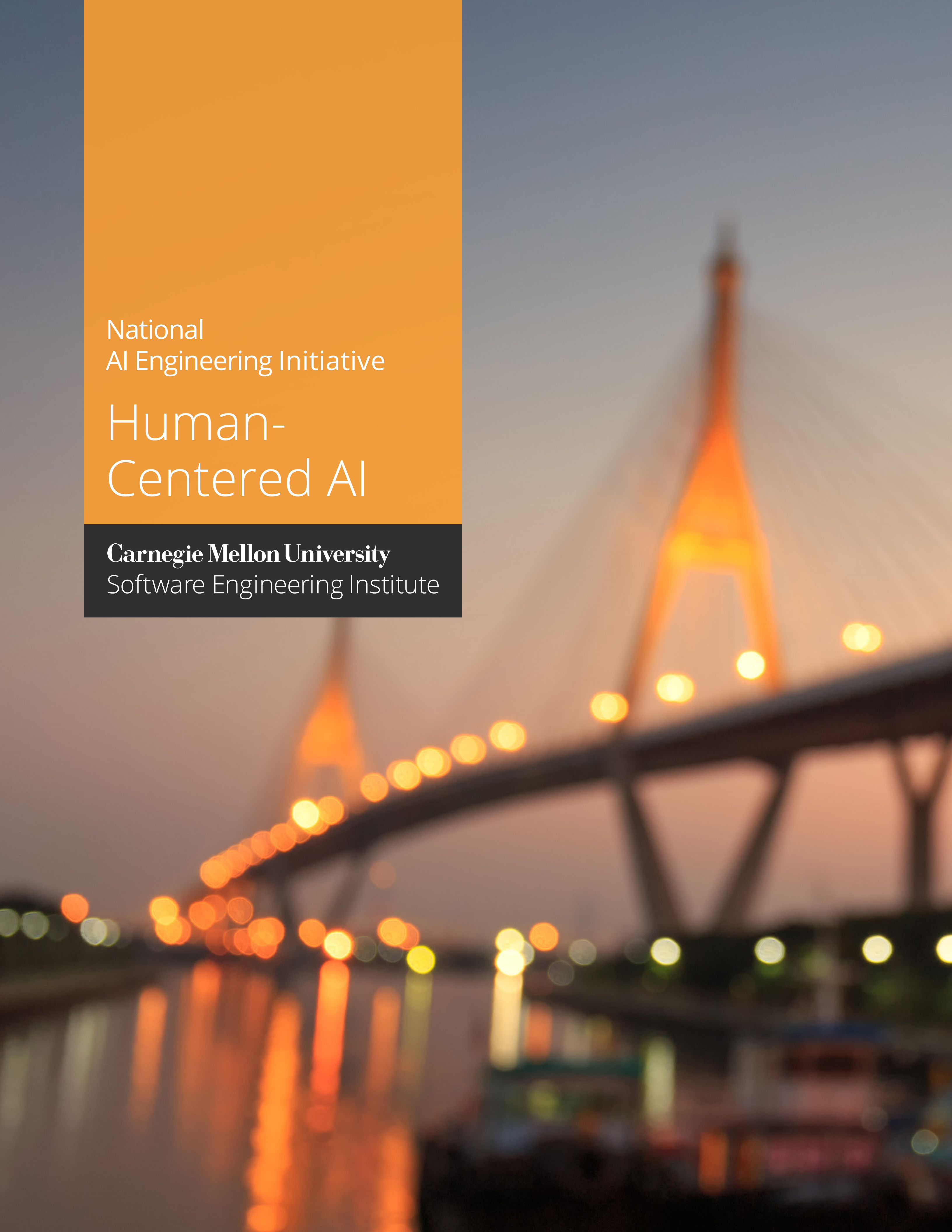

National
AI Engineering Initiative

Human-Centered AI

Carnegie Mellon University
Software Engineering Institute

Human-Centered AI

Contributors:
Hollen Barmer, Rachel Dzombak,
Matt Gaston, Jay Palat, Frank Redner,
Carol Smith, Tanisha Smith

Acknowledgements:
The team thanks Charles Holland, Ipek Ozkaya, and Joshua Poore for their review; and Nancy Ott for editing the paper.

