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.