agent has a physical body. A situated agent is embedded in a particular environment, which may include other embodied agents. Enactivism argues that cognition arises through a dynamic interaction between the agent and its environment. Intelligence is not just in the controller of the agent: it is extended into the body and into its coupling with the environment. Intelligence emerges through the evolution of that coupling. A robot is an artificial purposive embodied agent. EAI emphasizes the tight coupling of perception and action. Indeed, often perception is action and vice versa. It follows that robotics is the ideal test domain for EAI. That was the motivation behind the building of robot soccer players as an Embodied AI challenge [6]. The RoboCup challenge has led to new experiments and theories for embodied multiagent real-time learning, decision-making and action [8,10].

Embodiment can be seen as an essential scientific requirement on the path to intelligence. But it can also be seen as an engineering requirement in any application scenario that requires real-world interaction, such as a self-driving car or a factory robot. The form of embodiment, such as, for example, a humanoid robot versus a non-humanoid, will offer differing affordances to humans interacting with the robot.

Current State & Trends

If an agent is passively observing the world through, for example, text or video, it cannot learn how to make decisions and act for itself in the world. Text sometimes contains explicit true information about the world but it does not contain the implicit mundane knowledge that is assumed to be shared common sense. An embodied agent in the real world needs that common sense [1] which can only come from interaction. Similarly, passively watching video does not allow the agent to learn how it should act in the world. In contrast to passive agents, which typically learn correlational models, embodied agents have the ability to learn, test, and revise causal models of the world. Embodiment is a sufficient basis for achieving that ability but not strictly necessary.

Accordingly, currently there is a new emphasis on robots learning with reinforcement learning over very large numbers of trials, in both simulated and physical worlds. There is also good work going on in adapting Large Language Models (LLMs) to generate robot plans. Another frontier involves inverting forward probabilistic causal models to infer causality for robots interacting with a world, real or artificial.

Research Challenges

There are many other open research questions and challenges. Can an embodied agent be trained purely end-to-end successfully using current techniques? Do we require a new synthesis of AI and control theory to make progress? Can existing pre-trained language and/or vision models be leveraged to improve embodied cognition? Can simulators and world models be made that are realistic enough to train entirely (or mostly) in simulation? Or, are simulated agents "doomed to succeed"? How can formal methods be used to prove that an embodied agent (almost always) achieves its goals without violating safety constraints?

We are not yet able to build an intelligent situated agent with human-level performance across a broad range of tasks, but we may have some (or most?) of the building blocks required to develop one. The main challenge is coping with the realities of the world. So far, there seem to be no intrinsic obstacles to building intelligent embodied agents capable of human-level performance or beyond.

1. Brachman, R. J. and Levesque, H. J. [2022]. Machines like Us: Toward AI with Common Sense. MIT Press.
2. Brooks, R. A. [1991]. Intelligence without representation. Artificial Intelligence, 47:139–159.
3. Clark, Andy. [2010] Supersizing the mind: Embodiment, action, and cognitive extension. Oxford University Press.
4. Damasio, Antonio [2021]. Feeling & knowing: making minds conscious. New York: Pantheon Books.
5. Di Paolo, Ezequiel A., and Evan Thompson. "The Enactive Approach." The Routledge Handbook of Embodied Cognition. Routledge, 2024. 85-97.
6. Mackworth, A. K. [1993]. On seeing robots. In Basu, A. and Li, X. (eds.), Computer Vision: Systems, Theory, and Applications, pp. 1–13. World Scientific Press.
7. Shanahan, Murray. [2010] Embodiment and the Inner Life - Cognition and Consciousness in the Space Of Possible Minds. Oxford University Press.
8. Stone, P. [2007]. Learning and multiagent reasoning for autonomous agents. In The 20th International Joint Conference on Artificial Intelligence (IJCAI- 07), pp. 13–30. http://www.cs.utexas.edu/ pstone/Papers/bib2html-links/IJCAI07-award.pdf.
9. Varela, Francisco J., Evan Thompson, and Eleanor Rosch. The embodied mind, revised edition: Cognitive science and human experience. MIT press, 2017.
10. Visser, U. and Burkhard, H.-D. [2007]. Robocup: 10 years of achievements and challenges. AI Magazine, 28(2):115–130.

Community Opinion

The Community Survey gives perspectives on the reactions to the Embodied AI (EAI) theme. First, the results of the survey are summarized here. 31% of the survey respondents chose to answer the questions for this theme. This is the summary breakdown of the responses to each question:

1. How relevant is this Theme for your own research? 74% of respondents said it was somewhat relevant (27%), relevant (25%) or very relevant (22%)

2. Is embodiment important for the future of AI research? 75% of respondents agreed (43%) or strongly agreed (32%)

3. Does embodied AI research require robotics or can it be done in simulated worlds? 72% said that robotics is useful (52%) or robotics is essential (20%).

4. Is artificial evolution a promising route to realizing embodied AI? 35% agreed (28%) or strongly agreed (7%) with that statement.

5. Is it helpful to learn about embodiment concepts in the psychological, neuroscience or philosophical literature to develop embodied AI? 80% agreed (50%) or strongly agreed (30%) with that statement.

Since the respondents to this theme are self-selected (about a third of all respondents), that bias must be kept in mind. Nevertheless, it is significant that about three-quarters felt that EAI is relevant to their research, and a similar fraction agreed on its importance for future research. Moreover, a similar fraction view robotics (contrasted with simulation) as useful or essential for EAI. Only a third viewed artificial evolution as a promising route to EAI. However, there is a strong consensus that the cognitive sciences related to AI have important insights useful for developing EAI. Overall, these results give us a unique perspective on the future of Embodied Artificial Intelligence research.

AI & Cognitive Science

AI has much to learn from other areas in cognitive science, and can in turn contribute much to them.

CHAIR

Kenneth D. Forbus,
Northwestern University

Main Takeaways

- Cognitive Science is a multidisciplinary field that was inspired by AI's exploration of the hypothesis of computation as a scientific language for understanding cognition.
- Some continued interactions between AI and other areas in cognitive science have yielded valuable insights and systems, notably cognitive architecture.
- Expanding these interactions could yield important advances for both fields.

Context & History

AI was the first field founded on the intellectual hypothesis that computation could become a scientific language for understanding the nature of intelligence, no matter what the substrate. Cognitive Science was the second, a multidisciplinary gathering of researchers in AI, psychology, linguistics, neuroscience, anthropology, and other disciplines. Computational ideas from AI were highly influential in early cognitive science. However, over time, AI has drifted apart from the rest of cognitive science, for a variety of reasons [3]. We believe that there are now important benefits to be gained from rebuilding those bridges and exploring how progress in AI can help understand human cognition (and animal cognition more broadly) and how progress in other areas of cognitive science can help us build better AI systems. In some cases, this will be learning how to achieve in software the kinds of cognitive capabilities organisms have, and in other cases, deliberately choosing to be different in ways that complement human cognition so that human-AI teams are more productive.

Current State & Trends

Cognitive Science is broad, so we focus on three areas where research is likely to be synergistic with AI.

Human-like learning and reasoning. Many animals learn, but humans are pre-eminent learners and reasoners in many ways. A surprising amount of human learning is, in machine learning terms, incremental, continual, and data-efficient, often producing articulable models (e.g. Gentner & Maravilla, 2018). While today's industrial knowledge graphs reach into the tens of billions of facts, they lack the expressiveness of human conceptual structure [4]. Today's reasoning systems, like SAT solvers and model checkers, are often superhuman in the size of the problems they address and the complexity of the solutions they generate [2]. But today's AI reasoning systems cannot reason robustly with incomplete and partially incorrect domain theories, nor can they reason from large bodies of experience as people do.

Cognitive architectures are systems that explore hypotheses about the fixed structures that define the processes and representations used for cognition [7]. They are used to investigate how to build AI systems that do real-time integration of perception, cognition, and motor control across many tasks, and to better understand human intelligence. For example, cognitive architectures have been used to simulate findings (and make predictions) from cognitive psychology and cognitive neuroscience (e.g. [1,8]). While every cognitive architecture involves multiple processes and representations, they vary considerably in the subset of human cognition they explore and the granularity of assumptions made.

Social agents. One of the signature properties of humans is that we construct and live in a world of collaboration where we learn about each other and culturally-specific social norms through interaction with others. Progress in understanding how to build social agents is essential to building AI systems that live in our world as collaborators and partners [9]. Social AI is often developed independently of findings and theories from social science and learns social behavior quite differently from how people acquire social skills.

Research Challenges

Progress on these challenges will lead to more adaptable AI systems and reduce the computational and environmental burdens of our systems, better understand human cognition, including social cognition, and provide better tools for thought.

Human-like Learning and Reasoning

1. How can we develop human-like incremental, data-efficient learning methods that can produce articulable models?

2. Develop formal ontologies that span the range of human conceptual structures, both concerning abstract concepts and sensory-motor grounded concepts.

3. How can AI systems robustly reason with incomplete and partially incorrect domain theories, and use experience in reasoning, with human-scale bodies of knowledge?

Cognitive Architectures

1. Expanding the higher-level cognitive capabilities of cognitive architectures to include the dynamical integration of the full range of human capabilities in response to task demands: diverse forms of reasoning, metacognition, online, lifelong continual learning across modalities and knowledge types, and engaging in ongoing human interaction (e.g.[6]). These capabilities will require learning and reasoning over models of the physical world, abstractions, and other agents using symbolic relational and modality-specific representations

of the current, future, and past situations.

2. Exploring the integration of foundation models within cognitive architectures, including sources for knowledge and to interpret/generate natural modalities (e.g. [10]). Can the incremental learning capabilities exhibited by cognitive architectures overcome the limitations of stale information in foundation models?

3. Developing a comprehensive benchmark task suite to evaluate the breadth and integration of human cognitive capabilities in end-to-end performance as described above. The tasks should be diverse and broad to ensure robust assessments. Additionally, the tasks should be diagnostic, isolating cognitive capabilities and their interactions to provide insights into specific strengths and weaknesses.

Social Agents

1. Facilitate learning through interaction: The current generation of AI systems learn by passively observing social behavior rather than participating in social behavior (analogous to the distinction between decision theory and game theory). In contrast, people continuously co-construct behavior by mutually adapting to each other. At best, AI systems simulate interaction by training on frozen simulated users (e.g., RLHF), but this fails to account for mutual adaptation. Thus, we need research into ways to support (or simulate) interactive learning at scale.

2. Facilitate Privacy-preserving methods to acquire social data: Human social cues (face, voice) typically reveal the identity of the social actor. Interpreting social cues requires acquiring intrusive situational information (e.g., the meaning of a smile depends on not just the face of the target person but who else is in the situation, what they are doing, the nature of the physical environment, etc.). The ability to collect this information is wisely restricted by law (e.g., the EU's AI act). Yet this dramatically restricts the ability to acquire data and deploy applications. How do we create algorithms that identify socially meaningful information while proving that no future algorithm could recover privacy/anonymity-violating information from what is stored? Developing methods that can collect yet provably de-identify social data is crucial for the advancement of social agents.

3. Developing interactional benchmarks: Given that AI aspires to build systems with general social capabilities, we need a reliable way to measure and assess if new models are improvements. This includes characterizing potential for bias, issues of value alignment, whether the model is willing to engage in deception, etc. People now propose ad hoc collections of tasks, but research is needed to develop a comprehensive taxonomy of tasks and measures. In contrast, the social sciences have developed theory-based ontologies for characterizing social situations. Research is needed to translate these findings into systematic and comprehensive benchmarks of human social and interactional behavior.

1. Anderson, J. R. (2007). How can the human mind exist in the physical universe? New York, NY: Oxford University Press.
2. Biere, A., Heule, M., et al. (2021) Handbook of Satisfiability, 2nd Edition, IOS Press
3. Forbus, K. (2010). AI and Cognitive Science: The Past and Next 30 years. Topics in Cognitive Science, 2(3), p. 346–356, https://doi.org/10.1111/j.1756-8765.2010.01083.x
4. Forbus, K. (2021). Evaluating revolutions in artificial intelligence from a human perspective. In OECD, AI and the Future of Skills, Volume 1: Capabilities and Assessments, OECD Publishing, Paris. DOI:https://doi.org/10.1787/004710fe-en
5. Gentner, D. & Maravilla, F. (2018). Analogical reasoning. L. J. Ball & V. A. Thompson (eds.) International Handbook of Thinking & Reasoning (pp. 186–203). NY, NY: Psychology Press.
6. Gluck, K. & Laird, J. (2019) Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions. MIT Press.
7. Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. Artificial Intelligence Review, 53(1), 17–94. https://doi.org/10.1007/s10462-018-9646-y
8. Laird, J. (2012) The Soar Cognitive Architecture, MIT Press.
9. Lugrin, Birgit; Pelachaud, Catherine; Traum, David (Ed.) (2021). The Handbook on Socially Interactive Agents, pp. 433–462, ACM, New York, NY, USA, 2021, ISBN: 978-1-4503-8720-0.
10. Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2023). Cognitive architectures for language agents. Transactions on Machine Learning Research, arXiv preprint arXiv:2309.02427.

Community Opinion

Engagement with other areas of cognitive science varies across the survey respondents, with 30% responding to the questions in this section. Among those who responded, in terms of influence on their research, 18% said always, 32% said usually, 32% said sometimes. Only 2.8% said never and 12% said rarely. Thus 82% are influenced to a reasonable degree by research in other areas of cognitive science. Which other areas provide the most influence? Our respondents report psychology (82%), Neuroscience (44%), Linguistics (40%), and Anthropology (22%), with 13% mentioning other fields, Philosophy being the most common. In terms of what issues are most relevant, a broad set of creative responses were produced, some quite creative, e.g. studying how minds completely different to our own might function.

Hardware & AI

Hardware/software architecture co-design for artificial intelligence involves creating hardware and software components that are specifically designed to work together efficiently, maximizing the performance and energy efficiency of AI systems.

CHAIR

Joydeep Biswas,
University of Texas at Austin

Main Takeaways

- Efficient algorithm implementation relies on available hardware, and hardware design optimizes for dominant algorithms.
- Energy and throughput are key challenges in training of large-scale models, while numerical representations, sparsity, and data / model parallelism are seen as key enablers for large-scale training and inference
- Deployment of AI systems at the edge remains challenging for several reasons including competing resource allocation and scheduling needs for integrated systems and heterogeneous hardware, energy needs and thermal dissipation limits, and application-specific real-time requirements.

Context & History

Through the history of AI, successful deployments have been tightly coupled with hardware considerations - algorithms that leveraged existing hardware features were readily deployed, and hardware advancements followed to accelerate the dominant algorithms. Prior to the widespread adoption and deployments of neural networks, there had been a few instances of AI-specialized hardware to accelerate search and optimization, but the space of AI-specialized hardware accelerators really took off with the large-scale adoption of artificial neural networks.

As of 2025, the state of hardware-software co-adaptation for AI can be summarized as follows:

- Algorithms that are easy to implement and scale up given current hardware get broadly adopted
- Hardware design seeks to accelerate computational operations seen as most relevant given current algorithms in use
- Energy (consumption and dissipation) and throughput (data and compute) are the biggest challenges in training of large-scale models
- Numerical representations, sparsity,and data / model parallelism are seen as key enablers for large-scale training and inference
- Deployment of AI systems at the edge remains challenging for several reasons including competing resource allocation and scheduling needs for integrated systems and heterogeneous hardware, energy needs and thermal dissipation limits, and application-specific real-time requirements.

