PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-kok9-bm6o **Comments Due:** May 28, 2025 **Submission Type:** API

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0198
Comment on FR Doc # 2025-07332

Submitter Information

Organization: RAND

General Comment

Please find our inputs to this RFI attached in a PDF.

Attachments

NSF_RFI_AI_CriticalInfrastructure

ISMAEL ARCINIEGAS RUEDA, DANIEL TAPIA, JOSHUA KAVNER, HENRI VAN SOEST

Input on the 2025 National AI R&D Strategic Plan

RAND

May 2025

Summary

This document provides public comments from the authors in response to the 'Request for Information on the Development of a 2025 National Artificial Intelligence (AI) Research and development (R&D) Strategic Plan," issued by the National Science Foundation (NSF).

In line with Executive Order 14179 (Removing Barriers to American Leadership in Artificial Intelligence), the Office of Science and Technology Policy (OSTP), the Networking and Information Technology Research and Development (NITRD) National Coordination Office (NCO) are soliciting input on how the previous administration's National Artificial Intelligence Research and Development Strategic Plan (2023 Update) (NAIRD) can be rewritten so that the United States can secure its position as the unrivaled world leader in AI. The stated goals are to promote R&D to accelerate AI-driven innovation, enhance U.S. economic and national security, promote human flourishing, and maintain the United States' dominance in AI while focusing on the Federal government's unique role in AI R&D over the next 3 to 5 years.

This submission provides input for the National Artificial Intelligence Research and Development Strategic Plan (2025 Update) as it relates to two strategies in the 2023 NAIRD pertaining to public-private partnerships (Strategy 8) and sharing data and testbeds (Strategy 5). In particular, this submission discusses research and development of AI applications in the power grid as it both serves as an illustration of our motivations and recommendations and is an example area of interest as stated in the RFI.

We make our recommendations so that the mixed (e.g. competitive and cooperative) incentives of principals or stakeholders are considered in the research, development, and deployment of AI systems they employ. In the context of the power grid, it is important that AI systems are not designed with the implicit assumption that stakeholders have only cooperative incentives and no competitive incentives. Guiding the development and deployment of AI solutions to avoid miscoordination, competition, and collusion[i] requires conducting research that is sensitive to mixed incentives, but also directly engaging and informing private and public stakeholders in the power grid. Recognizing that our discussion is focused on AI in the power grid but is not necessarily limited to it, we recommend that the 2025 NAIRD build on the following strategies from previous NAIRD:

- Expand Public-Private Partnerships to Accelerate Advances in AI to integrate and acclimate Stakeholders (Strategy 8)
    - The 2025 NAIRD should support research on multiagent systems where different types of agents interact to better understand the resulting dynamics and guide development of those systems, especially for those on the power grid. This should be informed by engagement with power grid stakeholders.
    - The 2025 NAIRD should support research that develops and evaluates tools, techniques, materials, and methods that could be useful for informing

stakeholders with operational understandings of AI systems. These research outputs could take the form of white papers, technical publications, symposiums or tabletop exercises involving AI systems to inform stakeholders' understandings.

- Develop Shared Public Datasets and Environments for AI Training and Testing which Capture Stakeholder Incentives and Behaviors (Strategy 5)
    - The 2025 NAIRD should support R&D that identifies the minimum set of necessary information required from actors to preserve privacy or other interests and still contribute to improvements in multiagent systems. In the power grid to support AI systems oversight by regulators and Independent System Operators (ISO). For companies in the power grid, information sharing (e.g., scheduled outages) and privileged (e.g., bidding strategies of cost data) is both critical for transparency and competition. Development algorithms or applications that account for information privacy and still confer benefits to the grid and companies will be important to ensure that R&D and innovation is self-sustaining.
    - The 2025 NAIRD should support R&D that stress tests AI models within a high-fidelity simulation, that includes economic and human behavioral elements, differential adoption, as well as supporting but not themselves critical operations, like maintenance. While possible, it is not guaranteed that power grid stakeholders naturally converge on a set of well-known and understood models. It is likely that technology adoption will not be uniform.