The Pillars of AI Engineering

1. **Human-Centered**
2. Robust and Secure
3. Scalable

The emergent discipline of AI Engineering is focused on three pillars: human-centered AI, robust and secure AI, and scalable AI.

To learn more about AI Engineering, visit **sei.cmu.edu/our-work/artificial-intelligence-engineering**.

Human-Centered AI

Many artificial intelligence (AI) systems in use today were originally designed for low-stakes situations, or situations with a low cost of failure. If your streaming service recommends a movie that isn't to your taste, or if your device misinterprets a request to play the Beach Boys, and the Beastie Boys starts blaring instead, you may be temporarily frustrated or laugh off the experience—but you can quickly move on. What about the design and development of AI systems when the stakes are far higher? For example, how do we design and develop AI systems that have the potential to dictate whether or not someone receives a mortgage for a home purchase, or where to send first responders during a rapidly spreading wildfire? In high-stakes scenarios, humans and AI systems must work together in human-machine teams with trust, ethical integrity, and mutual understanding of a shared goal.

Human-centered AI systems **are designed to work with, and for, people.** The real-world, operational environments in which teams and organizations hope to embed AI systems are dynamic and complex, with ambiguous decisions spaces. Humans—their behaviors, influences, decisions, and actions—are central to such environments [1]. As the desire to use AI systems grows, human-centered engineering principles are critical to guide system development toward effective implementation and minimize unintended consequences.

The development and deployment of human-centered AI systems is challenging due to a variety of factors: (1) the state of the art in AI technology is continuously evolving, (2) the intended level of interdependence between humans and machines leads to trust and transparency challenges, (3) AI systems will be deployed alongside humans in unfamiliar and unpredictable operational contexts, and (4) guidelines and heuristics are scarce for understanding and implementing the level of oversight needed to create and maintain ethical AI systems. While bodies of work address some of these factors individually, viewing them together presents a host of new challenges.

We identify three specific areas of focus to advance human-centered AI:

- **Designers and systems must understand the context of use and sense changes over time:** Successful AI Engineering depends on the team's ability to identify and articulate the desired system outcome and understand human and contextual factors affecting the outcome. The system itself must be able to learn when shifts in context have occurred. What are the best ways to maintain clarity around operational intent and mechanisms for adapting and evolving systems based on dynamic contexts and user needs?
- **Development of tools, processes, and practices to scope and facilitate human-machine teaming:** Implementation of AI systems entails high levels of interdependence between human and machine. Adoption of AI systems requires the primary users to interact with and understand systems, gaining appropriate levels of trust. Every AI system needs to be designed to recognize boundaries and unfamiliar scenarios, and to provide transparency regarding its limitations.
- **Methods, mechanisms, and mindsets to engage in critical oversight:** AI systems learn through data and observations, rather than being explicitly programmed for a deterministic outcome. Critical and reflective oversight by organizations, teams, and individuals that create and use AI systems is needed to uphold ethical principles and proactively consider the risks of bias, misuse, abuse, and unintended consequences through design, development, and ongoing deployment.

For each area, we identify ongoing work as well and challenges and opportunities in developing and deploying AI systems with confidence.

Understanding Context of Use

Traditional software systems behave according to a carefully crafted set of rules that direct the system to produce an outcome. To create a machine learning (ML) system, which is a common type of AI system, developers select data, a learning algorithm, and an objective function (typically mathematical descriptions of the desired value outcome and purpose) in order to produce a model that can make decisions automatically in the applied context. When all goes well, the learned model provides the desired value outcome (recommendations, pattern identification, etc.). Given the role of the model, one of the most powerful influences on an ML system is how the team defines the objective function or system purpose.

Unfortunately for AI system developers, accurately capturing what humans value as outcomes is a significant challenge, and those values are likely to change over time. The fields of human-centered design and human-computer interaction (HCI) have long established the difficulty for humans to articulate the decisions they make and the motivators that guide behavior.