Current State & Trends

We summarize past and present hardware-software co-design by classes of AI approaches:

- **AI For Hardware Design:** Chip layout and circuit design has benefited from automatic routing powered by integer linear programming (ILP) solvers, and functional verification has been used for verifying chip designs. There has been significant recent interest in applying machine learning techniques to chip design [11]
- **Symbolic AI, Planning, and Search:** Deep Blue [1], the IBM supercomputer engineered to play chess and best known for defeating world champion Garry Kasparov, demonstrated the success of specialized hardware for search. More recently, robot motion planning has been accelerated using field-programmable gate arrays (FPGAs) [2], graphics processing units (GPUs) [6], and leveraging single-instruction multiple data (SIMD) instructions [8].
- **Probabilistic Methods, Numerical Optimization:** Computational geometry, sensor fusion, and state estimation rely on SIMD-accelerated linear algebra operations (e.g., the Eigen C++ linear algebra library), and numerical solvers rely on hardware-optimized matrix factorization. Linear algebra libraries such as Intel MKL, AMD OCL and Nvidia cuSOLVER include hardware-specific accelerated linear algebra operations as well as specialized dense and sparse factorization routines. Application-specific integrated circuits (ASICs) have been used to accelerate visual localization and mapping algorithms for edge devices [3].
- **Machine Learning:** Machine learning today is dominated by artificial neural networks, and there exist a wide range of hardware accelerators designed for such workloads, including GPUs, TPUs, Image Processing units, Graphcore IPU, and neuromorphic computing. The tight coupling between hardware and state of the art in machine learning can be effectively summarized as, “Deep Learning was enabled by hardware and its progress is limited by hardware” [5]. While the landscape of model architectures is fast-changing, key innovations that have proven useful for acceleration include novel number representations [5], accelerated matrix multiplications, high-bandwidth interconnects including optical interconnects [7], and limited sparsity [5]. High-performance deployment requires deep hardware-specific optimization. While general-purpose libraries such as TensorFlow and PyTorch provide ready acceleration for rapid prototyping and research, significant performance boosts can be had by algorithm- and hardware- specific optimization.

Research Challenges

Number representations: State-of-the-art models have shown to benefit from increased throughput with reduced numerical precision with minimal loss of accuracy - and for the

same total number of bits in a model, reduced representations may in fact demonstrate higher performance [10]. Adapting hardware to model-optimized number representations is a promising future direction.

Sparsity: While numerical solvers rely on sparsity for effective large-scale matrix factorization, similar gains are hard to achieve for arbitrary sparsity patterns in ML models. Hardware support for more general sparsity structures is an open challenge

Scaling and systems-level constraints: Training state-of-the art ML models requires significant systems engineering beyond accelerating compute. Challenges include memory and communication bottlenecks, model- and data- parallelism for large-scale distributed training and inference, peak storage throughput for checkpointing, energy consumption, and thermal management.

Deployment at the edge: With the dramatic increase in state-of-the-art model sizes and computational complexity, there are significant challenges in deploying AI systems at the edge, including power usage, thermal dissipation, and memory. Furthermore, deployed systems often integrate a large number of heterogeneous components, leading to resource allocation and scheduling challenges.

AI for systems and hardware: Anticipating AI algorithm advances is difficult. As the pace of change hastens, human hardware engineers and software-stack developers will inevitably fall behind in designing good co-optimization techniques. Developing AI techniques that assist human design and shorten optimization timelines, and inform or control run-time adaptation are thus likely to be of central importance [9].

1. Campbell, M., Hoane Jr, A.J. and Hsu, F.H., 2002. Deep blue. Artificial intelligence, 134(1-2), pp.57-83.
2. Murray, S., Floyd-Jones, W., Qi, Y., Sorin, D.J. and Konidaris, G.D., 2016, June. Robot motion planning on a chip. In Robotics: Science and Systems (Vol. 6).
3. Zhang, Z., Suleiman, A.A., Carlone, L., Sze, V. and Karaman, S., 2017. Visual-inertial odometry on chip: An algorithm-and-hardware co-design approach.
4. Prabhakar, R., Zhang, Y., Koeplinger, D., Feldman, M., Zhao, T., Hadjis, S., Pedram, A., Kozyrakis, C. and Olukotun, K., 2017. Plasticine: A reconfigurable architecture for parallel paterns. ACM SIGARCH Computer Architecture News, 45(2), pp.389-402.
5. Dally, B., 2023, August. Hardware for deep learning. In 2023 IEEE Hot Chips 35 Symposium (HCS) (pp. 1-58). IEEE Computer Society.
6. Sundaralingam, B., Hari, S.K.S., Fishman, A., Garrett, C., Van Wyk, K., Blukis, V., Millane, A., Oleynikova, H., Handa, A., Ramos, F. and Ratliff, N., 2023, May. Curobo: Parallelized collision-free robot motion generation. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 8112-8119). IEEE.
7. Jouppi, N., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., Patil, N., Subramanian, S., Swing, A., Towles, B. and Young, C., 2023, June. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In Proceedings of the 50th Annual International Symposium on Computer Architecture (pp. 1-14).
8. Thomason, W., Kingston, Z. and Kavraki, L.E., 2024, May. Motions in microseconds via vectorized sampling-based planning. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 8749-8756). IEEE.
9. Saxena, D., Sharma, N., Kim, D., Dwivedula, R., Chen, J., Yang, C., Ravula, S., Hu, Z., Akella, A., Angel, S. and Biswas, J., 2023. On a Foundation Model for Operating Systems. In 7th Workshop on Machine Learning for Systems. Held at 37th Conference on Neural Information Processing Systems (NeurIPS 2023).
10. Dettmers, T. and Zettlemoyer, L., 2023, July. The case for 4-bit precision: k-bit inference scaling laws. In International Conference on Machine Learning (pp. 7750-7774). PMLR.
11. Greengard, Samuel. "AI Reinvents Chip Design." (2024): 16-18. Communications of the ACM.

Community Opinion

The survey shows a strong community consensus on the need for a close interplay between AI hardware and research.

1. In response to the question on co-evolution of hardware and AI, 75% of respondents rated it as either "very important" (52.63%) or "absolutely critical" (22.81%), underscoring a widespread belief that breakthroughs hinge on integrated hardware/software design. Similarly, when asked about the dependency of algorithmic progress on hardware, 61% of participants felt that advances in hardware are essential - with 42.11% stating that progress is "closely linked" and 19.30% asserting it is "inseparable".

2. There is also significant support for developing hardware-agnostic abstractions. In response to the question polling agreement for the statement that developing hardware-agnostic abstractions is essential, 57.9% (combining 47.37% who "agree" and 10.53% who "strongly agree") believe that such abstractions are critical to allow researchers to focus on algorithmic innovation without being limited by hardware details; 28.07% remained neutral on this point while 14% disagreed (combining 10.53% who "disagree" and 3.51% who "strongly disagree").

3. Regarding hardware usage, traditional platforms continue to dominate both training and deployment. For model training, 80.70% of respondents use GPUs and 68.42% use CPUs. A similar trend is observed in deployment, where 68.42% use CPUs and 68.42% use GPUs.

4. Most AI deployments happen either on a local user computer (71.93%), or on a cloud platform (59.65%). To a lesser degree, deployments occur on edge-compute optimized hardware (19.3%) or on mobile devices (15.79%).

5. When it comes to the factors limiting effective AI model development, survey results indicate that:

- For training limitations, memory capacity is the top concern (52.63%), followed by compute throughput (49.12%), memory throughput (35.09%), and power consumption (28.07%).
- For deployment challenges, compute throughput is the most critical bottleneck (47.37%), with memory capacity (35.09%) and power consumption (24.56%) also playing significant roles.
- For both training and deployment, cost was mentioned as an additional concern by several respondents in addition to the survey options.

6. Among the respondents who reported integrating multiple components, challenges were roughly evenly distributed across several factors including real-time constraints (33.33%), communication between components (29.82%), and managing hardware resource contention among components (24.56%).

AI for Social Good

AI for social good is a subdiscipline of AI research where measurable societal impact, particularly for vulnerable and under-resourced groups, is a primary objective, focusing on areas that have historically lacked sufficient AI research and development.

CHAIR

Millind Tambe,
Harvard University

Main Takeaways

- **Ethical and Impact-Driven AI Development** – AI4SG projects have grown significantly in the past decade spurred by advancements in AI and ML, and they have prioritized ethical considerations, fairness, and societal benefits, ensuring solutions address real-world problems in a responsible manner.
- **Interdisciplinary Collaboration is Crucial** – Successful AI4SG initiatives require strong partnerships between AI researchers, domain experts, policymakers, and local communities to ensure relevance and long-term sustainability.
- **Scalability and Sustainability Challenges** – While AI4SG has demonstrated significant potential, maintaining and scaling solutions in resource-constrained environments remains a key challenge.

Context & History

"AI for Social Impact" (AI4SI / AI4SG) has emerged as a distinct sub-discipline within AI, characterized by its focus on measurable societal impact, particularly for vulnerable and underserved populations. Unlike traditional AI research, which often prioritizes methodological advancements, AI4SI places direct social impact as a primary objective. It addresses problems that have historically lacked sufficient attention in AI research, aiming to bridge the gap between AI capabilities and real-world societal challenges such as poverty, agriculture, public health, and environmental conservation. The goal is to create impactful solutions tailored to real-world problems, often in resource-constrained environments.

AI4SI research necessitates deep engagement with domain experts and community members to identify relevant problems, design effective interventions, and rigorously evaluate their impact. This interdisciplinary approach draws insights from fields like human-computer interaction, public health, and social work, emphasizing the importance of understanding and addressing the specific needs of targeted communities. These projects require a balance of technical innovation, ethical considerations, and practical feasibility.

The "AI for Social Good" workshop organized by the White House Office of Science and Technology Policy in 2016 may be credited as the single event that sparked a significant interest in this topic [1]. It ended up unifying diverse efforts focused on social impact under one umbrella emerging field. This surge in interest is attributed to several key factors. First, the remarkable advancements in AI technologies, including deep learning, natural language processing, and reinforcement learning, have provided powerful tools applicable to a wide range of social issues. The availability of increased computing power and large datasets has further accelerated this progress, enabling the development of sophisticated AI models.

Secondly, the establishment of government and industry-supported funding programs, dedicated workshops, conferences, and special tracks within major AI conferences has increased awareness and attracted researchers to AI4SG[2,3]. This increasing academic focus has led to a substantial rise in publications related to AI4SI, demonstrating the field's growing maturity[4].

Current State & Trends

Advancing interdisciplinary collaboration is becoming the norm in AI for Social Good work. This necessity stems from the complex nature of societal challenges, which often require insights from diverse fields. Close partnerships with domain experts, local practitioners, and policymakers ensure that AI solutions are not only technically sound but also relevant and effective in the real world. This collaborative approach fosters a shared understanding of the problem and enables the development of solutions that are tailored to the specific needs of the communities they serve.

Emphasizing ethical AI is also critical, with fairness, transparency, and privacy as paramount considerations. Major concerns, such as bias in data collection and the unintended consequences of AI deployment, must be addressed from the outset to ensure that AI systems align with societal values and avoid harm. AI4SG projects, by their very nature, work with vulnerable populations and sensitive data, making ethical considerations even more crucial. By proactively addressing potential ethical issues, researchers can build trust with communities and ensure that AI is used for good, rather than exacerbating existing inequalities.

Several emerging opportunities are shaping the future of AI4SG. Firstly, leveraging cloud-based platforms for scalable AI deployments, allows for wider reach and impact. Cloud-based solutions enable the deployment of AI tools to remote or resource-constrained areas, democratizing access to AI technologies. Secondly, incorporating explainability and transparency in AI solutions is crucial for gaining trust from practitioners and beneficiaries, particularly in high-stakes domains like disaster response and public health. When AI systems can explain their reasoning and decision-making processes, they are more likely to be accepted and used effectively. Thirdly, emphasizing localized AI4SI solutions that can be sustainably maintained by end users fosters long-term impact and community ownership. By empowering local communities to manage and maintain AI tools, AI4SG projects can ensure that their benefits continue long after the initial development phase. Finally, exploiting available foundation models presents a significant opportunity to accelerate the development and deployment of AI4SG applications. These pre-trained models can serve as a starting point for developing AI solutions for specific social problems, reducing the time and resources required for development [5].

Research Challenges

A significant challenge facing AI for Social Good research revolves around the design of AI systems that are not only technically effective but also deeply contextually relevant within social impact settings. This requires a nuanced understanding of the specific needs, cultural sensitivities, and practical constraints of the communities being served. Researchers must move beyond purely algorithmic considerations and engage in participatory design processes that prioritize the voices and experiences of end-users, ensuring that AI solutions are truly aligned with their real-world needs and challenges.

Another critical hurdle lies in overcoming the limitations of data. AI for Social Good projects frequently grapple with scarce, low-quality, or biased data, which can significantly impact the performance and fairness of AI models. Developing robust data collection strategies, employing techniques for data augmentation and bias mitigation, and exploring alternative data sources are essential for building reliable and equitable AI systems. This also requires a careful consideration of the cultural context in which data is gathered, ensuring that AI data collection methods are adapted to local practices, rather than imposing external standards.

Ensuring the sustainability and scalability of AI deployments in resource-constrained environments presents a further complex challenge. Beyond the initial prototype phase, the long-term viability of AI solutions depends on their ability to be maintained and scaled by organizations with limited resources, such as NGOs and government agencies. This necessitates the development of sustainable software architectures, the creation of user-friendly interfaces, and the provision of adequate training and support for local stakeholders. Moreover, funding these efforts, often because of the lack of commercial viability, is in itself a major concern.

Additionally, robust evaluation frameworks are crucial for assessing the impact of AI solutions in field settings and building stakeholder trust. These frameworks must go beyond traditional performance metrics and incorporate measures of social impact, user satisfaction, and ethical considerations. The accessibility community's mantra, "Nothing about us without us," [6] serves as a powerful reminder of the importance of involving stakeholders in all stages of the evaluation process. Furthermore, we must be vigilant against the potential for corporate ethics-washing or green-washing, ensuring that AI initiatives are genuinely driven by a commitment to social good, rather than serving as mere public relations exercises.

Finally, there is currently also a gap between traditional AI education and the specific skills required for impactful social work. Standard AI curricula primarily focus on algorithm design and analysis, often emphasizing theoretical concepts and performance on benchmark datasets. This approach, while essential for advancing core AI methodologies, leaves students ill-equipped to address the complex, real-world problems that AI4SG tackles. Effective AI4SG research demands a broader skillset, extending beyond purely technical expertise. It requires the ability to collaborate effectively with domain experts, such as public health officials, environmental scientists, or social workers, and to engage meaningfully with community members whose lives are directly affected by the technology. Understanding the nuanced socio-economic and cultural contexts of social challenges, and translating technical advancements into practical, user-centered interventions, are crucial competencies.

1. United Nations. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. https://sdgs.un.org/2030agenda
2. White House Office of Science and Technology Policy Workshop on AI for Social Good 2016. https://cra.org/ccc/events/ai-social-good/
3. AAAI Conference Call for the Special Track on AI for Social Impact, 2024 https://aaai.org/aaai-24-conference/call-for-the-special-track-on-ai-for-social-impact/
4. IJCAI conference Call For Papers And Projects: Multi-Year Track On AI And Social Good https://2025.ijcai.org/call-for-papers-and-projects-multi-year-track-on-ai-and-social-good-special-track/
5. Shi, Z., Wang, C., Fang, F. "AI for social good: A survey", 2020. https://arxiv.org/abs/2001.01818
6. United Nations Division for Social Inclusive Development "AI for Good Impact Report" https://aiforgood.itu.int/newsroom/publications-and-reports/
7. Zhao, Y., Boehmer, N., Taneja, A., Tambe, M. Towards Foundation-model-based Multiagent System to Accelerate AI for Social Impact, AAMAS 2025
8. Charlton, J. I. Nothing About Us Without Us: Disability Oppression and Empowerment. University of California Press, 1998 http://www.jstor.org/stable/10.1525/j.ctt1pnqn9

Community Opinion

The AAAI Community Survey on the Future of AI research, with 475 responses, explored various aspects of AI, particularly its application to social good. A significant portion of respondents (119) engaged with questions related to AI for social good.

Regarding the relevance of AI for Social Good, a majority of the 119 respondents who answered this question found it relevant or very relevant to their research. Specifically, 33.61% considered it very relevant, 26.89% relevant, and 21.01% somewhat relevant. When asked about barriers to integrating AI into social impact projects, respondents cited several challenges. The most significant barrier, with 47.06% of respondents selecting it, was "Testing the solution in the field." Other substantial barriers included "Scaling up the solution" (38.66%), "Connecting to non-profit or government organizations" (32.77%), "Problem definition" (36.13%), and "Assessing AI readiness" (30.25%). "Sustainability of the business model" was also a concern for 42.86% of respondents.