Opportunities and challenges of deploying AI in the power grid

This response contains insights the authors have developed through the study of AI, critical infrastructure, and specifically, the impact of AI adoption on energy security and operations of power grids in both US and Europe. Generally, AI systems can be potent but require careful attention to their unique risks and organizational changes that are needed to fully leverage applications.

Power grids need to operate as highly reliable networks[ii] as they are critical not only to communities, but to other life sustaining critical infrastructure sectors (e.g. water). Like other sectors and areas, research and development into the use of AI to further the robustness and efficiency of the power grid is active. Applications of AI systems into power grids include: (i) operational decisions about pricing, (ii) transmission, (iii) power generation, (iv) methods to prevent grid failure and improve efficiency,[iii] (v) stability assessment or grid stabilization, maintenance, or fault detection,[iv] (vi) demand and supply forecasting, (vii) energy use reduction,[v] and (viii) the interfaces of different sectors (such as energy and water).[vi] While the industry may be years away from deploying AI-based systems as a matter of course, research has demonstrated the potential to assist human decision-making in different facets of power grid operations in the

near-term.[vii] Indeed, one 2023 survey found that 220 AI companies were active in the energy sector, with the largest categories of application areas being data analytics, assets optimization, and operations and maintenance.[viii]

Even so, most AI applications (especially those based on neural networks) are opaque, and depending on their sensitivity, can introduce risks when integrated into larger and critical systems. As a result, the use of AI to boost the performance of critical infrastructure should be approached with caution.[ix] If deployed incorrectly on either a technical or organizational level, AI applications could add stress to the very sectors they aim to improve.[x] Electrical grid stakeholders already face challenges like aging infrastructure, uneven AI adoption leading to compatibility issues, and the need to balance electricity generation with consumption in real time. Research shows that increasing autonomy in systems can make them more fragile instead of improving them.[xi] For example, algorithms designed to maximize profits in competitive markets could lead to decisions that reduce redundancy in the power grid, which is necessary to maintain reliability.[xii] This could increase risks for both operators and consumers, in ways that may not be initially discernable.

There are several challenges that must be overcome to safely and effectively incorporate AI into the power grid. At a basic level, firms that have not yet digitized their equipment may struggle to comply with data requirements, in terms of quality, volume, or having readily available computation for training and inference. Even if data requirements are satisfied, the range of AI solutions are likely to be varied even for the same or related set of operations. Model-specific weaknesses that are benign for some applications may be severe for others.[xiii] This poses additional risks such as out-of-sample problems, black swan events, or other interruptions.[xiv] Beyond the technical deployment of AI, safe and effective implementation of AI will depend on organizations' ability to modify their businesses processes and policies to fully leverage new technological capabilities[xv] alongside robust, secure, AI-ready infrastructure.[xvi] Not only would such policies and infrastructure have to guide use, but given the importance of human oversight on the power grid, procedures and policies about disagreements between human and AI systems need to be outlined and developed ahead of time.[xvii]

The energy sector is not alone in having to simultaneously navigate realizing the benefits, mitigating risks, and surmounting obstacles when employing AI. However, without proper preparation, AI deployment could adversely affect different measures of energy security such as the cost of energy (i.e., affordability), energy source reliance (i.e., sustainability) and provision of energy without interruption, brownouts or blackouts (i.e., reliability).[xviii]

Recommended Focus Areas for the 2025 NAIRD

The successful implementation of AI applications to the electrical grid and broader critical infrastructure depends on the development and dissemination of best practices. Those best practices are partly informed by the research literature, but also by practitioners and in this case, power grid operators. Since the major impediment to technological adoption is user infrastructure and the capacity to absorb innovations, we recommend that the NAIRD support two separate but interrelated research efforts in the 2025 National Artificial Intelligence Research and Development Strategic Plan:

- Accelerate Advances in AI to integrate and acclimate Stakeholders (NAIRD 2023 Strategy 8) by (1) supporting research on multiagent systems that better reflect the operation of power grid, and (2) developing and evaluating tools, techniques, materials, and methods that inform stakeholders about the implementation of AI systems.
- Develop Shared Public Datasets and Environments for AI Training and Testing which Capture Stakeholder Incentives and Behaviors (NAIRD Strategy 5) by: (1) identifying the set of necessary information required from actors in the power grid to support AI systems, and (2) supporting research and development of power grid stress tests that include economic and human behavioral elements and differential AI adoption.