"What users say that they do, and what they actually do, can vary significantly, especially when they are removed from the relevant context. People have a hard time reflecting on, even noticing, the ordinary and habitual activities of everyday life. They report with conviction on what they believe to be true, not necessarily on what is true." [2]

As a result, aligning system values and actions with the objective of the human continues to be harder than we think [3], [4]. While techniques including inverse reinforcement learning [5] and mechanism design [6] are attempting to address the value-alignment challenge, many challenges remain. Developers need methods to determine appropriateness of data and observed behavior for use in an AI system as well as greater awareness of the frequent mismatch between humans' behavior and their true intentions. Other critical gaps include ways of measuring how system behaviors reflect humans' understanding of overall functionality and intentions, how to maintain human-machine clarity in operational intent, and mechanisms for adapting and evolving systems based on dynamic contexts and user needs. Meeting these challenges can contribute to the ability of systems to learn from humans, to improve performance over time, or to recognize when systems are no longer appropriate due to a shift in user needs.

Human-centered AI requires understanding factors such as the primary user's needs, the context of use, and the user's mental model of the work. Research methods from the fields of design and HCI that are rooted in ethnography provide ways to understand why humans take actions and make the decisions they do, as well as the underlying beliefs that guide their behaviors [7].

"The main task of [user research] is not only to watch, but also to decode human experience—to move from unstructured observations to discover the underlying meanings behind behavior; ...to deduce logical implications for strategic decisions." [8]

A key need for the AI Engineering field is to develop processes, practices, and frameworks that can integrate human insights surfaced during user research into the correct reward function for the AI system [9]. At the same time, we need human-centered research on strategies for AI systems to approximate humans' actual objectives (as opposed to stated objectives), sense over time how users' objectives are evolving, and continuously improve approximations of the desired outcomes. We need to understand user context and needs, and then translate and operationalize that user research into the design and operations of the system itself.

Currently, human-centered research is conducted by people across a broad set of job titles, including human-machine interaction and human-computer interaction professionals; user-centered researchers and designers; user experience (UX) researchers and designers; interaction designers; design researchers; digital anthropologists; ethnographers; information architects; and cognitive psychologists. Teams can ensure they are knowledgeable about the users of the system by conducting human-centered research (i.e., observations, interviews, and other methods drawn from ethnography) with the intended primary users prior to beginning the development efforts. In doing so, teams are more likely to design and develop AI systems that match users' needs. Such human-centered research is not an insignificant effort, but the alternative is building a system that is frustrating to use, or worse, doesn't match the initial challenge and misrepresents users' needs [8].

Understanding context is a dynamic process that cannot be completed at a single point in time. The developers of AI systems need to ensure capacity exists to continue user research over time, both by humans and by the AI system itself as it captures and learns from user behavior in context. It is also critical to note that AI is often designed as part of a broader system. In understanding context, there is a need to toggle focus between the immediate users' behaviors and the system in which the user exists. Research is needed on opportunities and mechanisms to collect information and sense or infer user intention from other parts of larger systems.

Designing for Human-Machine Teaming

Human-centered AI requires a significant level of attention and intention to design systems "from the outset to team with human capabilities, providing assistance where human intelligence has limits and leveraging that intelligence where it is uniquely powerful" [10]. The complex and uncertain situations inherent to national security require the dynamic activity of teaming, where team members and embedded systems must make rapid decisions [11]. As such, understanding and implementing factors that support effective human-machine teams is a significant research area for AI Engineering.

Human teams have a common purpose, a need for interdependence to accomplish goals, and mutual accountability for the results of the goal [12]. Defining the common purpose of human-machine teams relates to understanding the system context and system objectives. Interdependence between humans and machines could involve the capability to observe the other's state, to bidirectionally share information, to request help, or to flag an anomaly event [10], [13]. AI systems are often designed as components of larger systems. Interdependence exists then on two levels: between the human and machine as well as between the human-machine team and the larger network of connected teams and systems. Rather than viewing the human-machine team in isolation, we need research on what it practically means to adopt a whole system's perspective and what tools can help to ensure focus on integration of components into systems. Open questions to examine both within a single human-machine team and across networks of human-teams include (1) how to facilitate transparent information sharing and responsibilities, (2) how to align purpose and strategy, (3) how to enable adaptability in response to a shifting environment, and (4) methods and guidance for decentralized decision making.