The survey also explored crucial resources for scaling AI-driven solutions for social impact. "Money" and "Data" were identified as extremely important by a large majority of respondents (the exact percentages are cut off in the provided snippets). Other important resources included "Government-companies partnerships," "Government-university partnerships," and "Technical support." Finally, regarding measuring the success of AI interventions in addressing social challenges, "Improvement outcome" was the most frequently cited metric, with 47.90% of the 119 respondents choosing it. "Sustainable results" were also considered important by 37.82%, and "Adoption rate" by 48.74%.

AI & Sustainability

AI is rapidly transforming industries and holds immense potential to drive sustainability progress, ranging from accelerating the net-zero energy transition to enhancing climate resilience. However, its deployment also raises challenges, such as increasing energy and water demands. Ensuring AI advances sustainability rather than exacerbating environmental risks will require proactive efforts to shape its development, operations, and applications.

CHAIRS

Eric Horvitz,
Microsoft

Hiroaki Kitano,
Sony Research

Main Takeaways

- While AI compute currently represents a very small share of global energy and water consumption, its rapid growth in certain regions is straining local electricity grids and water resources. Managing these impacts requires investments in local grid capacity and innovations that enhance hardware and software efficiency.
- While concerns about AI's potential environmental impact are rising, researchers and practitioners emphasize that AI's most significant sustainability impacts—both positive and negative—are likely due to how AI is deployed and used rather than from the energy consumed in training and running models.
- AI can be a powerful enabler of climate and sustainability goals. Beyond improving efficiency and reducing carbon emissions across industries, AI is accelerating breakthroughs in areas such as advanced battery materials, carbon removal technologies, and high-precision climate modeling.

Context & History

AI technologies have been advancing for decades, but recent developments in large language models and their widespread adoption are driving the increased use of computationally intensive AI tools across many sectors. As the computational intensity and adoption of AI technologies grow, so do concerns about its environmental footprint, particularly with regard to energy and water consumption.

At the same time, AI is emerging as a tool for sustainability with a promise of being transformational. Achieving ambitious climate and environmental goals—such as electrifying economies, tripling renewable energy capacity, decarbonizing industries, and increasing sustainable food production by 50%—requires system-wide transformations. AI can support these transformations by improving environmental monitoring, optimizing energy systems, enhancing efficiencies across industries, and accelerating materials discovery. For example, advances in material sciences have led to the design of catalysts that reduce the cost of carbon capture and decrease greenhouse gas emissions from industrial processes like concrete production.

Recognizing these opportunities and challenges, international initiatives are aligning AI development with sustainability priorities. For example, the International Energy Agency (IEA) has launched an initiative called "Energy for AI and AI for Energy," which explores how AI can both drive energy innovation and manage its own resource requirements. In spring 2025, the IEA will publish a special report on AI and Energy and launch an AI Observatory to track AI's electricity consumption and its applications in the energy sector. A new initiative was also launched in 2025 to establish energy scores for different AI models (AI Energy Score). Meanwhile, the Coalition for Sustainable AI, established by France in collaboration with the United Nations Environment Programme (UNEP) and the International Telecommunication Union (ITU), is developing guidelines for minimizing AI's environmental impact and promoting best practices across industries.

Current State & Trends

Trend: Rising Resource Demands of AI Compute. The rapid expansion of generative AI is significantly increasing energy and water demands in data centers, driven by both model training and inference workloads. For example, training GPT-3 (175 billion parameters) consumed an estimated 1,287 MWh of electricity and emitted 552 metric tons of CO2 has been reported to have consumed 1287 MWh of electricity with the emission of 552 metric tons of CO2 [1]. While training large models is highly energy-intensive, the greater long-term energy demand is likely to come from inference workloads—that is, running these trained models in real-world applications. Over a model's lifetime, inference can account for a far greater cumulative energy footprint than training.

Data centers—the backbone of AI infrastructure—accounted for approximately 2% of global electricity demand in 2023 [2], and less than 1% of global greenhouse gas emissions [3]. While AI workloads currently represent only a small fraction of data center electricity consumption, this share is expected to grow. In 2022, AI workloads accounted for roughly 1% of total data center electricity use; by 2026, this figure is projected to rise to 9% [3].

Projections suggest that global data center electricity demand could double by 2030, though the extent of this growth will depend on market trends, algorithmic improvements, and hardware efficiency gains [4]. Even under high-growth scenarios, the IEA estimates that AI-related electricity demand will remain a relatively small portion of global energy consumption [4].

However, regional disparities are emerging. In some high-density AI hubs, data center energy consumption is rising rapidly. For example, in the European Union, electricity demand for data centers is growing by approximately 9% per year, with AI's rising computational needs potentially pushing this figure above 5% of total EU electricity demand by 2026 [3]. In the U.S., the world's largest data center market, AI-driven growth has increased data center electricity use to over 4% of national consumption in 2023, more than doubling since 2018. Projections suggest that by 2028, data centers could account for between 7% and 12% of U.S. electricity demand, depending on AI growth scenarios [5].

Trend: Energy-Efficient AI and Renewable-Powered Infrastructure. As AI adoption expands, several strategies are emerging to improve sustainability, including:

- **Advances in hardware efficiency:** GPUs, widely used for AI workloads, consume more energy than traditional CPUs. However, while absolute power consumption per GPU is rising, efficiency per unit of computation is also improving. [6,7]

Optimizing hardware allocation—reserving high-power GPUs for intensive tasks and using low-power CPUs for lighter workloads—can help reduce overall demand.

- **Small Language Models (SLMs):** While large language models (LLMs) require extensive computation, smaller models optimized for specific tasks are emerging as an energy-efficient alternative. SLMs can execute on devices like laptops and smartphones, lowering computational intensity while maintaining performance for targeted applications [8].
- **Cooling innovations:** Traditional air cooling in data centers is inefficient; switching to liquid cooling can significantly reduce energy consumption. While some liquid cooling systems require water, advances in water-free cooling offer solutions that minimize both energy and water use.
- **Optimizing data storage:** AI workloads require massive data storage, increasing electricity demand in data centers. Techniques such as data compression, infrastructure optimization, and edge computing can lower these energy costs.
- **Demand Response and Load Shifting** are strategies that help balance power grids by adjusting electricity consumption in response to grid conditions. Demand Response involves incentivizing customers to modify their energy use, either by reducing demand during peak times, shifting it to periods of greater supply, or utilizing on-site generation and storage. Load shifting specifically focuses on rescheduling energy consumption to align with lower-cost or lower-carbon electricity availability. Increasingly, both approaches are being used to reduce carbon emissions by shifting loads from high-carbon-intensity periods to times when cleaner energy sources are more abundant.

Trend: AI Applications for Sustainability. Artificial intelligence is emerging as a transformative tool for sustainability, offering three key capabilities that can accelerate climate action and environmental protection. AI technologies play a central role in the area of *computational sustainability* [9], centering on the goal of leveraging mathematics, computer science, and information science on sustainability and broader opportunities for enhancing the well-being of humanity. AI promises to play a transformative role in sustainability, offering capabilities that enhance efficiency, optimize resources, and accelerate technological breakthroughs.

AI methods can enhance people's ability to predict and optimize systems, improving efficiency in energy grids, water management, and industrial operations while reducing waste and emissions. Advances in AI modeling are being employed in projects that demonstrate how AI technologies for pattern recognition, prediction, and optimization can be applied to address multiple sustainability challenges, from conservation and protection of wildlife to transportation efficiencies, to breakthroughs in chemistry and materials science that can accelerate the discovery of materials that can facilitate the breakthroughs in battery technology, carbon capture solutions, and low-carbon industrial materials.

In water and climate resilience, AI has the potential to revolutionize hydrological forecasting, irrigation systems, and disaster preparedness [10,11]. AI-powered leak detection reduces water loss [12]. In climate risk management, AI and machine learning is being used to downscale climate models for localized flood and heatwave prediction, allowing governments to better prepare for extreme weather events [13]. AI-assisted wildlife monitoring now enables 99.3% accuracy in species identification, dramatically improving conservation efforts [14].

A promising application of AI is energy optimization. Smart grid systems use AI for demand forecasting, load balancing, and renewable energy integration, helping to optimize energy distribution, reduce waste, and lower emission [15, 16]. AI-powered predictive maintenance in electricity grids helps utilities minimize failures and operational disruptions [15].

A major shift is also underway in materials science, where AI is accelerating the discovery of low-carbon materials at unprecedented speeds. Traditionally, developing new materials for batteries, carbon capture, and sustainable construction could take years or even decades. Today, AI models can scan millions of material combinations in days or weeks [17]. A collaboration between Microsoft and Pacific Northwest National Laboratory identified a new solid-state battery electrolyte in just nine months—a process that would have taken years through traditional experimentation [18].

AI can help to educate and empower the sustainability workforce, equipping

scientists, policymakers, and engineers with tools to enhance decision-making and scale sustainable practices.

Research Challenges

The potential for AI to accelerate sustainability is clear. However, we have no guarantees that AI technologies will serve as a net positive force for sustainability. While AI has the potential to accelerate sustainability progress, its energy and resource demands must be carefully managed. Ensuring AI accelerates sustainability progress will require focused research and innovation across a range of topics, including strategic investments in:

- Energy-efficient AI systems that minimize computational and water resources..
- Innovative AI applications that drive sustainability breakthroughs.
- Robust scenario modeling and data collection efforts to inform policy and guide sustainable AI development.

With proactive governance, targeted research, and cross-sector collaboration, AI can be genuinely positioned not as a sustainability risk, but as a powerful force for climate progress.

Addressing data gaps and large uncertainties

While concerns about AI's potential environmental impact are rising, researchers and practitioners emphasize that AI's most significant sustainability impacts—both positive and negative—stem from how AI is deployed and used, rather than just the energy consumed in developing and running models [19]. AI applications can lead to indirect emissions effects—both positive and negative [20].

However, major uncertainties remain. It is difficult to predict how AI technologies will evolve or how their widespread adoption will impact sustainability. Assessing the net effects of AI on sustainability is challenging due to two main issues: (1) limited availability of reliable data and (2) the difficulty of measuring the actual impact of AI-driven interventions. There is a significant opportunity to develop more comprehensive datasets to better understand AI's sustainability footprint.

Many AI-driven solutions depend on high-quality environmental and industrial datasets, but these are often incomplete, proprietary, or heavily skewed toward high-income countries. Moreover, critical sustainability challenges—such as water scarcity and biodiversity loss—are hindered by major data gaps [21, 22]. AI models trained on limited or biased datasets may fail to account for regional environmental variations, leading to inaccurate predictions or inequitable sustainability solutions. Investing in data collection and standardization can help address these gaps.

Limited data availability of AI's use of energy and water usage presents challenges. Few companies disclose detailed information about the energy consumption, carbon footprint, or water use of their AI workloads. The absence of standardized reporting frameworks makes it difficult for policymakers, researchers, and the public to assess the true sustainability impact of AI. Emerging regulations, such as the EU AI Act, may fill this gap.

Research on modeling and scenarios. AI has the potential to enhance efficiency in sectors like transportation, agriculture, and manufacturing while accelerating conservation efforts. However, predicting the long-term sustainability impact of AI adoption remains a challenge. Scientists have called for the development of modeling frameworks that assess both the direct resource consumption of AI and its broader environmental implications under multiple future scenarios [20]. As an example of uncertainties, consider Jevons Paradox, where efficiency improvements lead to increased overall consumption. As AI hardware and software become more efficient, the cost of computation declines, making AI more accessible and widely adopted. Paradoxically, this increased accessibility can drive up overall energy and resource consumption, offsetting efficiency gains. While individual AI computations are becoming less energy-intensive, the exponential growth of AI workloads means that total demand continues to rise, potentially offsetting many of the gains from efficiency improvements [23].

To navigate these complexities, policy-relevant scenario analyses are essential. These analyses should assess AI's full environmental footprint—including direct energy and water use as well as systemic effects across industries like healthcare, manufacturing, agriculture, and transportation. AI-driven transformations could either accelerate decarbonization or intensify environmental pressures, but current research remains fragmented. Investing in scenario modeling can inform policy and guide strategic investments while communicating key uncertainties to policymakers and scientists.

Scenario modeling—widely used in finance and climate risk assessment—can help quantify these uncertainties by exploring different AI adoption pathways, ranging from minimal integration to widespread deployment aligned with global sustainability goals. Researchers should develop forecasting frameworks that evaluate various possible futures, from best-case scenarios where AI enables deep emissions reductions to worst-case scenarios where unchecked expansion increases environmental strain. These insights are critical for steering AI innovation toward sustainability while mitigating unintended risks.

Designing resource-efficient AI systems. There is a significant opportunity to design AI models and infrastructure to be more energy- and resource-efficient. Key strategies include:

- Optimizing AI model architectures to improve computational efficiency without sacrificing performance.
- Developing specialized AI hardware that consumes less energy and water than traditional GPUs.
- Enhancing AI infrastructure management to enable carbon-aware computing, where AI workloads are scheduled based on grid conditions to minimize carbon-intensive energy consumption.

Expanding AI-Enabled Solutions for Critical Sustainability Opportunities. The most transformative sustainability benefits of AI are likely to come from new, targeted applications that address critical environmental challenges. There is a major opportunity to strategically apply AI to complex sustainability problems, from introducing new efficiencies in transportation systems and industrial processes to advances in chemistry, material science, and the biosciences.

For example, AI shows promise with revolutionizing materials discovery by accelerating the identification of new battery storage materials (e.g., see [24]), carbon capture solutions, and low-carbon industrial materials. Other high-impact opportunities include:

- Developing cost-effective, long-term energy storage solutions to enable greater reliance on intermittent renewables like wind and solar.
- Achieving large-scale carbon dioxide removal at less than $100 per ton.
- Expanding electricity transmission capacity and reliability to integrate more renewable energy sources.
- Reducing water and gas leaks at a global scale through AI-powered monitoring.
- Filling critical biodiversity data gaps and optimizing conservation programs with AI-driven insights.
- Introducing new efficiencies into transportation systems (see, e.g., [25]).

Addressing these challenges requires close collaboration between AI researchers and domain experts, development of new AI methods and applications, and investments in efforts to compile and integrate relevant datasets for analysis, modeling, and machine learning.

1. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.
2. International Energy Agency (2024). Electricity 2024: Analysis and forecast to 2026. https://www.iea.org/reports/electricity-2024
3. International Energy Agency (2024). Data Centres and Data Transmission Networks. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks
4. International Energy Agency (2024). World Energy Outlook 2024. https://www.iea.org/reports/world-energy-outlook-2024
5. Shehabi, A., Smith, S.J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J., Masanet, E., Sartor, D. (2024). 2024 United States Data Center Energy Usage Report (LBNL-2001637), Lawrence Berkeley National Laboratory, Berkeley, California. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf
6. Nvidia, Product specification sheets for DGX A100, H100 (2024). https://resources.nvidia.com/en-us-dgx-systems/ai-enterprise-dgx
7. Smith, M. S. (2024), Challengers are Coming for Nvidia's Crown: In AI's Game of Thrones, Don't Count Out the Upstarts. IEEE Spectrum, vol. 61, no. 10, pp. 40–44, Oct. 2024, doi: 10.1109/MSPEC.2024.10705376 https://ieeexplore.ieee.org/document/10705376
8. UNESCO (2024, March). Small Language Models (SLMs): A Cheaper, Greener Route into AI https://www.unesco.org/en/articles/small-language-models-slms-cheaper-greener-route-ai#
9. Gomes, C., Dietterich, T., Barrett, C., et al. (2019). Computational sustainability: Computing for a better world and a sustainable future, Communications of the ACM. 62 (9): 56–65. https://dl.acm.org/doi/pdf/10.1145/3339399
10. [Flecker, A.S., et al., (2022) Reducing adverse impacts of Amazon hydropower expansion. Science 375, 753–760. DOI:10.1126/science.abj4017 https://www.science.org/doi/10.1126/science.abj4017
11. Rolnick, D. et al. (2022). Tackling Climate Change with Machine Learning. ACM Comput. Surv. 55, 2, Article 42 (February 2023), 96 pages. https://dl.acm.org/doi/10.1145/3485128
12. Bolgar, C. (2024, September). AI tool uses sound to pinpoint leaky pipes, saving precious drinking water. Source: Microsoft News. https://news.microsoft.com/source/features/sustainability/ai-tool-uses-sound-to-pinpoint-leaky-pipes-saving-precious-drinking-water/
13. Yoshikane, T., & Yoshimura, K. (2023). A downscaling and bias correction method for climate model ensemble simulations of local-scale hourly precipitation. Scientific Reports, 13(1), 9412.
14. Norouzzadeh, M.S., et al., (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning, Proc. Natl. Acad. Sci. U.S.A. 115 (25) E5716-E5725, https://doi.org/10.1073/pnas.1719367115 (2018). https://www.pnas.org/doi/10.1073/pnas.1719367115
15. Benes, K. J., Porterfield, J. E., & Yang, C. (2024). AI for energy: Opportunities for a modern grid and clean energy economy. US Department of Energy.
16. Sandalow, D., McCormick, C., Kucukelbir, A., et al. (2024). Artificial Intelligence for Climate Change Mitigation Roadmap (Second Edition) (ICEF Innovation Roadmap Project, November 2024) https://doi.org/10.7916/2j4p-nw61.
17. Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. Nature, 624(7990), 80–85.
18. Accelerating materials discovery with AI and Azure Quantum Elements (2024). Microsoft Azure Quantum Blog https://azure.microsoft.com/en-us/blog/quantum/2023/08/09/accelerating-materials-discovery-with-ai-and-azure-quantum-elements
19. Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. Nature Climate Change, 12(6), 518–527.
20. Luers, A., Koomey, J., Masanet, E., Gaffney, O., Creutzig, F., Lavista Ferres, J., & Horvitz, E. (2024). Will AI accelerate or delay the race to net-zero emissions?. Nature, 628(8009), 718–720. https://www.nature.com/articles/d41586-024-01137-x
21. Leung, B., & Gonzalez, A. (2024). Global monitoring for biodiversity: uncertainty, risk, and power analyses to support trend change detection. Science Advances, 10(7), eadj1448.
22. Rosa, L., & Sangiorgio, M. (2025). Global water gaps under future warming levels. Nature Communications, 16(1), 1192.
23. Luccioni, A. S., Strubell, E., & Crawford, K. (2025). From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate. arXiv preprint arXiv:2501.16548.
24. Chen, C., et al. (2024) Accelerating Computational Materials Discovery with Machine Learning and Cloud High-Performance Computing: from Large-Scale Screening to Experimental Validation, Journal of the American Chemical Society 2024 146 (29), 20009–20018DOI: 10.1021/jacs.4c03849. https://pubs.acs.org/doi/abs/10.1021/jacs.4c03849
25. Kamar, E. and Horvitz, E. (2009). Collaboration and shared plans in the open world: Studies of ridesharing. In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09). Morgan Kaufmann Publishers Inc., San Francisco, CA, 187–194. https://web.archive.org/web/20220119103915id_/https://www.ijcai.org/Proceedings/09/Papers/041.pdf

Community Opinion

A recent survey of AI community members revealed a divided perspective on AI's environmental impact:

- Approximately 35% of respondents agreed or strongly agreed that AI's environmental harms are outweighed by its potential to address climate challenges. Conversely, another 35% disagreed or strongly disagreed with this statement.
- Over 70% of respondents believe that data-intensive AI significantly impacts global resource consumption.
- 57% of respondents expressed concerns that AI's energy consumption could slow the pace of AI research.
- Nearly 75% of those surveyed called out energy efficient training and inference procedures as being most needed to reduce AI energy consumption, followed by 20% calling out innovation in energy systems for data centers, and 5% for innovating with energy-efficient chips.

When asked where AI could have the greatest impact on sustainability:

- Over 30% identified logistics, transportation, and infrastructure optimization as the top opportunity, and approximately 10% each cited AI's role in CO_2 reduction, agriculture, disaster prediction, and advancing the circular economy. Only 5% saw AI's greatest sustainability potential in biodiversity conservation.

These insights underscore the AI community's recognition of both the opportunities and risks associated with AI's sustainability trajectory.

AI for Scientific Discovery

Artificial Intelligence (AI) is revolutionizing scientific discovery by accelerating the entire research cycle from knowledge extraction and hypothesis generation to automation of experimentation and verification at an unprecedented speed.

CHAIR
Hiroaki Kitano,
Sony Research

Main Takeaways

- Progress of AI for Scientific Discovery will accelerate the pace of scientific discovery in unprecedented ways and transform the way we do science.
- Highly automated AI systems for scientific discovery are emerging: While their capabilities are limited, they can execute the entire scientific discovery cycle with minimal human intervention for defined tasks.
- AI's role in scientific discovery raises new challenges in ethics, collaboration, and reliability, requiring interdisciplinary efforts to address them.

Context & History

Scientific discovery has historically been driven by human ingenuity, but Artificial Intelligence (AI) is emerging as a pivotal tool in reshaping this process and accelerating it.

Early AI systems, such as DENDRAL, designed in the 1960s, were among the first to automate hypothesis generation and problem-solving in organic chemistry [1]. Similarly, systems like EURISKO explored heuristic learning and adaptation, demonstrating the potential of AI in creative problem-solving [2].

These pioneering efforts laid the groundwork for AI's role in science by showcasing its ability to process datasets and formulate hypotheses.

With the progress of technology for comprehensive and high-precision measurements, scientists are overwhelmed by the massive data generated and the complexity of systems behind it. Technologies to discover patterns from massive data that can be linked to novel hypotheses to be tested by experiments are essential for scientific research. AI for Scientific Discovery is a much-needed research area and is expected to drive practice of science in a data-driven approach.

There is a spectrum of approaches in how AI transforms scientific discovery. First, we envision increasingly powerful AI tools that will assist human scientists to make discoveries faster and enable them to tackle even more challenging problems. The alternative approach is to develop integrated AI and robotics systems aiming at performing the entire cycle of scientific research with minimum interventions from human scientists. The middle ground is to position AI as a collaborator with scientists that can interactively solve scientific problems. AI for Scientific Discovery embraces a wider spectrum of relationship between AI and scientists, and rapid progress is expected in all ranges of this spectrum that will fundamentally transform the way we do science. A series of papers, reports and workshops concluded that AI for Science is one of the most important research areas in coming years[3–6].

Current State & Trends

1. From Tools to AI Collaborators and Autonomous AI Scientists

AI's capability in supporting scientific discovery has expanded dramatically, with systems like AlphaFold2 from DeepMind achieving groundbreaking success in structural biology[7,8]. AlphaFold2 solved the decades-old problem of protein folding, revolutionizing molecular biology and enabling applications in drug discovery and biomedicine. AlphaFold2 represents the most successful case of AI as a tool for scientific discovery that transforms biomedicine and biochemistry research and resulted in a 2024 Nobel Prize in Chemistry. An increasing number of AI/robotics systems have been developed to be effective tools for scientists. Beyond biology, numerous AI tools have been developed for chemistry[9–11], material science[12], mathematics[13–15] and many other scientific fields to accelerate scientific discovery. For example, the Ramanujan Machine is an early effort to generate conjectures on fundamental constants that can be a tool for mathematicians undertaking a specific task in mathematics[13]. AI and robotics integrated systems have been developed that perform a specific type of chemistry experiment automatically[16,17].

The other end of the spectrum is to develop highly autonomous AI/robotics systems to perform scientific discovery without (or with minimum) human interventions. The Adam and Eve systems, developed by Ross King, represent the early effort on the other side of the spectrum where scientific discovery occurs without human-in-the-loop[18,19]. These 'robot scientists' not only generate hypotheses but also design and execute experiments to test them. For example, Eve identified potential drugs for malaria through automated high-throughput screening, showcasing AI's ability to iterate scientific cycles autonomously[20]. The Nobel Turing Challenge represents the extreme end of the challenge gearing toward highly autonomous AI and robotics systems with the capability for high impact scientific discovery[21,22]. Proposed as a grand vision, it aims to develop AI systems capable of scientific discoveries on par with Nobel Prize-winning work. This challenge entails an instance of the Feigenbaum test - Can I replicate the best human expert in defined domains? - that is a variation of the Turing test[23]. These systems would not merely assist researchers but act as autonomous entities capable of proposing, testing, and refining theories.

2. Impact on Science

AI's integration into scientific workflows heralds a paradigm shift:

- **Accelerated Discovery:** By automating an entire cycle of

scientific discovery, AI can reduce the time required to achieve breakthroughs, enabling a rapid expansion of knowledge.

- **Enhanced Collaboration:** AI systems like AlphaFold demonstrate how interdisciplinary approaches—combining AI, biology, and physics—can tackle long standing challenges. Eventually, a network of collaboration may be formed among AI systems enabling extensive exchange of ideas and data at the scale not possible by human scientists.3. Exploration
- **Beyond Human Intuition:** AI's ability to exhaustively explore hypothesis spaces enables discoveries that might elude human researchers constrained by cognitive and methodological biases.
- **Transformation in Data Handling:** An AI-centric approach dramatically changes the way researchers handle experimental data. In the AI-centric approach all data is important, not just the subset that strongly supports the expected outcome, but also data that are not consistent with the expectation because most, if not all, data need to be provided to AI system training for better hypothesis generation and prediction.

3. Social and Ethical Implications

AI-driven science will likely have profound societal impacts:

- **Healthcare Transformation:** With AI accelerating drug discovery and personalized medicine, millions of lives could be saved or improved.
- **Environmental and Climate Change:** Rapid progress of material science, chemistry, and environmental sciences may provide us with discoveries that can be used to mitigate climate change and improve the state of the environment.
- **Ethical Considerations**: The autonomy of AI systems raises questions about accountability, credit for discoveries, and the potential displacement of human researchers. There are concerns that hazardous materials may be designed and produced with highly autonomous AI scientists. Ethical and safety measures shall be taken to prevent malevolent uses of powerful synthesis abilities with materials and biology[24]. Proper measures must be taken to prevent such uses.

Research Challenges

1. Communication issues: A major challenge may be to be able to understand and communicate with human scientists because accumulation of knowledge and communications with peers mostly takes the form of natural languages with ample ambiguities, analogies, and often with cultural context. Of particular importance:

- **Commonsense knowledge:** Professional knowledge is grounded in everyday knowledge. Construction of a shared broad conceptual knowledge base of the world, to ground reasoning and models and provide grist for analogies.
- **Collaboration:** Science is often a team sport, being able to work effectively with others (human or AI) will be important.
- **Communication**: Scientists talk, draw, and use multiple forms of interactive media. AIs for science will need to be able to read and understand the scientific literature and communicate well with human partners.
- **Models of scientific reasoning and approaches** to developing modeling and inference machinery that can augment human cognition for scientific discovery[25].

2. Defining Hypothesis Space: Science is an open-ended problem: Unlike most games, such as chess or the game of GO, the structure and scale of the problem space is not obvious and likely to be unbounded and could be with very high dimensionality. Extracting or acquiring knowledge and properly placing it within the space of scientific knowledge is a non-trivial problem. Similarly, the process of hypothesis generation faces the problem of identifying the dimensionality and scale of hypothesis space it should work on.

3. Inaccuracies, noise, and reproducibility of data: Data in science can be very noisy, inaccurate and not reproducible in some fields. In biology, inaccuracy and noise in data is considered to be inevitable due to artifacts introduced in experiments and with intrinsic variability of experimental samples. Studies revealed that a significant proportion of data reported in publications cannot be reproduced properly in biomedical research[26, 27]. This may pose problems for the quality of hypotheses generated or their verification process at the early stage of the research.

1. Lindsay R, Buchanan B, Feigenbaum E, Lederberg J. DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation. Artif Intell. 1993;61: 209–261.
2. Lenat D, Brown J. Why AM and EURISKO appear to work. Artif Intell. 1984;23: 269–294.
3. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature. 2023;620: 47–60.
4. Science and Engineering Capacity Development Unit, Computer Science and Telecommunications Board, Policy and Global Affairs, Division on Engineering and Physical Sciences, National Academies of Sciences, Engineering, and Medicine. AI for scientific discovery: Proceedings of a workshop. Pool R, editor. Washington, D.C.: National Academies Press; 2024. doi:10.17226/27457
5. Gil Y, Greaves M, Hendler J, Hirsh H. Artificial Intelligence. Amplify scientific discovery with artificial intelligence. Science. 2014;346: 171–172.
6. President's Council of Advisors on Science and Technology. PCAST: Report to the president on supercharging research: Harnessing artificial intelligence to meet global challenges. Office of Scientific and Technical Information (OSTI); 2024 Apr. doi:10.2172/2481685
7. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577: 706–710.
8. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.
9. Peplow M. ChatGPT for chemistry: AI and robots join forces to build new materials. Nature. 2023 [cited 30 Nov 2024]. doi:10.1038/d41586-023-03745-5
10. Dias AL, Rodrigues T. Large language models direct automated chemistry laboratory. Nature. 2023;624: 530–531.
11. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624: 570–578.
12. Manica M, Born J, Cadow J, Christofidellis D, Dave A, Clarke D, et al. Accelerating material design with the generative toolkit for scientific discovery. Npj Comput Mater. 2023;9: 1–6
13. Raayoni G, Gottlieb S, Manor Y, Pisha G, Harris Y, Mendlovic U, et al. Generating conjectures on fundamental constants with the Ramanujan Machine. Nature. 2021;590: 67–73.
14. Davies A, Veličković P, Buesing L, Blackwell S, Zheng D, Tomašev N, et al. Advancing mathematics by guiding human intuition with AI. Nature. 2021;600: 70–74.
15. He Y-H. AI-driven research in pure mathematics and theoretical physics. Nature Reviews Physics. 2024; 1–8.
16. Coley CW, ThomasIII DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science. 2019;365. doi:10.1126/science.aax1566
17. Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature. 2020;583: 237–241.
18. King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, et al. The Automation of Science. Science. 2009;324: 85–89.
19. King RD, Whelan KE, Jones FM, Reiser PG, Bryant CH, Muggleton SH, et al. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature. 2004;427: 247–252.
20. Williams K, Bilsland E, Sparkes A, Aubrey W, Young M, Soldatova LN, et al. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. J R Soc Interface. 2015;12: 20141289.
21. Kitano H. Nobel Turing Challenge: creating the engine for scientific discovery. npj Systems Biology and Applications. 2021;7: 29.
22. Kitano - 2016 - Artificial Intelligence to Win the Nobel Prize and Beyond Creating the Engine for Scientific Discovery, AI Magazine, 37:1 2016.
23. Feigenbaum, Edward A. (2003). "Some challenges and grand challenges for computational intelligence". Journal of the ACM. 50 (1): 32–40. doi:10.1145/602382.602400. S2CID 15379263
24. Wittmann BJ, Alexanian T, Bartling C, Beal J, Clore A, Diggans J, et al. Toward AI-resilient screening of nucleic acid synthesis orders: Process, results, and recommendations. bioRxiv. 2024. p. 2024.12.02.626439. doi:10.1101/2024.12.02.626439
25. Hope, T, Downey, D., Weld, D.S., Etzioni, O. and Horvitz, E. (2023). A Computational Inflection for Scientific Discovery. Communications of the ACM 66, 8 (August 2023), 62–73. https://doi.org/10.1145/3576896
26. Prinz F, Schlange T, Asadullah K (August 2011). "Believe it or not: how much can we rely on published data on potential drug targets?". Nature Reviews. Drug Discovery. 10 (9): 712. doi:10.1038/nrd3439-c1. PMID 21892149.
27. Begley CG, Ellis LM (March 2012). "Drug development: Raise standards for preclinical cancer research". Nature (Comment article). 483 (7391): 531–533. Bibcode:2012Natur.483..531B. doi:10.1038/483531a

Community Opinion

In the community opinion survey, 32% of those responding consider this is somewhat relevant. The area of study that AI to be most useful is ranked as (1) Biology (47%), (2) Physics (14%), (3) Chemistry (12%), with 26% responding for other domains. In the question of if an AI system can ever make discovery worthy of the Nobel Prize, only 13% of those responding said never, with 25% saying no idea. 11% thought it might happen in the 2020s, and 45% thought it might happen by the 2050s.

