In: *Realizing the Promise and Minimizing the Perils of Artificial Intelligence for the Scientific Community*. Eds Kathleen Hall Jamieson, Anne-Marie Mazza, and William Kearney. University of Pennsylvania Press. Forthcoming.

A Moment in Time for AI: Reflections on Science and Society

Eric Horvitz
June 2024

Reflecting on the current state of artificial intelligence (AI), I find myself immersed in two interrelated realms: the scientific advancements of AI and their societal impacts. We are in an exciting period for AI, with the capabilities of neural network models emerging faster than our understanding of their underlying principles. These advancements have stimulated scientific curiosity and catalyzed new directions for AI research, bringing novel questions, energy, and intensity to colleagues and teams I collaborate with. Simultaneously, the rapid diffusion of AI tools into everyday life has deepened my sense of responsibility regarding their societal influences. I have invested increasing time and resources in reflecting on and addressing potential disruptions, ethical concerns, and the opportunities AI presents in various realms.

Scientific Journey

I was drawn to do my doctoral work in AI as a path to gaining an understanding of the mysteries of human cognition. I contributed to the ignition of a probabilistic revolution in AI, moving away from the dominant logic-based methods of the time, and working to advance the development of AI systems based on a foundation of probability and utility theory. The axioms of probability, extended to taking ideal actions in the world via the axioms of utility theory, form a widely assumed and celebrated normative basis for reasoning and decision making.[1] I focused during my doctoral efforts on developing models of bounded rationality built on probability and utility that could enable systems with limited computational resources to perform well amidst the complexity of the open world.[2] The work included the development of formal mechanisms for guiding evidence gathering and inference. Other teams explored numerous other approaches for leveraging probability in representations and reasoning. This shift to a *rationalist* approach to AI—harnessing a normative foundation of probability and utility—became central in advancing machine learning, perception, reasoning, and decision making.[3] The approach enabled the community to build systems that could address real-world challenges, such as providing physicians with well-justified recommendations on medical diagnoses and decisions. The rationalist approach provided clear semantics and a strong theoretical foundation for building systems operating on understandable and sound principles.[4]

---

[1] von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press, 1944.
[2] Horvitz, E. J. (1991). Computation and action under bounded resources. Stanford University.
[3] Horvitz, E. J., Breese, J. S., & Henrion, M. (1988). Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, 2(3), 247-302.
[4] Heckerman, D. E., Horvitz, E. J., & Nathwani, B. N. (1992). Toward normative expert systems: Part I the Pathfinder project. *Methods of Information in Medicine*, 31(02), 90-105.

Recent advancements in neural network models mark a significant inflection point in AI's trajectory.[5] Impressive capabilities and rates of improvement are seen in vision, speech recognition, and language understanding benchmarks. *Generative AI* has recently emerged with models being built at increasing scales demonstrating surprising powers in generating language, images, video, and molecules. Neural network models are being harnessed in numerous areas, including the sciences. For example, advances in predicting protein structure and drug design are accelerating research in the biosciences, including efforts to design new therapeutics.

Despite the excitement, we grapple with the relationship of neural models to prior advances. We have a poor understanding of the capabilities of AI systems. In distinction to the clarity of previous work on the rationalist approach, much of the detailed operation of generative models remains a mystery. Neural networks have thrust us into empirical studies of these large-scale systems, akin to methodologies for studying nervous systems.[6,7,8] This jump, from a successful multi-decade trajectory of advances with rationalist approaches to the mysteries of neural networks, frames intriguing and interesting opportunities in the science of AI to pursue answers to significant questions that remain largely unanswered today. We face a critical scientific challenge of bridging the gap between empirical observations of the behavior of neural networks and foundational principles of well-understood theories of inference and action.[9] I hope to see bridges constructed over the next decade.

Societal Implications and Responsibilities

AI scientists and engineers have an important role and responsibility to identify and share technical developments that have implications for people and society more broadly. This includes informing and engaging with multiple stakeholders across domains and sectors and working to broaden awareness and participation. This work involves being available for expert consultations, organizing and participating in special meetings and engagements around milestone developments, and establishing organizations and initiatives for tracking, guiding, and communicating AI advances over time.