These two groupings of research efforts are connected: without a detailed account and representation of the incentives that individual energy sector stakeholders face, it may be difficult to understand the implications of adopting different AI features at the aggregate level. This suggests the need to identify, aggregate, and share access to the right data. Electrical grids are growing, integrating additional sources of energy and affecting other critical infrastructure sectors, such as communications and transportation systems. Therefore, the information needed to inform users about their AI applications must be explicitly identified ahead of time. However, it is infeasible to ask for total information sharing. Energy sector stakeholders, including distribution service operators and power generation companies, have market incentives that would preclude full transparency for competition. With a representative model of power grid dynamics and understanding of available data, it can then be possible to stress test the effect of differential AI adoption on a power grid given the specific risk profiles of AI applications and algorithms and their intersections with human behavior.

These two efforts can jointly inform stakeholders discussion: the United States has a variety of energy market structures and regulations, so no analysis or policy will represent a “one-size-fits-all” recommendation for AI guidelines. However, the capabilities that are generated in response to these modifications can inform stakeholders discussions and policies across the United States and help prevent catastrophic or cascading failures in critical infrastructure sectors resulting from the unique risks that are posed by AI applications. History has shown that the promise of technologies can be undercut by either ineffective or overbearing regulatory regimes. For example, nuclear incidents have led to skepticism about the safety of technology that could have reduced reliance on hydrocarbons. Similarly, de-regularization efforts in the power sector in California did not account for market incentives, resulting in the return of greater regulation.[xix]

These two points show that it is especially important to adequately account for stakeholder incentives to ensure they can meaningfully and effectively contribute to the functioning of an electrical grid utilizing the best technologies available. Because AI is relatively nascent in relation to typical technology adoption timeline, deliberate research could expedite this.

Strategy 8 Recommendations: Involving stakeholders

Adoption of new technologies into organizations requires change management, infrastructure, and the absorptive capacity to implement and maintain them. As case studies of technological adoption have shown, achieving technical performance is separate from effective utilization of them.[xx] Effective utilization will depend on power grid stakeholders' abilities to implement and utilize and maintain AI systems effectively, but will ultimately be constrained by the systems they operate within. To better improve the nation's understanding of AI as applied to the power grid and inform stakeholders, we propose two actions.

First, while research into the different dimensions of risks and failures of AI is ongoing, research to include scenarios where AI is distributed across a system like power grid stakeholders have to contend with is needed.[xxi] In the power grid, the understanding is that a smart grid system would involve decentralized intelligence at multiple points in the network, and so it would constitute a multi-agent system (MAS).[xxii] Even though MASs are designed to navigate this environment, they still have possible failure modes that can arise from miscoordination (where agents cannot coordinate successfully), conflict (where agents' or their principals' incentives are not aligned), or collusion (where agents inadvertently take advantage of other actors to advance their interests).[xxiii] There are multiple risk factors that can contribute to these failures, and in particular we are concerned with those stemming from a lack of adequate information, training to different objectives that are sensitive to incentives but not common welfare, and user-related issues (stemming either from lack of understanding or trust). Understanding how these risks manifest in the power grid will need to be informed by engagements with power grid stakeholders.

Second, integrating AI within the energy sector in a trustworthy manner is more complicated than with other technologies, and efforts to understand how power grid stakeholders react to AI systems and to improve their understanding is vital to achieve safe and effective integration. As noted by Rogers et al. (2025), "the autonomous behavior expected of the smart grid, its distributed nature, and the existence of multiple stakeholders each with their own incentives and interests, challenges existing engineering approaches."[xxiv] Directly engaging stakeholders will not only better inform systems and educate potential users, but with their inclusion in the development of those technologies, sector stakeholders would have the opportunity to come to their own rules and procedures regarding its use.[xxv]

The effect of both of these efforts is to draw closer connections between researchers and power grid stakeholders to the operational implications of AI systems for power distribution.[xxvi] AI systems are already employed in the sector in the form of decision-analysis tools of varying architectures, but as the electrical grid increases in complexity with management of various

energy sources, those energy sources' own networks, the inclusion of "prosumers" of energy and "smart grid" systems within an aging infrastructure with analog components, there is both a need and a risk to integrating greater autonomous systems to improve efficiency.[xxvii] This risk is much greater if researchers are not sensitive to the limitations and operational requirements (e.g. demand profiles, age of infrastructure, revenue models) of energy stakeholders, and stakeholders are not sufficiently informed of the risks associated with increasing autonomy in electrical grids. Without bridging this information gap, it will be difficult, and potentially legally fraught, to integrate AI systems within the electrical sector.