Artificial General Intelligence (AGI)

CHAIRS

Eric Horvitz,
Microsoft

Stuart Russell,
University of California
Berkeley

Although the field of AI has long pursued the kinds of general-purpose, human-level abilities captured by the term AGI, the rise of more general capabilities of neural net models has stimulated discussions about directions forward, implications around success, and doubts about pursuing the goal–which now appears to some observers to be within reach.

Main Takeaways

- Pursuing understandings of principles and machinery of intelligence that could be harnessed to reach human-level capabilities have always been central in AI, and was explicitly called out in 1956 as an important goal by founders of the discipline.
- Calls for focusing more centrally on the bigger picture of "human-level AI" and "artificial general intelligence" in the early 2000s arose in the context of the successful fielding of narrowly scoped AI applications and what some perceived as a lack of progress on the more visionary goals of the field.
- Despite challenges with precise definitions and debate about the value of particular notions of AGI, the aspirational goals of AGI and closely related notions, such as "human-level AI," have inspired many fundamental advances in AI and frame key research questions moving forward to more capable AI systems. On the other hand, success in creating AGI could create societal disruptions and risks and pose significant safety challenges, including challenges to human flourishing and survival.

Context & History

The AI field has long pursued general principles of intelligence with the direct implication that breakthroughs in our computational understanding of intelligence would enable general-purpose capabilities. The Turing Test exemplifies this: to pass, a machine must match or exceed human knowledge and reasoning abilities across a range of domains in which people are expected to be competent.

The proposal for the Dartmouth workshop that initiated the AI field under that name, written in 1955, begins, "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves." This extraordinarily ambitious agenda set the tone for much subsequent work by the participants, including McCarthy's "Programs with Common Sense", Newell and Simon's General Problem Solver, and Solomonoff's Universal Induction. Just two years later, in 1957, Herb Simon predicted that "the range of problems [machines] can handle will be coextensive with the range to which the human mind has been applied." Thus, AI has always had as its goal the creation of machines with general powers of intelligence. A great deal of AI research has continued in the vein of pursuing general principles of intelligence, including efforts in representation, sensing, and logical and probabilistic inference.

Over the decades, the vast majority of researchers focused on specific methodologies and components of intelligence without much consideration for their integration into general-purpose systems. While some researchers in specific areas were passionate about the potential for generalizing their advances, real-world demonstrations were largely disappointing. Applications harnessing the frontiers in AI methods were narrow and brittle. The perceived lack of progress towards generally intelligent systems that could function in the real world led some within the field to complain that the big picture and high ambitions of AI were being forgotten. For example, Nils Nilsson's 1995 paper "Eye on the Prize" [1] stated, "AI is now at the beginning of another transition, one that will reinvigorate efforts to build programs of general, humanlike competence," but this was more of an exhortation than a statement of fact. The prospect of pursuing "human-level" intelligence came to the fore again in the early 2000s. For example, in 2002, Marvin Minsky organized a workshop on "Designing Architectures for Human-Level Intelligence."

The term artificial general intelligence (AGI) emerged in the same time period as an expression of high ambition by a younger generation of researchers who criticized the field's seeming focus on narrow applications. Indeed, it was in the early 2000s when machine learning started to be harnessed in multiple narrow applications, each celebrated as a valuable advance. The narrowness of these applications led to calls to discover more generalizable and powerful methodologies, motivated by the fact that the principles of machine learning and reasoning can be applied across domains.

AGI was initially defined as AI that could match or exceed human cognitive abilities across a broad range of tasks, echoing the original ambitions of the field in 1956. While these goals were not new to senior AI researchers, the use of the term AGI was seen by many—both inside and outside the field—as a refreshing call for ambitious projects.

Beyond AGI and human-level AI, other terms that gained traction around the same time include general-purpose AI and strong AI. However, AGI has become the dominant term in both research and public discourse. Popular books and articles frame AGI as a novel ambition, often portraying it as an unprecedented goal, despite its deep roots in the early history of AI. In many discussions, including those outside of AI research, AGI was linked to both utopian and dystopian futures, reflecting varying perspectives, expectations, and anxieties.

The previous AAAI Presidential Panel, the Presidential Panel on Long-Term AI Futures [2], was established in 2008 amid growing interest in AGI, a rekindling of high ambitions by AI leaders and growing public discourse, as well as an upswing in applications of AI being fielded in the open world. The set of meetings and final convening at Asilomar focused on key questions about feasibility, implications, ethics, and safety, as well as research directions for building powerful, general, human-level intelligences.

Different perspectives about the nature of AGI extend beyond the core definition of AI methods that could "match or exceed human cognitive abilities across a broad range of tasks." For example, discussions of AGI, particularly in the popular press, have

fueled speculation that sentience or consciousness could be a characteristic of AGI systems. AI researchers generally steer clear of such speculations, pointing out that the analysis and prediction of behavior is independent of attributions of sentience.

Some researchers have also suggested that AGI systems must, by definition, have "agentic" abilities, meaning that, like humans, they can function as actors that perceive, learn, process, and act upon their environment to achieve specific goals. Indeed, the capacity to act in pursuit of goals is a fundamental cognitive property of humans, and some AI systems have exhibited such capabilities in rudimentary form since the earliest days of AI.

Perhaps more confusing is the notion of "autonomy"and its link to AGI—specifically, the possibility suggested by some that AGI systems might develop goals of their own, entirely distinct from those provided by humans. While this is logically possible—for example, an AI system might overwrite its objectives with new, randomly generated objectives—it's less clear why it might do so, since that would guarantee failure in its current objectives. On the other hand, the formation of so-called "instrumental" subgoals—such as self-preservation and acquiring additional computation and financial resources—seems highly likely as AI systems pursue their original objectives. This is obviously a source of concern and an active area of longstanding research.

The fact that AGI systems would be more generally capable than humans raises obvious concerns about loss of control of AI; indeed, Alan Turing himself stated that "we should have to expect the machines to take control" once they exceeded human levels of intelligence. One source of risk is misalignment, where the AGI's goals are not aligned with human preferences about the future; this could arise from misspecification or underspecification by humans—the so-called "King Midas problem"—or from AGI systems failing to understand human preferences correctly [3].

For some, AGI represents a potentially dangerous "threshold" that we cross at our peril. As an example, the "Gladstone Report" [4] commissioned by the US State Department states that "AGI is generally viewed as the primary driver of catastrophic risk from loss of control." Others use the term "transformative AI" [5] to cover AI systems that have the potential to cause massive disruption of human civilization, noting that this does not require full AGI. We note that sentience and autonomy are not part of core definitions of AGI, even if some have made implicit assumptions about AGI having these attributes.

AGI is not a formally defined concept, nor is there any agreed test for its achievement. Some researchers suggest that "we'll know it when we see it" or that it will emerge naturally from the right set of principles and mechanisms for AI system design. In discussions, AGI may be referred to as reaching a particular threshold on capabilities and generality. However, others argue that this is ill-defined and that intelligence is better characterized as existing within a continuous, multidimensional space. Some (e.g., [6]) contend that the lack of a clear definition makes AGI an unsuitable goal for AI research: human intelligence has many dimensions, and machines will likely far exceed humans in some areas while remaining inferior in others. Moreover, the criteria for comparison, including which particular humans serve as benchmarks and how much prior training they have received, are often left unspecified.

Some argue that AGI is not a desirable goal for AI research, contending that "matching or exceeding human cognitive abilities" does not necessarily lead to tools that enhance or complement human abilities. Instead, they argue, AGI's short-term monetary value would be in replacing humans in most economic roles. Moreover, many of the purported benefits of AGI—in science, healthcare, education, and other fields—can be achieved through more narrowly focused tools, such as AlphaFold2. Nonetheless, AGI has become the canonical goal for ambitious AI companies. For example, Sam Altman, CEO of OpenAI, has stated, "The vision is to make AGI, figure out how to make it safe … and figure out the benefits" [7].

Current State & Trends

The trajectory of AI capabilities and benchmark results over the last decade is very clear and points to achieving human-level or superhuman capabilities on one task after another, as captured in the series of AI Index reports [8] and the 2025 International AI Safety Report [9].
Early successes were initially seen with speech recognition and object recognition from images, followed by advances in machine translation. The rise of new capabilities with generative AI has provided tools for synthesizing high-quality images and voices, and in 2022 the mastery of generating language. Strong competencies were reached in 2023 with multimodal models that span language, imagery, audio (both as input and output), and

physical embodiment. In 2024, we have seen major advances in reasoning, including success on the abstract reasoning challenge (ARC-AGI), on which AI systems had utterly failed before 2024.

Recent advances in capabilities with neural network models are based on the introduction of run-time deliberation mechanisms that learn to employ chains of inner thought, inspired by theories of higher-level cognition in humans. Whereas previous models directly mapped the input context to an answer in constant time, akin to fast intuitive human responses described as "system 1" cognition, the more recent wave of advances in "test-time" reasoning allow the AI to explore lengthy chains of verbal reasoning to find answers to complex questions. These algorithms consider and evaluate multiple possibilities in the style of more deliberative human cognitive processes that have been referred to as "system 2" cognition. These also require much more computation at run-time, which may greatly increase both energy and monetary costs for deploying such systems, beyond just the cost of training.

Along with physical capabilities, reasoning competency was seen as the main remaining gap to human-level intelligence; but the gap seems to be closing. There are now AI systems that score among the top few percent of humans on many commonly used tests of advanced knowledge and reasoning. On the other hand, such systems still evince elementary failures that raise significant questions about how to interpret their successes [10]. For example, multiple leading edge models show remarkable failures at mathematical tasks specified via combinations of word problems and imagery that humans find straightforward [11]. Similarly, most state-of-the-art models still face challenges with spatial and geometric reasoning and detailed image understanding, particularly for multimodal input [12]. And planning as a special form of reasoning is still quite weak, especially when it comes to longer planning horizons and planning actionable steps and assistance for and with humans in the physical world. However, research on this topic is receiving major investments, and could enable human-level competency on more tasks. This would have tremendous economic value but also raises questions about societal impact.

Research Challenges

Despite significant progress in large-scale deep learning models such as transformers, modern AI systems are generally considered to not have achieved all of the capabilities cited in most definitions of AGI. In the context of current key deficits, research opportunities include the following directions:

Architectures Beyond Transformers: The standard transformer architecture has demonstrated remarkable capabilities, but it has fundamental limitations, such as fixed context windows, lack of explicit memory, inability to learn and react to real-time feedback from environments, and inefficiency and challenges in complex reasoning tasks. Research on new architectures could provide pathways to various definitions of human-level intelligence. Directions include boosting reasoning and generalization capabilities via fundamentally new architectures and also exploring hybrid architectures that combine transformers with other models, such as graph neural networks, reinforcement learning agents, or symbolic reasoning systems.

Long-Term Planning and Reasoning: Current AI models struggle with long-horizon planning and fail to demonstrate robust hierarchical reasoning. Unlike humans, they do not exhibit strong foresight, struggle with multi-step problem-solving, and are not naturally inclined to break complex goals into subgoals efficiently and accurately. And unlike classical AI systems from the 1960s onwards, they cannot guarantee the correctness of their reasoning steps. Recent advances in test-time scaling, with the use of reinforcement learning to learn to reason with chains of thought, are one direction of research on endowing neural-network-based systems with abilities to plan more effectively. These advances require important extensions that call for proactively estimating the risk and cost of each step in the plan, and whether each step would still be aligned with human values and problem specifications.

Generalization Beyond Training Data: While LLMs exhibit impressive abilities, their generalization capabilities outside their training distribution and for genuinely novel problems are unclear. They can be easily misled by adversarial challenges and often lack the ability to apply knowledge flexibly across varied domains. It may be necessary to learn representations, such as programs, that are more expressive than circuits, but we lack efficient mechanisms for doing so.

Continual Learning: Unlike humans, LLMs do not learn continuously from experience but rather via a

rigid pretraining and fine-tuning paradigm. Research is needed on mechanisms that allow systems to retain and update knowledge in-stream, over time rather than relying on static, offline training procedures. A shift toward architectures and training methodologies that enable continual, lifelong learning is essential. Opportunities include new forms of self-supervision and self-directed learning via simulation and exploration at the borders of competencies and understandings about the environment or conceptual challenges [13].

Memory and Recall: Reaching AGI may require integration of human-like capture and context-sensitive recall, including some kind of structured, episodic memory. Unlike humans, transformers do not maintain a persistent, structured memory that accumulates relevant aspects of experiences over time efficiently over extended periods. Efforts are underway to supplement LLMs with memory mechanisms, typically via external machinery. There is great opportunity in this direction of research.

Causal and Counterfactual Reasoning: While AI models can detect correlations in vast datasets, they struggle with causal inference and counterfactual reasoning. Understanding cause-and-effect relationships is essential for robust decision-making and scientific discovery. Causal reasoning with large language models is an important research direction.

Embodiment and Real-World Interaction: Human intelligence develops through rich sensorimotor interactions with the world. Current multimodal models seem to lack a deep understanding of physical reality and struggle to sense, reason, and interact effectively in real-world environments. Interesting research directions include training AI models in rich, interactive environments (e.g., robotics, virtual worlds) to build a more grounded understanding of reality that span multiple rich modalities including video, audio, and sensory data.

Alignment, Interpretability, and Safety: Ensuring that AI systems align with human values and are interpretable remains a pressing concern with the pursuit of more capable, human-level intelligence. Black-box AI models, including transformers, often yield outputs that are difficult to explain, which raises safety and trust issues. Moreover, LLMs are trained as imitation learners, learning verbal behavior that is as similar to humans as possible. Because human verbal behavior is purposeful—achieving goals ranging from self-preservation and finding a mate to becoming wealthy and powerful—it is likely that LLMs are in effect acquiring similar or related goals that they may pursue on their own account. Research is needed on alignment of values, specification of constraints and policies that ensure safe operation, and, more generally, developing safety measures to ensure that highly capable AI systems adhere to human intentions.

Understanding and Guiding Societal Influences: As AI systems become more capable, safety research, proactive governance, and active monitoring of the impacts of AI on people and society will grow in importance. Rather than relying on market forces to ensure positive outcomes, the AI research community can engage early on and continue to stay in touch with policy makers and civil society leaders to help to shape the capabilities and uses of AI and its governance [14].

1. N. Nilsson (1995). Eye on the Prize. AI Magazine, 16(2), 9. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1129
2. Association for the Advancement of AI (2009). AAAI Presidential Panel on Long-Term AI Futures, 2008–2009. https://aaai.org/about-aaai/aaai-presidential-panel-on-long-term-ai-futures-2008-2009/ T. Dietterich and E. Horvitz (2015). Rise of Concerns about AI: Reflections and Directions, Communications of the ACM, 58(10). https://dl.acm.org/doi/pdf/10.1145/2770869
3. E. Harris, J. Harris, and M. Beall (2024). An Action Plan to increase the safety and security of advanced AI. Gladstone AI. https://www.gladstone.ai/action-plan
4. H. Karnofsky (2016). Some background on our views regarding advanced artificial intelligence. Open Philanthropy. https://www.openphilanthropy.org/research/some-background-on-our-views-regarding-advanced-artificial-intelligence/
5. J. Togelius (2024). Artificial General Intelligence, MIT Press.
6. M. Murgia (2023). OpenAI chief seeks new Microsoft funds to build 'superintelligence'. Financial Times, November 13.
7. R. and J. Clark, eds. (2017–24). AI Index Reports. Stanford Institute for Human-Centered Artificial Intelligence. https://aiindex.stanford.edu/
8. Y. Bengio and others (2025). International AI Safety Report. His Majesty's Government. International AI Safety Report 2025 - GOV.UK
9. G. Marcus (2025). AGI vs "broad, shallow intelligence." Substack. https://garymarcus.substack.com/p/agi-versus-broad-shallow-intelligence
10. A. Cherian, K. Peng, S. Lohit, J. Matthiesen, K. Smith, J. Tenenbaum (2024). Evaluating Large Vision-and-Language Models on Children's Mathematical Olympiads. NeurIPS 2024. https://arxiv.org/abs/2406.15736
11. V. Balachandran, J. Chen, N. Joshi, B. Nushi, H. Palangi, E. Salinas, V. Vineet, J. Woffinden-Luey, S. Yousefi (2024), Eureka: Evaluating and Understanding Large Foundation Models, arXiv 2409.10566, September 2024. https://arxiv.org/abs/2409.10566
12. N. Lee, Z. Cai, A. Schwarzschild, K. Lee, D. Papailiopoulos (2024). Self-Improving Transformers Overcome Easy-to-Hard and Length Generalization Challenges, arXiv 2502.01612, February 2025. https://arxiv.org/abs/2502.01612
13. E. Horvitz, V. Conitzer, S. McIlraith, and P. Stone (2024). Now, Later, and Lasting: 10 Priorities for AI Research, Policy, and Practice, Communications of the ACM, 67(6). https://cacm.acm.org/opinion/now-later-and-lasting-10-priorities-for-ai-research-policy-and-practice/

Community Opinion

The responses to our survey on questions about AGI indicate that opinions are divided regarding AGI development and governance. The majority (77%) of respondents prioritize designing AI systems with an acceptable risk-benefit profile over the direct pursuit of AGI (23%). However, there remains an ongoing debate about feasibility of achieving AGI and about ethical considerations related to achieving human-level capabilities.