Fifteen years ago, AI was beginning to make its way into real-world applications as I assumed the presidency of the Association for the Advancement of Artificial Intelligence (AAAI). I themed my presidency "AI in the Open World," highlighting the need to develop AI systems that could perform robustly and in a trustworthy manner on real-world tasks, and also our responsibility to understand and address the potential societal impacts of AI systems.[10] To explore societal influences, I commissioned the AAAI *Presidential Panel on Long-Term AI*

---

[5] Bubeck, S., Chandrasekaran, V., Eldan, R., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4, arXiv preprint arXiv:2303.12712, March 22, 2023. https://arxiv.org/abs/2303.12712

[6] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 17359-17372.

[7] Olsson, C., Elhage, N., Nanda, N., et al. (2022). In-context learning and induction heads. CoRR, abs/2209.11895.

[8] Yuksekgonul, M., Chandrasekaran, V., Jones, E., et al. (2024). Attention satisfies: A constraint-satisfaction lens on factual errors of language models, *ICLR* 2024.

[9] Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273-278.

[10] Horvitz, E. (2008, July 15) *Presidential Address: AI in the Open World*, *AAAI National Conference*, July 2008, Chicago, Il. Transcript: https://erichorvitz.com/AAAI_Presidential%20Address_Eric_Horvitz.pdf

Futures. This initiative culminated in a retreat at Asilomar in 2009, chosen for its symbolic connection to the historic meeting on recombinant DNA.[11] The clear value of the discussions and collaborations at the AAAI Asilomar retreat and pre-meetings sparked the establishment five years later of the *One Hundred Year Study* on AI at Stanford, which was created to bring experts together every five years to observe, synthesize, and provide assessments and guidance in the spirit of the AAAI Asilomar meeting.[12] The study is endowed to continue this process for the lifespan of Stanford University. Projects of the study include the creation of faster-paced analyses, including the *AI Index*, an annual assessment of AI capabilities and influences.[13]

Beyond recurrent studies by experts, the ubiquity of AI's influences requires that diverse voices participate and help to guide AI. AI scientists have a responsibility to organize, alert, and educate a spectrum of stakeholders—as well as to establish venues for listening and responding. In 2016, AI scientists from industry, academia, and non-profit research centers co-founded the Partnership on AI, bringing together stakeholders from industry, academia, and civil society to foster discussions, analyses, and make recommendations on the responsible advancement of AI.[14] As the founding chair, I've observed the power of bringing scientists together with policymakers, civil liberties experts, and a broad swath of civil society organizations. While still in its first decade, the Partnership on AI has made significant contributions to multiparty collaboration on key topics.

With potential fast-paced developments, AI scientists may need to engage quickly at times and bring diverse expertise to the table as early as possible when new capabilities and issues arise. Given the behaviors I saw in our internal studies of an early pre-release version of GPT-4 in August of 2022, I felt it important to gain permission to share the confidential pre-release model with experts across a spectrum of disciplines. This initiative led to the *AI Anthology* effort, which provides multiple viewpoints on how the new capabilities might be best leveraged for human flourishing.[15]

AI scientists also need to inform and provide guidance to government agencies and leaders about technical advancements with AI and work with policymakers on steps forward. It's been an honor to testify on AI at both open hearings and closed sessions of Congress[16,17]—and to have

---

[11] Presidential Panel on Long-Term AI Futures (2009). Association for the Advancement of Artificial Intelligence (AAAI), Asilomar Conference Center, Pacific Grove, CA, United States, February 2009. https://aaai.org/about-aaai/aaai-presidential-panel-on-long-term-ai-futures-2008-2009

[12] Horvitz, E. (2014) One hundred year study on AI: Reflections and framing, Stanford University. https://ai100.stanford.edu/about/reflections-and-framing

[13] *The AI Index Annual Report* 2024 (2024), Stanford University. https://aiindex.stanford.edu/report/