These two threads complement one another: the more informed researchers are of these methods, their risk profiles, and appropriate use cases, the less work stakeholders must do to operationalize these technologies; the more informed stakeholders are, the better information they can provide to researchers.

Strategy 5 Recommendations: Representative testbeds for the electrical sector

As mentioned above, AI applications require accurate and plentiful data to perform successfully, but lack of accurate or timely information can make algorithms' performance suffer. As a matter of practice, power grid stakeholders will need to constantly update and ensure their AI models are fresh, demonstrating the need for improved data curation strategies. In particular, stakeholders will need to share data even though there may be market incentives to keep data private. Due to the interdependent nature of the electrical grid, events at one point in the grid can have implications at the other. Without adequate accounting of seemingly distinct events, there is a risk that failure can occur. This is especially the case since different critical infrastructure sectors are becoming increasingly intertwined. For instance, communication systems rely on and support the electrical grid, information systems rely on the water system for cooling, and the electrical grid may separately rely on hydroelectric power.

Additionally, assuming each critical infrastructure sector has some incentives to increase autonomation, their model architecture and implications for maintaining and use of their interconnections will also need to be considered. For example, the interaction of multiple, diverse agent sets could lead to complications, such as positive feedback loops between an expert system controlling power generation and another seeking maximum reliability. This means that data about operations and autonomous systems use need to be collected to prevent emergent issues in the operation of the electrical sector.

These data alone are not sufficient, however. Also needed are accurate testbeds for specialized sectors to model not only the physics of their operation, but also the incentives and social behavior that can become relevant, such as cognitive overload, revenue requirements, transmission failures, shifting data distributions, social behavior, and energy spikes.[xxviii] There are already some efforts at simulating and modeling electrical grids that include the implications of

AI agents, but these should be extended to consider the co-existence of non-AI enabled agents and AI enabled agents.[xxix] Without a realistic understanding of how the electrical grid will contend with a system of AI-enabled stakeholders, foreseeing and mitigating the risks of cascading failures will be a problem. To this end, the 2025 NAIRD should support research that:

- Identifies the minimum set of information required from actors within complex or multiagent systems to still function while protecting their privacy and other interests. In the power grid to support AI systems. This is inclusive of both the type of information that must be shared within a smart grid to ensure that its autonomous or semi-autonomous elements can function and studies of power grid stakeholders' capabilities to develop and maintain the data infrastructure to provide that information.
- Develops power grid stress tests that include economic and human behavioral elements and differential AI adoption. These research efforts should be able to simulate scenarios that include social and human behavioral elements and differential adoption, as well as supporting AI systems that are not themselves critical operations, like maintenance.

The outputs of these activities will provide researchers and stakeholders with a better understanding of the implications of autonomous systems in their operating environments. These two threads will also end up supporting the two threads involved with stakeholder outreach by primarily providing the type of actionable research and analysis that can help formulate agreements, policies, and understandings within an electrical grid.