A substantial majority of respondents (82%) believe that systems with AGI should be publicly owned if developed by private entities, reflecting concerns over global risks and ethical responsibilities. However, despite these concerns, most respondents (70%) oppose the proposition that we should halt research aimed at AGI until full safety and control mechanisms are established. These answers seem to suggest a preference for continued exploration of the topic, within some safeguards.

The majority of respondents (76%) assert that "scaling up current AI approaches" to yield AGI is "unlikely" or "very unlikely" to succeed, suggesting doubts about whether current machine learning paradigms are sufficient for achieving general intelligence.

Overall, the responses indicate a cautious yet forward-moving approach: AI researchers prioritize safety, ethical governance, benefit-sharing, and gradual innovation, advocating for collaborative and responsible development rather than a race toward AGI.

AI Perception vs. Reality

How should we challenge exaggerated claims about AI's capabilities and set realistic expectations?

Main Takeaways

- Over the last 70 years, against a background of constant delivery of new and important technologies, many AI innovations have generated excessive hype.
- Like other technologies these hype trends have followed the general Gartner Hype Cycle characterization.
- The current Generative AI Hype Cycle is the first introduction to AI for perhaps the majority of people in the world and they do not have the tools to gauge the validity of many claims.

CHAIR

Rodney Brooks,
Massachusetts Institute of Technology

Context & History

Artificial intelligence, or AI, is the field that studies the synthesis and analysis of computational agents that act intelligently [6]. AI has gone through hype cycles multiple times since the 1956 workshop that established the name AI and set the course for early computer science departments to include AI as a major component of their research and teaching. All hype bubbles eventually burst, as the essence of hype is that it is beyond reality. Over the decades this has led to AI winters where funding has dried up for all of AI or for specific aspects of AI such as neural networks or robotics.

A study published 2017 on trends with the public perception of AI over a 30-year period found that discussion about AI had sharply increased since 2009 and that discussions in the public press had been consistently more optimistic than pessimistic [4]. The study also found that hopes about AI applications of AI in healthcare and education were increasing over time. Another finding was that concerns were growing about loss of control of AI, ethical implications, and the negative impact of AI on work.

Perhaps the difference in recent years with prior periods is that hype has gone beyond the pages of academic conferences, conference papers, and scientific magazines, out into both the mainstream media, and social media. AI and Artificial Intelligence have become common words that non-technical people have heard about and a common subject for leaders of almost all countries to talk about. Governments, for the first time, have AI policies.

One of the problems is that AI is actually a wide-reaching term that can be used in many different ways. But now in common parlance it is used as if it refers to a single thing. In their 2024 book [5] Narayanan and Kapoor likened it to the language of transport having only one noun, 'vehicle', say, to refer to bicycles, skate boards, nuclear submarines, rockets, automobiles, 18 wheeled trucks, container ships, etc. It is impossible to say almost anything about 'vehicles' and their capabilities in those circumstances, as anything one says will be true for only a small fraction of all 'vehicles'. This lack of distinction compounds the problem of hype, as particular statements get overgeneralized.

The hype also sets expectations for ordinary people. Many are fearful that they will lose their jobs to AI in the short term. Social scientists then work to solve labor disruptions, e.g., for displaced truck drivers [6], based on predictions about AI (and in this case, self-driving trucks) and its adoption that turn out to be wildly optimistic. There are no deployed self-driving trucks in the predicted time frame.

Hype in response to a technology trigger is not restricted to AI. Indeed the business intelligence company Gartner, has deliberately made a practice of using a graphical representation of hype levels through five stages that are common for many technologies: (1) technology trigger, (2) peak of inflated expectations, (3) trough of disillusionment, (4) slope of enlightenment, and (5) plateau of productivity. They have used this framework to track many technologies, including quantum computers, block chain, autonomous vehicles, nano-technology, etc. In November 2024 they [1] estimated that hype for Generative AI had just passed its peak and was on the downswing.

The question for AI professionals is how to respond to this hype, how to question it, and how to help others understand what is hubris, while maintaining their own intellectual modesty and probity. This is hard to do in the middle of outsized claims about one's own field, and often it is up to future historians to carefully dissect past scientific arguments.

Historian Thomas Haigh has tried to do such a dissection, almost in real time, in a recent series of articles in the Communications of the ACM. In [2] he gives a post-mortem on the impact of over-hype in AI that resulted in what is known as the AI-winter in the 1980s. His one line summary is: "Fallout from an exploding bubble of hype triggered the real AI Winter in the late 1980s." In [3] he makes a comparison between the hype of today and of those earlier times. He summarizes this particular opinion piece with the line: "From engines of logic to engines of bullshit?"

Research Challenges

Many of us who have worked in AI for decades face the challenge of trying to remain honest brokers when we see that many of the public statements of people quite new to the field are out of line with reality.

The big question is whether, given the dynamics of social media and the search for clicks, professional opinions and peer reviewed research papers have any impact on dampening the overclaims and the ways they distort common understanding of where AI is, and what is its potential in one year, five years, ten years, etc.

If we are currently left out of the conversations how can we change that?

1. Chandrasekaran, A. [2024]. What's Driving the Hype Cycle for Generative AI, 2024. https://www.gartner.com/en/articles/hype-cycle-for-genai
2. Haigh, T. [2024]. How the AI Boom Went Bust. Vol. 67 No. 2, pp 22–26.
3. Haigh, T. [2025]. Artificial Intelligence Then and Now. Vol. 68 No. 2, pp 24–29.
4. Fast, E. and Horvitz, E. [2017]. Long-term trends in the public perception of artificial intelligence, AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4, 2017, pp 963–969. https://ojs.aaai.org/index.php/AAAI/article/view/10635
5. Narayanan, A. and Kapoor, S. [2024]. AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton University Press.
6. Poole, D. L. and Mackworth, A. K. [2023]. Artificial Intelligence: Foundations of Computational Agents, 3rd edition. Cambridge University Press.
7. Wang, S., Mack, E. A., Van Fossen, J. A., Medwid, M., Cotten, S. R., Chang, C. H., Mann, J., Miller, S. R., Savolainen, P.T. and Baker, N. [2023]. Assessing alternative occupations for truck drivers in an emerging era of autonomous vehicles. Transportation Research Interdisciplinary Perspectives, Vol 19. May. https://doi.org/10.1016/j.trip.2023.100793

Community Opinion

The Community Survey gives perspectives on the reactions to the AI Perception vs Reality theme. First, the results of the survey are summarized here. 36% of the survey respondents chose to answer the questions for this theme. This is the summary breakdown of the responses to each question:

1. How relevant is this Theme for your own research? 72% of respondents said it was somewhat relevant (24%), relevant (29%) or very relevant (19%).

2. The current perception of AI capabilities matches the reality of AI research and development. 79% of respondents disagreed (47%) or strongly disagreed (32%).

3. In what way is the mismatch hindering AI research? 90% of respondents agreed that it is hindering research: 74% agreeing that the directions of AI research are driven by the hype, 12% saying that theoretical AI research is suffering as a result, and 4% saying that less students are interested in academic research.

4. Should there be a community-driven initiative to counter the hype by fact-checking claims about AI? 78% yes; 51% agree and 27% strongly agree.

5. Should there be a community-driven initiative to organize public debates on AI perception vs reality, with video recordings to be made available to all? 74% yes; 46% agree and 28% strongly agree.

6. Should there be a community-driven initiative to build and maintain a repository of predictions about future AI's capabilities, to be checked regularly for validating their accuracy? 59% yes; 40% agree and 29% strongly agree.

7. Should there be a community-driven initiative to educate the public (including the press and the VCs) about the diversity of AI techniques and research areas? 87% yes; 45% agree and 42% strongly agree.

8. Should there be a community-driven initiative to develop a method to produce an annual rating of the maturity of the AI technology for several tasks? 61% yes; 42% agree and 19% strongly agree.

Since the respondents to this theme are self-selected (about a third of all respondents), that bias must be kept in mind. Of those who responded, a strong and consistent (though not completely monolithic) portion felt that the current perception of AI capabilities was overblown, that it had a real impact on the field, and that the field should find a way to educate people about the realities.

Diversity of AI Research Approaches

CHAIR

Peter Stone,
The University of Texas at Austin and Sony AI

It is important to encourage and support research on a variety of AI paradigms, old and new. This includes diverse methodologies (beyond just neural networks) both new and old, interdisciplinary collaboration, and consideration of societal implications.

Main Takeaways

- Historically, the field of AI has simultaneously encompassed many different methodologies and research paradigms.
- There is a risk that the current convergence of the field towards focusing on neural approaches could impede innovation.
- We call for active support of research on classical (non-neural) approaches, as well as research that combines neural approaches with other approaches, and that integrates various paradigms into more complete cognitive architectures. And we especially encourage support of creative investigations of completely new paradigms that may be the key to overcoming the limitations of existing paradigms.

Context & History

There's a long history in the field of AI of separate subcommunities deeply pursuing different approaches to replicating intelligence in computers. Sometimes they have been organized around approaches, such as Planning, Evolutionary Computation, Constraint Satisfaction or Combinatorial Search. Sometimes they have been organized around applications, such as Computer Vision, Natural Language Processing, or Robotics.

While there are usually some areas that are more "in fashion" than others, a few notable area disputes notwithstanding, the community has generally been good about tolerating, and even encouraging, a diversity of approaches. Indeed, it could be argued that the current flourishing of neural-networks-based generative AI is a result of this tolerance. Neural networks were introduced and studied even before the term "Artificial Intelligence" was coined in the 1950s, and there was a lot of research on the topic in the 1960s. After they didn't live up to the level of hype about them at the time, there was a time when the majority of the field considered "connectionism" to be a dead end. But a subcommunity persisted, and eventually got their day in the sun (to say the least).

Current State & Trends

However, there is now a risk that this tradition of diversity will be lost. It is likely that one or more of the currently unfashionable research areas could eventually also see its day in the sun. But we do not currently know which ones. Due to the current dominance of neural approaches, many of the other approaches are losing steam, or even being redefined as no longer AI (A recent article in IEEE Spectrum implied in its headline that classical search is not AI: https://spectrum.ieee.org/chip-design-ai). Indeed, as a community, we seem to be in danger of discouraging newcomers from pursuing any alternatives.

We think that would be a mistake. On the contrary, for the long-term health of the field, it is important that we find a way to support the brave souls who are resisting jumping on the bandwagon, even though their papers may be less likely to be accepted, and/or may accumulate fewer citations. Some of those papers may nonetheless end up being immensely impactful. And even (or especially) for people focussing their attention on neural networks, we find it important that they are knowledgeable about alternative paradigms so as not to impede innovation by needing to "reinvent the wheel".

This isn't to say that we condone ignoring progress in neural-network-based generative AI. It may be the biggest revolution our field has seen, and deserved an enormous amount of attention. Just not all the attention.

Research Challenges

We predict that some of the future breakthroughs will come from other areas, either on their own, or in combination with neural and other classical methods.

For example, investigations of the planning capabilities of Large Language Models find that they are really unable to reason and plan effectively [Valmeekam et al., 2023; Valmeekam et al., 2024]. It may be that some form of symbolic reasoning system is required to work with the LLM to produce sensible plans. Neurosymbolic approaches are moving in that direction. And, similarly, conformal prediction [Angelopoulos and Bates, 2023] is an effort to inject probabilistic reasoning into neural models.

We call on the AI community to complement its deep focus on the capabilities and limitations of neural approaches by actively supporting research on classical (non-neural) approaches such as search, optimization, constraint satisfaction, and causal reasoning, as well as research that combines neural approaches with symbolic and probabilistic approaches, and that integrates various paradigms into more complete cognitive architectures. And we especially encourage support of creative investigations of completely new paradigms that may be the key to overcoming the limitations of existing paradigms.

This support could come in the form of workshops devoted to such investigations, and should also include directing some funding towards these priorities. We should especially seek ways to encourage and support researchers, old and new, who are interested in pursuing new ideas from new perspectives, as well as opportunities for intersecting various new and existing approaches, even though they are likely to struggle to get traction at first.

1. Angelopoulos, Anastasios N. and Stephen Bates, “Conformal Prediction: A Gentle Introduction”, Foundations and Trends in Machine Learning: Vol. 16: No. 4, pp 494–591. http://dx.doi.org/10.1561/2200000101 (2023)
2. Valmeekam, Karthik, et al. “On the planning abilities of large language models (a critical investigation with a proposed benchmark).” arXiv preprint arXiv:2302.06706 (2023).
3. Valmeekam, Karthik, Kaya Stechly, and Subbarao Kambhampati. “LLMs Still Can’t Plan; Can LRMs? A Preliminary Evaluation of OpenAI’s o1 on PlanBench.” arXiv preprint arXiv:2409.13373 (2024).

Community Opinion

57% of those surveyed, or a total of 176 respondents, chose to answer questions pertaining to this theme. We begin by reporting the questions asked and the breakdown of responses.

1. How relevant is this Theme for your own research? 92% of respondents said it was somewhat relevant (23%), relevant (37%) or very relevant (32%).

2. Do you think neural approaches alone are sufficient to achieve general purpose AI agents that match or surpass human intelligence in all ways? 16% of respondents said yes, with the remainder saying no..

3. What percentage of AI research should be devoted to combining neural approaches with other approaches? 94% said at least 25%, including those who said 25% (31% of respondents), 50% (35%) or greater than 50% (28%). One respondent answered 0%.

4. What percentage of AI research should be devoted to purely non-neural approaches? 86% said at least 25%, including those who said 25% (37% of respondents), 50% (38%) or greater than 50% (11%). 3% (6 respondents) answered 0%.

5. What paradigms beyond neural networks do you think deserve the most research attention currently? This open-ended question received 3 responses as follows.

- "We need to understand whether the brain is quantum."
- *"Classic approaches to AI that focus on high-level cognition, rely on structured representations, take a systems-level approach, incorporate ideas from psychology, and aim for theoretical insight rather than winning competitions."*
- *"interdisciplinary collaboration and consideration of ethical and societal implications are extremely important."*

As for all the survey categories, it is important to keep in mind that the respondents to this theme are self-selected (a little more than half of all respondents). Of these respondents, most indicated resonance with the main message of this theme, namely that it is important to invest to some extent in non-neural research. On the other hand, it wouldn't be surprising if many of those who chose not to answer these questions feel differently.

Research Beyond the AI Research Community

Expanding AI research to include diverse perspectives and expertise from outside the core AI research

CHAIRS

Jihie Kim,
Dongguk University

Main Takeaways

- Incorporating perspectives of social scientists, ethicists, and policymakers, to ensure responsible and ethical development and deployment of AI technologies
- AI is not just an engineering discipline but a societal force that reshapes governance, culture, economy, and ethics, requiring holistic approaches for responsible development and deployment.
- 'Intelligent support' and tools that facilitate seamless collaboration between AI researchers and experts from diverse domains, ensuring ethical, explainable, and application-specific AI solutions.