[14] Partnership on AI, https://partnershiponai.org

[15] Horvitz, E. (ed.) (2023) *AI Anthology.* https://unlocked.microsoft.com/ai-anthology/eric-horvitz

[16] Horvitz, E. (2014) Reflections on the Status and Future of Artificial Intelligence, Testimony Before the United States Senate, *Hearing on the Dawn of Artificial Intelligence*, Committee on Commerce Subcommittee on Space, Science, and Competitiveness, November 30, 2016. Testimony: https://erichorvitz.com/Senate_Testimony_Eric_Horvitz.pdf Video: https://youtube.com/watch?v=fl-uYVnsEKc

[17] Horvitz, E. (2022) AI and Cybersecurity: Rising Challenges and Promising Directions, In: Hearing on AI Applications to Operations in Cyberspace before the Subcommittee on Cybersecurity, of the Senate Armed Services Committee, 117th Congress, May 3, 2022. https://erichorvitz.com/Testimony_Senate_AI_Cybersecurity_Eric_Horvitz.pdf

opportunities to engage with senior leadership at the White House and colleagues via my role as a member of the President's Council of Advisors on Science and Technology.

These diverse projects and engagements are examples of AI scientists' responsibilities to engage and inform across sectors, to work to broaden awareness and participation, and to promote research on AI's responsibilities, ensuring we include multiple voices in decisions and stay ahead of the innovation wave with technical, sociotechnical, and regulatory advancements.

Moving Forward

Looking ahead, the interplay between AI's scientific advancements and societal impacts will become even more critical. We urgently need to grow our scientific understanding of the operation of systems built on neural-network methodologies. Better scientific understandings will help us to shape the development and application of AI methods. We need to complement curiosity-driven research and the thrill of scientific breakthroughs in AI foundations with investments in technology and policy to understand, shape, and regulate influences of the technologies on people and society. This work includes ongoing study spanning technology, design, and psychology of human-AI interaction.[18]

The potential benefits of AI are immense—from enhancing scientific discovery to improving educational and raising the quality of healthcare outcomes. However, we have to consider recognized risks, particularly with information and media integrity, biosecurity, fairness and equity, safety and reliability, and privacy and security. We must also stay on top of "deep currents" of more complex interactions of AI with culture and society, such as how these systems may change and disrupt—in costly and in valuable ways—education, the creative arts, scientific discovery, jobs and the economy. We must work to monitor and come to better understandings of the subtle but potentially powerful influences of AI applications on the human psyche, including the impacts on our human dignity and agency.[19] Outcomes need not be dominated by situations and equilibria reached via laissez-fair flows of technology into society. With the maturation of AI and its applications, we have opportunities to manage and guide the technology with foresight and responsibility.

The current state of AI is marked by fast-paced progress and significant challenges. As a scientist driven by curiosity about human cognition and devoted to reaching understandings of computational principles of intelligence, I'm excited by potential AI discoveries, machinery, and new applications on the horizon. At the same time, I am cautious and concerned about how AI innovations might be harnessed. Our task is to steer AI's development to promote human well-being and societal progress. Through continued scientific exploration and a thoughtful, inclusive, multidisciplinary approach to applications and influences, we can leverage AI as a force for

---

[18] Sellen, A. and Horvitz, E. (2024) The rise of the AI co-pilot: Lessons for design from aviation and beyond, *Communications of the ACM*, July 2024. https://doi.org/10.1145/3637865

[19] Horvitz, E., Conitzer, V., McIlraith, S, and Stone, P. (2024) Now, later, and lasting: 10 priorities for AI research, Policy, and Practice. *Communications of the ACM,* June 2024, 39–40. https://doi.org/10.1145/3637866

good, advancing our understandings of the scientific foundations of intelligence and enriching human society. AI scientists, with their unique insights, must lead at the frontier, providing awareness of developments and implications, and engaging with the public, civil society organizations, government leaders and agencies, and experts across various fields to address these responsibilities and to help shape AI's future.