---

[i] Hammond, Lewis, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. *Multi-agent Risks from Advanced AI,* Cooperative AI Foundation, Technical Report #1. As of May 22, 2025: https://arxiv.org/abs/2502.14143

[ii] Aaron Clark-Ginsberg, David DeSmet, Ismael Arciniegas Rueda, Ryan Hagen, Brian Hayduk. "Disaster risk creation and cascading disasters within large technological systems: COVID-19 and the 2021 Texas blackouts." *Journal of Contingencies and Criss Management* Vol. 29 No. 4 December 2021: 445-449

[iii] Wei, Mingkui, Zhuo Lu, and Wenye Wang, "On Characterizing Information Dissemination During City-Wide Cascading Failures in Smart Grid," *IEEE Systems Journal*, Vol. 12, No. 4, December 2018, pp. 3404–3413. As of May 19, 2025: https://ieeexplore.ieee.org/document/8098607/; Bose 2017; Park, Jiyoung, and Dongheon Kang, "Artificial Intelligence and Smart Technologies in Safety Management: A Comprehensive Analysis Across Multiple Industries," *Applied Sciences*, Vol. 14, No. 24, December 2024, Article 11934. As of May 19, 2025: https://www.mdpi.com/2076-3417/14/24/11934

[iv] Shen, Zhiwei, Felipe Arraño-Vargas, and Georgios Konstantinou, "Virtual Testbed for Development and Evaluation of Power System Digital Twins and Their Applications," *Sustainable Energy, Grids and Networks*, Vol. 38, June 2024, Article 101331. As of May 19, 2025: https://linkinghub.elsevier.com/retrieve/pii/S2352467724000602; Ahmad, Ishtiaq, Jawad Haider Kazmi, Mohsin Shahzad, Peter Palensky, and Wolfgang Gawlik, "Co-Simulation Framework Based on Power System,

AI, and Communication Tools for Evaluating Smart Grid Applications," in *2015 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, Bangkok, Thailand, IEEE, November 2015, pp. 1–6. As of May 19, 2025: http://ieeexplore.ieee.org/document/7387092/; http://arxiv.org/abs/2406.05003
Rogers, Alex, Sarvapali Ramchurn, and Nicholas Jennings, "Delivering the Smart Grid: Challenges for Autonomous Agents and Multi-Agent Systems Research," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26, No. 1, September 2021, pp. 2166–2172. As of May 19, 2025: https://ojs.aaai.org/index.php/AAAI/article/view/8445

[v] McMillan, Lauren, and Liz Varga, "A Review of the Use of Artificial Intelligence Methods in Infrastructure Systems," *Engineering Applications of Artificial Intelligence*, Vol. 116, November 2022, Article 105472. As of May 19, 2025: https://linkinghub.elsevier.com/retrieve/pii/S0952197622004626; Bedi and Toshniwal 2019.; Lin, Zi, Xiaolei Liu, and Maurizio Collu, "Wind Power Prediction Based on High-Frequency SCADA Data Along with Isolation Forest and Deep Learning Neural Networks," *International Journal of Electrical Power & Energy Systems*, Vol. 118, June 2020, Article 105835. As of May 19, 2025: https://linkinghub.elsevier.com/retrieve/pii/S0142061519332491

[vi] Zaidi, Syed Mohammed Arshad, Varun Chandola, Melissa R. Allen, Jibonananda Sanyal, Robert N. Stewart, Budhendra L. Bhaduri, and Ryan A. McManamay, "Machine Learning for Energy-Water Nexus: Challenges and Opportunities," *Big Earth Data*, Vol. 2, No. 3, July 2018, pp. 228–267. As of May 19, 2025: https://www.ingentaconnect.com/content/10.1080/20964471.2018.1526057

[vii] van Soest, Henri and Ismael Arciniegas Rueda, Hye Min Park, Bryden Spurling, Austin Wyatt, Harper Fine, Joshua Steier, and Melusine Lebert, *The use of AI for improving energy security: Exploring the risks and opportunities of the deployment of AI applications in the electricity system.* RAND Report RR-A2907-1. As of August 1, 2024: https://www.rand.org/pubs/research_reports/RRA2907-1.html.

[viii] Franki, Vladimir, Darin Majnarić, and Alfredo Višković, "A Comprehensive Review of Artificial Intelligence (AI) Companies in the Power Sector," *Energies*, Vol. 16, No. 3, 2023, Article 1077.