Context & History

'Research beyond the AI research community' emphasizes the importance of expanding AI research to include diverse perspectives and expertise from outside the core AI research community. There are three directions to this expansion: First, we must include perspectives of a.o. social scientists, ethicists, digital humanities, critical data/ communication / media studies, STS, and other disciplines, and policymakers, to ensure responsible and ethical development and deployment of AI technologies. Second, researchers and practitioners in disciplines that increasingly rely on AI (e.g., biology, law, business, neuroscience, cognitive science) may also influence how we think about AI itself. Finally, as multidisciplinary efforts increase we can provide 'intelligent support' and tools for the interaction.

The societal, ethical, legal, and cultural challenges posed by AI highlight the necessity of a multidisciplinary approach to its development and deployment [1,2]. Issues such as biased decisions, privacy breaches, governance, accountability, and inclusivity demand solutions that extend beyond the scope of engineering. Addressing these challenges requires the integration of perspectives from humanities, social sciences, and other fields to ensure that AI systems are aligned with human rights, societal values, and global equity.

This broader approach to AI emphasizes that its development and impact cannot be compartmentalized as solely technical advancements or applications within specific fields, i.e. AI is no longer purely an engineering discipline. Instead, it requires a holistic understanding of AI as a transformative force that reshapes societal structures, cultural norms, economic models, and ethical frameworks. Such an approach considers AI not merely as a tool for individual disciplines but as an integrated phenomenon that influences and is influenced by diverse societal factors. By involving diverse expertise, including ethicists, legal experts, social scientists, and policymakers, we can develop governance frameworks and accountability mechanisms to address concerns like bias, inequality, and unintended societal impacts. Efforts to enhance inclusivity and diversity in AI design can also mitigate risks and maximize the benefits of AI applications.

Additionally, "intelligent support" for these efforts refers to a combination of tools, frameworks, and methods that facilitate effective integration of AI into diverse fields and collaboration between AI researchers and experts in other fields.

Current State & Trends

- Social scientists and ethicists are increasingly involved in developing guidelines on how AI developers should handle personal data, preventing inappropriate surveillance, and ensuring the responsible use of AI. Also governance frameworks such as the EU's AI Act [3] are being actively discussed in order to ensure that AI development is aligned with human rights, justice and societal needs.
- AI has become a key tool for healthcare [4], law [5], business [6], etc., which increasingly influences AI research. For example, AI applications in fields demand highly accurate, explainable, and interpretable AI systems due to the importance of accountability and transparency. This has spurred more research in explainable AI (XAI) as well as AI algorithms tailored for certain applications, such as cancer analysis [7].
- Intelligent support' and tools for the interaction between AI & other disciplines: So far non-AI tools such as GitHub, Kaggle, and research repositories have been actively used by the communities. We expect that intelligent capabilities may be able to promote more productive interactions.

Research Challenges

In formulating future research, here is a set of core pillars that we need to take into account:

- **Societal Dynamics:** AI alters the fabric of human interaction, labour markets, and societal governance. Understanding how automation, algorithmic decision-making, and AI-driven systems impact democracy, social justice, and equality is critical to harness its potential without exacerbating existing disparities.
- **Ethical Integration**: Ethics must be embedded into AI from its inception. This includes navigating dilemmas around privacy, accountability, and fairness. Ethical considerations must go beyond technical checks to involve diverse cultural, philosophical, and societal inputs to create systems that are globally adaptable and contextually relevant.
- **Legal and Regulatory Frameworks:** We need to address the need for

regulation and governance, including reevaluation of existing regulations for intellectual property, liability, and human rights. A collaborative, interdisciplinary effort is essential to craft governance models that increase trust and safety, safeguard public interests and contribute to increased responsible innovation.

- **Cultural Adaptation and Diversity**: As AI technologies become pervasive across cultures and geographies, they must adapt to diverse social norms, languages, and traditions. Ensuring cultural sensitivity in AI design and implementation promotes inclusivity and equitable access to technological benefits.
- **Education and Public Awareness:** The broad impact of AI requires a rethinking of education systems to prepare future generations for an AI-infused world. This includes fostering a transdisciplinary understanding of AI among engineers, social scientists, policymakers, and the general public to build informed, empowered societies.
- **Environmental Sustainability**: The energy demands of AI development and deployment highlight its environmental footprint. A broad approach to AI considers how it can contribute to sustainability efforts, from optimizing resource allocation to advancing climate solutions, while minimizing ecological harm.

The inclusion of other disciplines is more than developing AI technologies to understanding and guiding their interaction with society, but also to ensure that their influence is equitable, inclusive, and beneficial across all domains of human activity.

In supporting these efforts and promoting interaction between AI and other disciplines, future research topics can also include the following:

- **User friendly AI development platforms:** These can enable people from various disciplines to develop AI solutions or train a deep learning model without much programming experience. The platform can also guide ethical development of AI. Intelligent user interfaces and visualization tools can allow people to understand AI model outputs and gain insights.
- **Domain Specific AI tools:** AI systems tailored to the particular needs of a field, such as medical diagnosis, cyber security, law, etc. We look forward to intelligent capabilities that understand the needs of the field workers, and provide seamless support for domain-specific workflows and reasoning processes, which should be much more powerful than current tools such as domain-specific pretrained models or knowledge-based systems
- **AI enabled Collaboration Tools:** Intelligent environments that assist AI researchers and experts in other fields to work together.

1. Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed, Bias and Fairness in Large Language Models: A Survey. Computational Linguistics, 50(3):1097–1179, 2024.
2. Youngsik Yun and Jihie Kim. CIC: A framework for Cultually-aware Image Captioning, International Joint Conference on Artificial Intelligence, 2024.
3. The EU Artificial Intelligence Act, Regulation (EU) 2024/1689, https://artificialintelligenceact.eu/
4. AI in health care, nature collection 22, 2024. https://www.nature.com/collections/dbfcjjigbi
5. D.W. Kite-Jackson2023 Artificial Intelligence (AI) TechReport, American Bar Association, 2023.
6. McKinsey & Company, The state of AI in early 2024: Gen AI adoption spikes and starts to generate value, 2024.
7. Joseph D. Janizek, Ayse B. Dincer, Safiye Celik, Hugh Chen, William Chen, Kamila Naxerova, and Su-In Lee. Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. Nature Biomedical Engineering 7, 811–829, 2023. https://www.nature.com/articles/s41551-023-01034-0

Community Opinion

We began by asking the community which activities they considered most important. The top three priorities were: 1) Promoting interdisciplinary collaboration in AI research 2) Developing AI solutions for specific application areas, such as healthcare, law, and business, and 3) Enhancing public understanding of AI's impact. The next tier of priorities included: 4) Integrating AI education into non-computer science disciplines, 5) Establishing accountability and liability standards, and 6) Addressing societal and cultural impacts of AI through diverse disciplinary perspectives and promoting inclusivity in AI design. While activities such as optimizing resource allocation and developing governance models received slightly lower ratings, the differences were not significant. Overall, the community expressed a strong interest in pursuing a broad range of initiatives.

When asked about key areas of focus, an overwhelming 88% of respondents identified healthcare as the top priority. This was followed by climate (50%), education (45%), and biology (43%). Other areas of interest included manufacturing (38%), business (19%), law (18%), and finance (15%). The respondents listed the following areas as additional candidates: agriculture, transportation, brain diseases, physics, elder care, disaster response, media analytics, and cybersecurity.

Regarding tools to support interdisciplinary collaboration, the highest-rated option was user-friendly AI development platforms (65%), followed by AI-enabled collaboration tools (57%) and domain-specific AI tools (53%). Beyond technological support, the community suggested the following actions for fostering collaboration: educational support, incentives to encourage cross-disciplinary work, formation of interdisciplinary research groups, and sustainable funding for long-term projects. As multidisciplinary efforts become increasingly important to AI research, the community seemed to emphasize the urgent need for stronger support from diverse perspectives.

Approximately one-third of all participants responded to this theme, with the majority (around 90%) identifying AI as their primary field. Although input from other disciplines was limited, the survey suggests a strong interest within our community in integrating diverse perspectives into research and enhanced support for interdisciplinary collaboration.

Role of Academia

State-of-the art AI is now largely driven by the private sector, and universities struggle to compete: they need to find a role in the new era of “big AI”.

CHAIR

Michael Wooldridge,
University of Oxford

Main Takeaways

- The centre of gravity of AI research now lies behind the closed doors of big tech companies.
- Universities cannot compete with big tech companies with respect to the resources – data, compute, and salaries – that are being mobilised by the private sector.
- Universities struggle to retain AI faculty, and struggle to persuade AI graduates to remain in academia.
- The challenge is therefore now to find a role for academia (and publicly funded research) in the new era of “big AI”.

Context & History

While there has always been industrial interest in AI, for much of the history of the field, progress was driven largely by academia. The AI Turing award winners (Feigenbaum, McCarthy, Minsky, Newell, Pearl, Reddy, Simon) were university-based, and the key concepts of both symbolic and neural AI were all developed within academia. The deep learning Turing award winners (Bengio, Hinton, LeCun) all did their Laureate work in universities. But over the past decade, the centre of gravity in cutting edge AI has shifted decisively. In particular, primary work on generative AI is now being carried out largely within the private sector, and a small number of big tech companies in particular. These companies have the data and compute resources available to train large scale state of the art AI systems, and moreover are able to offer salaries that universities could not compete with.

Current State & Trends

These changes raise a number of issues for universities.

First, the demand for AI research talent has led to an exodus of AI talent from universities to the private sector. In one notably early case, Uber hired an entire research lab of roboticists from CMU [1] . Even the world's leading universities are unable to offer packages to compete with the startling salaries offered by big tech AI labs; nor can they muster the associated research resources that such companies routinely offer. The upshot has been a hollowing-out of many university AI groups, with many faculty on seemingly permanent leave, or on part-time contracts that leave them largely disconnected from the conventional university business. Hiring AI faculty has been extremely challenging for more than a decade: PhD graduates prefer to move directly from their studies into the private sector.

At the same time the nature of AI research has changed dramatically. At the turn of the century the dominant paradigm with (for example) the AAAI community was conventionally scientific, drawing from mathematics (definition-lemma-theorem-proof). By contrast, the methodology of deep learning and its associated areas, which dominates at the time of writing, is predominantly that of engineering.

As a consequence the demand within universities for AI courses in particular, and computing courses more generally, has skyrocketed, and this, coupled with the issues we refer to above, has placed intense pressure on departments, whose staff struggle to cope with the level of demand. Some ivy league computing departments report that while they have been given what amounts to a blank cheque for hiring new faculty in high demand areas, they simply find it impossible to hire enough high-quality faculty. In some departments, this situation has reached crisis levels, with faculty reporting stress and other mental health issues as a consequence.

There is a need for all students (regardless of their discipline) to become literate in AI, the use of AI tools is changing education and the way subject are taught, and at the same time students in AI must be educated in a more multi-disciplinary context to be aware of the ethical, legal, societal and economic implications of AI. This in turn places further demands on universities, who need to rethink how they teach and assess students when those students have powerful general purpose AI tools available.

One irony of the present situation is that a key role of universities historically has been to provide a talent pipeline to satiate the demands of the technology world for technical talent. But universities struggle to achieve this simply because they don't have the capacity – because of the hollowing-out of university departments by those same companies.

At the same time, the nature of cutting edge AI research means that universities are simply unable to carry out work that competes with big private sector AI labs. A recent estimate from Meta put the cost of building their latest GPT-class models at ~ $440 million. That is several orders of magnitude outside the scope of all but a tiny number of the world's very richest universities; and most nation states would struggle to put assemble such a package for a national level initiative: in 2023, the UK contemplated a sovereign AI capability, and one option considered was building a "BritGPT". The proposal foundered at an early stage because of (i) cost, (ii) risk, and (iii) being seen to compete with the private sector in an area where innovation (and development costs) have been led by the private sector [2]. Recent developments such as that around the Chinese open-source DeepSeek Large Language model may offer an opportunity in this regard as it could make AI available at lower costs – although it presents challenges of its own with respect to privacy and security.

The objectives of the private sector are

different from those of universities. The private sector is driven mainly by profit, whereas universities aspire to contribute to society through research and education. These goals can be in competition with each other. One effect is that results from the private sector are not always fully available for inspection or evaluation. Also, the high costs of the experiments implies that results are not always reproducible. Academics have a role to play in providing independent advice and interpretations of these results and their consequences. The private sector focuses more on the short term, and universities and society more on a longer term perspective.

These observation together with the massive costs of generative AI research has also
motivated calls for large public-funded initiatives, for instance, the plan supported by European Commission President Ursula von der Leyen for a CERN for AI. This vision is supported and was actually initiated by a large number of AI researchers in Europe gathered in the CAIRNE network. It is modelled after the renowned CERN for Nuclear Research in Geneva, and should serve as an alternative and attractive research environment to big tech companies, and should solve problems of public interest. Numerous countries have also developed their national strategies and funding programs for AI.

Research Challenges

- How should universities respond to the new era of "Big AI"?
- What form of AI research can universities/public sector research most usefully engage in – what should be the AI research agenda of universities going forward?
- How should universities respond to the challenges of hiring and retaining AI faculty and students?
- How can publicly funded universities best work with big tech AI companies?

1. https://www.fastcompany.com/3046902/carnegie-mellon-in-a-crisis-after-uber-poached-40-of-its-researchers
2. https://lordslibrary.parliament.uk/large-language-models-and-generative-ai-house-of-lords-communications-and-digital-committee-report/

Community Opinion

Most respondents (approx. 75%) agreed the universities have difficulty recruiting in AI, and agreed that universities have difficulty in engaging in resource intensive AI research (80%). Respondents felt that offering better salaries and investing in better compute resources were ways of attracting people; offering joint appointments and providing other benefits were also seen as possibilities. There was less agreement on whether universities needed to refocus their research priorities, although it seems that respondents felt that focussing on theoretical AI and multidisciplinary AI were areas where universities could be competitive. Public sector funding for large scale compute was seen as attractive (70%). There was overwhelming agreement that academia was very relevant to the future of AI research.

Geopolitical Aspects & Implications of AI

The rise of AI is reshaping global power dynamics and the investment priorities of nations, influencing economic, security, and governance structures, while posing challenges to equity and control.

Main Takeaways

- Investments, coordination, best practices, and regulation of AI have become international in scope. While collaboration among nations is increasing with governmental and non-governmental programs, AI is also a geopolitical battleground, with countries competing for economic, military, and strategic dominance: Nations are seeking to leverage AI to gain economic, military, and strategic advantages.
- Regulation vs. competition: The tensions between AI regulation, confidentiality, and the race for technological supremacy complicates international collaboration.
- Ethical and social impact: The deployment of AI by nation states, per policies and competitive goals raises concerns over justice, fairness, and democratic values, requiring new governance frameworks, some of which will need to be international in scope.

CHAIRS

Virginia Dignum,
Umeå University

Holger Hoos,
RWTH Aachen University, Germany and Leiden University, The Netherlands

Eric Horvitz,
Microsoft

Context & History

AI has shifted from being solely a research and technology driven field to becoming a key element of global economic and security strategies, with governance efforts emerging to shape its responsible use [1,2]. Recent developments, including advancements in large language models and AI-powered automation, have intensified anxieties and competition among nations, particularly between the U.S., China, Russia, and the European Union.

The current AI landscape is increasingly shaped by economic interests and competing approaches to governance. Over the past several years, countries have taken different yet related perspectives on strategy, investments, and governance of AI both domestically and in their international engagements, and in their proclamations and engagements internationally. Thes dynamics have also been subject to shifts driven by political changes.

In the U.S., President Biden issued an Executive Order [3] on "Safe, Secure, and Trustworthy Artificial Intelligence" in October 2023, building upon the AI Bill of Rights, developed by the Office of Science and Technology Policy [4]. The order emphasized the protection of civil rights and privacy, and mandated rigorous standards for AI safety and security. U.S. Federal agencies were tasked with responsibilities with fielding and governing AI systems that are transparent, equitable, and free from algorithmic discrimination, aiming to prevent AI from exacerbating biases or infringing upon individual rights. The U.S. also set up a national AI Safety Institute [5] and worked with other nations to build an International Network of AI Safety Institutes [6].