[ix] Gerstein Daniel, Erin Leidy. *Emerging Technology and Risk Analysis: Artificial Intelligence and Critical Infrastructure.* RAND Report RRA-2873-1. As of June 21, 2024: https://www.rand.org/pubs/research_reports/RRA2873-1.html

[x] Ammanath, Beena, "How to Manage AI's Energy Demand – Today, Tomorrow, and Beyond," World Economic Forum, April 2024. As of August 1, 2024: https://www.weforum.org/agenda/2024/04/how-to-manage-ais-energy-demand-today-tomorrow-and-in-the-future/; Saul, Josh, Leonardo Nicoletti, Saritha Rai, Dina Bass, Ian King, Jennifer Duggan, "AI Is Already Wrecking Havoc on Global Power Systems," Bloomberg, 2024. As of August 1, 2024: https://www.bloomberg.com/graphics/2024-ai-data-centers-power-grids

[xi] Druce, Jeff, James Niehaus, Vanessa Moody, David Jensen, and Michael L. Littman, "Brittle AI, Causal Confusion, and Bad Mental Models: Challenges and Successes in the XAI Program," arXiv, June 2021. As of May 9, 2025: http://arxiv.org/abs/2106.05506

[xii] Clark-Ginsberg Aaron, Arciniegas Rueda, Ismael, Jonathon Monken, Jay Liu, Hong Cheng, "Maintaining critical infrastructure resilience to natural hazards during the COVID-19 pandemic: hurricane preparations by US energy companies," *Journal of Infrastructure Preservation and Resilience*, Vol. 1. No. 1, 2020

[xiii] Bedi & Toshniwal 2019.

[xiv] Burton, Simon, Ibrahim Habli, Tom Lawton, John McDermid, Phillip Morgan, and Zoe Porter, "Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective," *Artificial Intelligence*, Vol. 279, February 2020, Article 103201. As of May 19, 2025: https://linkinghub.elsevier.com/retrieve/pii/S0004370219301109; Dev, Jayati, Nuray Baltaci Akhuseyinoglu, Golam Kayas, Bahman Rashidi, and Vaibhav Garg, "Building Guardrails in AI Systems with Threat Modeling," *Digital Government: Research and Practice*, Vol. 6, No. 1, March 2025, pp. 1–18. As of May 19, 2025: https://dl.acm.org/doi/10.1145/3674845

[xv] McMillan & Varga 2022.

[xvi] Smadi, Abdallah A., Babatunde Tobi Ajao, Brian K. Johnson, Hangtian Lei, Yacine Chakhchoukh, and Qasem Abu Al-Haija, "A Comprehensive Survey on Cyber-Physical Smart Grid Testbed Architectures: Requirements

and Challenges," *Electronics*, Vol. 10, No. 9, April 2021, Article 1043. As of May 19, 2025: https://www.mdpi.com/2079-9292/10/9/1043

[xvii] Subías-Beltrán, Paula, Oriol Pujol, and Itziar de Lecuona, "Safeguarding Autonomy: A Focus on Machine Learning Decision Systems", arXiv, March 2025. As of May 19, 2025: http://arxiv.org/abs/2503.22023

[xviii] van Soest et al. (2024).

[xix] Joskow P, "California Electricity Crisis," *National Bureau of Economic Research Digest*, Working Paper 8442, 2001

[xx] Arthur, W. Brian. *The nature of technology: What it is and how it evolves*. Simon and Schuster, 2009..

[xxi] Ji, Jiaming, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao, "AI Alignment: A Comprehensive Survey", May 2024. As of May 5, 2025: https://arxiv.org/abs/2310.19852v5; Hammond et al. 2025.

[xxii] van Soest et al. 2024

[xxiii] Hammond et al. 2025.

[xxiv] Rogers, Alex et al. 2021.

[xxv] Subías-Beltrán et al. 2025.

[xxvi] van Soest et al. 2024

[xxvii] van Soest et al. 2024; Kondor, Daniel, Valerie Hafez, Sudhang Shankar, Rania Wazir, and Fariba Karimi, "Complex Systems Perspective in Assessing Risks in Artificial Intelligence," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 382, No. 2285, December 2024, Article 20240109. As of May 19, 2025: https://royalsocietypublishing.org/doi/10.1098/rsta.2024.0109

[xxviii] Bainbridge, Lisanne, "Ironies of Automation," *Automatica*, Vol. 19, No. 6, November 1983, 775-779.

[xxix] Arciniegas Rueda et al. 2024