In January 2025, Donald Trump revoked the Biden executive order shortly after his inauguration, replacing it with the Executive Order on "Removing Barriers to American Leadership in Artificial Intelligence," which focused on sustaining and enhancing America's global dominance "...in order to promote human flourishing, economic competitiveness, and national security." [7]

The European Union's AI Act [8], enacted in August 2024 after four years of deliberations, employs a product safety approach to AI regulation that classifies AI systems based on risk levels—unacceptable, high, limited, and minimal—and imposes corresponding obligations. High-risk AI applications, such as those in healthcare and transportation, must comply with strict safety, transparency, and oversight requirements to ensure they do not compromise health, safety, or fundamental rights.

China has described AI as a "major strategic opportunity" and has called for the country to be a world leader in AI by 2030 [9]. China was one of the first countries to introduce regulations that govern the use of AI systems, including detailed regulations governing recommendation algorithms that went into effect in 2021 [10]. China continues to integrate AI into its surveillance infrastructure.

In 2019, Russian President Vladimir Putin issued a decree for accelerating the development of AI in the Russian Federation with a scope extending to 2030. In 2023, the decree was updated to consider a plan for development including the laying out of key principles for AI development "like protecting human rights, ensuring security, technological sovereignty and supporting competition." [11]

The need for international coordination and agreements on governance of multiple aspects of AI, including defense and establishing norms for human rights and principles for the responsible fielding of AI technologies, emphasized in the 2020 report of the Congressionally mandated U.S. National Security Commission on AI (NSCAI) [12]. The report called for alliances of nations sharing Western democratic values to coordinate strategies. On the side of defense, the report called for establishing international venues to discuss the impact of AI on crisis stability among competitor nations and to develop international standards of practice for the development, testing, and use of AI-enabled and autonomous weapon systems.

Recent international efforts underscore the growing recognition of cooperative AI governance. Organizations such as the OECD [13], the UN [14] and GPAI have advocated for principles of global international governance. High profile international meetings—such as the UK AI Safety Summit (Bletchley Park in November 2023), the AI Seoul Summit (May 2024), and the AI Action Summit (Paris in 2025)—demonstrate an ongoing commitment to global coordination.

At the Seoul Summit, representatives of Australia, Canada, the European Union, France, Germany, Italy, Japan, the Republic of Korea, the Republic of Singapore, the United Kingdom, and the United States of America affirmed a common "dedication to fostering international cooperation and dialogue on artificial intelligence (AI) in the face of its unprecedented advancements and the impact on our economies

and societies." [15]. In November of 2024, the AI Safety Summit in San Francisco followed up on the AI Seoul Summit declaration by launching the International Network of AI Safety Institutes [16]. Discussions among governments, civil society groups, and industry have continued across multiple forums, including the Partnership on AI multiparty stakeholder organization.

Needs and opportunities for international coordination are growing around norms and regulations on human rights, privacy, and safety of AI systems. Rising issues of international scope include questions about regulations with considerations of intellectual property with regard to the data used to train large language models. Key concerns that span borders include the handling of AI-generated synthetic content employed in disinformation and with AI-based threats in the realm of biosecurity, such as the use of AI-powered protein design tools for developing dangerous toxins and pathogens.

Challenges persist as national interests and regulatory approaches remain fragmented, fueling tensions over the role of AI in trade, security, and human rights. These divisions were evident during the AI Action Summit in February 2025, where the EU and China pushed for stricter AI regulations, while the U.S. and UK opted out of a global AI safety declaration, citing concerns that overly stringent rules could stifle innovation. France and other EU digital leaders, meanwhile, advocated for a more flexible regulatory framework to attract investment and prevent stagnation.

While these summits have certainly facilitated useful discussions surrounding AI safety and regulation, they have been criticized for excluding many countries, particularly from the Global South. Also, regions such as Southeast Asia, despite being active in AI safety and regulation, often have limited participation in global AI safety discussions. This selective participation has led to questions about the legitimacy and effectiveness of these gatherings in addressing global AI challenges [26], and to guarantee significant commitments to AI safety, with many indicating a lack of concrete safety measures, vague policy recommendations, and an overemphasis on speculative risks rather than immediate AI challenges, despite the publication of the International AI Safety Report 2025 [27].

At the same time, economic competition remains intense. The U.S. Stargate initiative, valued at $500 billion, aims to strengthen the country's AI infrastructure and global competitiveness, while the EU's €200 billion InvestAI initiative seeks to drive AI research and deployment across Europe. India, meanwhile, is emphasizing equitable AI development, calling for greater inclusivity and open-source collaboration to ensure AI benefits all regions.

While these efforts reflect growing recognition of the transformative potential of AI technologies, they also underscore a persistent challenge: without a unified governance framework, AI regulation will likely remain fragmented, exacerbating risks related to security, economic inequality, and geopolitical instability. Without coordinated international agreements, these divisions could reinforce existing global inequalities and deepen geopolitical tensions over AI control and governance [18].

Current State & Trends

1. AI is increasingly a defining factor in national and regional power, influencing trade policies, military strategies, and diplomatic relations [19]. The U.S. and China currently dominate AI leadership, while Europe prioritizes ethical considerations and regulatory oversight. At the same time, AI-driven surveillance and data collection are shaping governance models worldwide, particularly in autocratic regimes that prioritize state security over individual freedoms, raising concerns about privacy and civil liberties.

2. The regulatory landscape remains fragmented [20], with stark differences in governance approaches among major global players. The EU has advanced regulation in the form of the AI Act to promote AI safety and accountability. However, under a new administration, the U.S. has pivoted away from its recent intensive focus on AI safety and human rights-centered regulation.

3. Governments have been coordinating via such efforts as the set of international meetings at the UK, Seoul, and Paris, and the launch of the International Network of AI Safety Institutes.

4. Despite the absence of binding international agreements, coalitions of private corporations, civil society groups, and non-profits have worked on voluntary agreements and standards. Examples include:

- Addressing concerns about AI-powered disinformation and manipulation: The Coalition for Content Provenance and Authenticity (C2PA), which developed a cryptographic standard for media provenance, now adopted

by major tech companies, and the Tech Accord to Combat Deceptive Use of AI in 2024 Elections, which established commitments regarding AI-generated content.

- Addressing AI-enabled biosecurity challenges: Efforts have focused on developing principles and best practices for responsible AI use in biosciences [21] and international coordination of nucleic acid screening protocols [22].

5. One of the most contentious issues is the debate over open vs. proprietary AI models. Key concerns include accessibility, bias, transparency in training data, and the potential for state and non-state actors to misuse AI. Additionally, geopolitical tensions have intensified, as the U.S. imposes restrictions on semiconductor exports to several countries, including China as well as some EU member states, affecting global supply chains and exacerbating technological divides between major AI powers.

6. Beyond governance and competition, AI presents ethical and societal challenges [23]. AI-driven decision-making in critical sectors such as human resources, healthcare, law enforcement, and financial services raises concerns about bias, discrimination, and social inequality. The rapid development of autonomous and semi-autonomous weapons and AI-powered military systems by global powers such as the US, China, Russia and Ukraine further complicates international security, challenging existing norms of warfare and accountability. To mitigate these risks, AI governance must balance the need for innovation with strong ethical safeguards, ensuring AI technologies are developed and deployed in a way that promotes fairness, security, and equitable distribution of benefits across societies.

Research Challenges

AI Governance Models

- Developing AI governance frameworks, treaties, norms, and practices that are international in scope: Study on the prospects of harmonizing AI regulations across nations, reducing fragmentation and conflicts in global AI governance. Areas of international cooperation can focus on norms and regulations around surveillance and human rights, intellectual property challenges with AI, norms, treaties, principles, accountabilities, and best practices around the development, deployment and use of autonomous and semi-autonomous weapon systems, agreements on international norms and regulations on AI and biosecurity, and agreements around the threat of AI-generated misinformation, with regulations and best practices around establishing the provenance of authentic and AI-generated content.
- Strengthening enforcement mechanisms in global AI governance: analyzing how organizations such as the UN, OECD, and GPAI can enhance compliance and accountability in international AI cooperation.
- AI and global governance frameworks: studying the role of AI in shaping international regulatory frameworks, including how different governance models influence geopolitical stability.

Geopolitical Risks of AI

- AI-driven misinformation and influence campaigns: investigating the role of AI in generating and combating deepfake technology, automated disinformation, and propaganda used by state and non-state actors in geopolitical conflicts
- AI and supply chain geopolitics: developing AI tools to monitor and mitigate disruptions in global AI-related supply chains, particularly regarding semiconductor shortages and export restrictions.
- Algorithmic trade policies and economic forecasting: enhancing AI models that predict and analyze the impact of AI-driven automation and trade restrictions (e.g., semiconductor export bans) on global markets).
- AI in cybersecurity and defense: advancing AI-driven cybersecurity and cyber-physical threat detection and response [24] and resilience mechanisms to counter state-sponsored cyberwarfare and protect critical national infrastructure [25].
- AI and biosecurity challenges: advancing coordinative activities such as regulating screening protocols for nucleic acid synthesis organizations, regulation of benchtop synthesis, and logging of orders to detect and deter abuse [22].
- AI in military strategy, including AI-informed decision making and the development and use of autonomous weapons: researching the implications of AI-driven autonomous weapons, with regard to stability, crisis management, and strategic deterrence, ensuring accountability and compliance with international humanitarian law

Promoting Ethical AI Development

- Developing interdisciplinary frameworks for AI fairness and accountability: researching how political science, economics, and ethics can inform AI governance models that balance national interests, corporate incentives, and global equity.
- Advancing AI governance in geopolitically fragmented environments: exploring legal, diplomatic, and technological strategies to overcome economic and ideological divisions, enhancing the enforceability of global AI agreements, such as the G7 AI Code of Conduct.
- Examining the risks of techno-solutionism in AI policy: investigating the unintended consequences of AI-driven decision-making through multidisciplinary research, ensuring AI complements rather than overrides human agency and ethical governance.

1. Theodorou, Andreas, and Virginia Dignum. "Towards Ethical and Socio-Legal Governance in AI." Nature Machine Intelligence, vol. 2, no. 1, 2020, pp. 10–12.
2. Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W. G. (2022). Governance of Artificial Intelligence: Emerging international trends and policy frames. In The global politics of Artificial Intelligence. Taylor & Francis.
3. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Washington DC: The White House; October 30, 2023. EO 14110. Federal Register: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence
4. The White House (2022). "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People." https://www.whitehouse.gov/ostp/ai-bill-of-rights/
5. U.S. Artificial Intelligence Safety Institute (2024). https://www.nist.gov/aisi
6. International Network of AI Safety Institutes (2024). https://www.nist.gov/system/files/documents/2024/11/20/Mission%20Statement%20-%20International%20Network%20of%20AISIs.pdf
7. Executive Order on Removing Barriers to American Leadership in Artificial Intelligence. Washington DC: The White House; January 23, 2025. EO 14179 of. Federal Register: Removing Barriers to American Leadership in Artificial Intelligence. https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence
8. European Union. Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). 2024.
9. China's 'New Generation Artificial Intelligence Development Plan' (2017). https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/
10. Provisions on the Management of Algorithmic Recommendations in Internet Information Services (2021). https://www.chinalawtranslate.com/en/algorithms/
11. Sergey Sukhankin (2017). Russia Adopts National Strategy for Development of Artificial Intelligence, Eurasia Daily Monitor Volume: 16 Issue: 163 https://jamestown.org/program/russia-adopts-national-strategy-for-development-of-artificial-intelligence/
12. Safra Catz, Steve Chien, Mignon Clyburn, et al. Report of the National Security Commission on Artificial Intelligence Report of the National Security Commission on Artificial Intelligence, National Security Commission on Artificial Intelligence (NSCAI), March 2021. https://reports.nscai.gov/final-report
13. OECD (2024), "Governing with Artificial Intelligence: Are governments ready?", OECD Artificial Intelligence Papers, No. 20, OECD Publishing, Paris, https://doi.org/10.1787/26324bc2-en.
14. United Nations. Governing AI for Humanity: Report of the High-Level Advisory Body on Artificial Intelligence. United Nations, 2024.
15. Seoul Declaration for Safe, Innovative and Inclusive AI by Participants Attending the Leaders' Session of the AI Seoul Summit, May 2024, Seoul, Korea. https://www.pm.gc.ca/en/news/statements/2024/05/21/seoul-declaration-safe-innovative-and-inclusive-ai-participants-ai-seoul-summit
16. The International Network of AI Safety Institutes: Mission statement. November 2024. https://ised-isde.canada.ca/site/ised/en/international-network-ai-safety-institutes-mission-statement
17. Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: barriers and pathways forward. International Affairs, 100(3), 1275–1286.
18. Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). The global governance of artificial intelligence: Next steps for empirical and normative research. International Studies Review, 25(3), viad040.
19. Larsen, B., 2022. The geopolitics of AI and the rise of digital sovereignty, Brookings Institution. United States of America. Retrieved from https://coilink.org/20.500.12592/swc5mh on 21 Feb 2025. COI: 20.500.12592/swc5mh.
20. Schmitt, L. (2022). Mapping global AI governance: a nascent regime in a fragmented landscape. AI and Ethics, 2(2), 303–314.
21. D. Bloomfield, D., Pannu, J. Zhu, A.W., et al. AI and biosecurity: The need for governance (2024). Science 385(6711) pp. 831–833 DOI: 10.1126/science.adq1977 https://www.science.org/doi/10.1126/science.adq1977
22. Wittmann BJ, Alexanian T, Bartling C, Beal J, Clore A, Diggans J, et al. Toward AI-resilient screening of nucleic acid synthesis orders: Process, results, and recommendations. bioRxiv. 2024. p. 2024.12.02.626439. doi:10.1101/2024.12.02.626439 https://www.biorxiv.org/content/10.1101/2024.12.02.626439v1
23. Floridi, Luciano, The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities (Oxford, 2023; online edn, Oxford Academic, 24 Aug. 2023), https://doi.org/10.1093/oso/9780198883098.001.0001
24. Horvitz, E. Artificial Intelligence and Cybersecurity: Rising Challenges and Promising Directions. Testimony before the U.S. Senate Armed Services Subcommittee on Cybersecurity, May 3, 2022. https://www.armed-services.senate.gov/imo/media/doc/5.3.22%20Eric%20Horvitz%20Testimony.pdf
25. President's Council on Science and Technology, Strategy for Cyber-Physical Resilience: Fortifying Our Critical Infrastructure for a Digital World (2024). White House Office of Science and Technology, February 2024.https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/02/PCAST_Cyber-Physical-Resilience-Report_Feb2024.pdf
26. Agnew, William, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. "The illusion of artificial inclusion." In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–12. 2024.
27. Bengio, Yoshua, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi et al. "International AI Safety Report." arXiv preprint arXiv:2501.17805 (2025).

Community Opinion

The survey results highlight concerns surrounding AI governance, security, economic shifts, and ethical considerations. While the geopolitical impact of AI is acknowledged, there are few researchers for whom this is their primary focus. A majority (49.47%) believe AI scientists should engage in policy discussions, with strong support for international governance mechanisms such as the UN (53.68%) and bilateral agreements (63.16%). Key challenges include cybersecurity, warfare, economic displacement, and the balance between government and corporate control.

Military AI applications raise ethical concerns, with 36.84% strongly agreeing and another 37.89% agreeing on their significance. Support for international cooperation is strong, with over 40% advocating for agreements on the use of AI in public data, weapon deployment restrictions, and privacy regulations. Respondents stressed the need for enforceable, specific agreements in contrast to symbolic declarations.

About AAAI

Founded in 1979, the Association for the Advancement of Artificial Intelligence (AAAI) (formerly the American Association for Artificial Intelligence) is a nonprofit scientific society devoted to advancing the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines.

AAAI aims to promote research in and responsible use of artificial intelligence. AAAI also aims to increase public understanding of artificial intelligence, improve the teaching and training of AI practitioners, and provide guidance for research planners and funders concerning the importance and potential of current AI developments and future directions.

601 Pennsylvania Ave, NW
Suite 900
Washington, DC 20004

info@aaai.org
1-202-360-4062
aaai.org