In contrast to human teams, the human-machine team is not an equal one when it comes to accountability. In line with the Department of Defense (DoD) Ethical Principles for Artificial Intelligence [1], [14] and other sets of technical ethics [15], humans must have the ultimate responsibility for decisions and outcomes. AI systems do not (and must not) have rights or responsibilities. This means that humans need to be able to understand the AI sufficiently well to predict system behavior [16]. AI developers must avoid creating inscrutable systems by using tools such as model cards to describe the model in detail [17]. While "modern [AI and ML] techniques were born and bred for low-stakes decisions such as online advertising and web search where individual decisions do not deeply affect human lives" [18], AI Engineering in high-consequence contexts necessitates system transparency, predictability, and interpretability.

Interactions with AI systems require transparency with regard to when the system is in control, and when control transitions to the human partner. Unambiguous cues need to be appropriate given the context and urgency of those transitions. The human-machine team needs to have explicitly defined responsibilities. In some cases, there may be a need for explicit transition, similar to a circuit breaker lock-out that keeps electrical workers safe during system maintenance. At present, strategies to foster trust center on information sharing. Providing the system's purpose, limitations, data sources, and biases known to exist within the data, the algorithms, and across the system can help users and operators to build confidence and trust in the system. However, the way that information is provided must vary depending on the intended users and their context of their use. For example, a flight test engineer expects to have visibility and awareness of appropriate systems; over-simplification of the data in the pursuit of ease of use may cause them to miss important information, which erodes trust. Research and tools are needed to understand the level of specificity and classes of information to meet the needs of different user groups.

Another research challenge is to devise strategies for explaining significant decisions made by AI systems, for assessing the appropriateness of complex models in high-stakes scenarios, and for ensuring humans' abilities to override or reverse AI decisions. Furthermore, tools, processes, and frameworks are needed to guide the use of AI for irreversible decisions that affect a person's life, quality of life, health, or reputation [19]. Users' willingness to adopt an AI system depends on their perception of the system being accountable and transparent. Open research questions include how to build appropriate trust with the AI system, what appropriate trust means in a particular context, how we can design AI to recognize when system limits are surpassed or anomalies are detected, and how we train a workforce to engage with AI.

Engaging in Critical Oversight

AI systems resemble the contexts in which they are intended to be used—they are filled with complex interrelationships and many moving parts, and they evolve with time.

Implementation of AI systems requires a continuous level of oversight by humans who ask, "What are we doing? Why are we doing it, and for whom?"

At present, researchers are exploring the work of oversight and related issues, including regulatory policy and data ownership, through subfields such as responsible AI [20], [21], algorithmic fairness [22], ethical AI [23], and explainable AI [24]. Increasingly, organizations are adopting sets of technical ethics as key mechanisms for accountability, risk reduction, and the ultimate success of AI systems [25].

Risk can be present in the goal and design of AI systems (e.g., unintended and purposeful bias in data, algorithm selection, and training), and it can come from misuse and abuse of the system. The opportunities for risk embedded in systems mean that critical oversight must be integrated into all parts of the AI system life cycle. Engaging with ethics requires the entire team to have conversations that are sometimes difficult, because they can alter the intended work and require deep introspection. Establishing the psychological safety among team members to engage in reflective work takes time and intentional effort within organizations [26]. Preventing unintended and potentially harmful consequences requires careful vetting and critique of the work in the early stages of planning (before creating the system), as the system is built, and throughout its use. Prior to algorithm selection, AI engineers should fully explore the existing data to ascertain the inherent bias and the amount of variance within the data [27]. Documenting the data's motivation, composition, collection process, recommended uses, and so on is necessary to prioritize transparency and accountability [27]. Research into best practices should examine possible unintended consequences and define strategies for documenting design team decisions. Empowering diverse teams to do critical oversight work, and to authentically engage with ethics, will help to build ethical and de-risked AI systems.

Achieving success for AI systems depends on the ability to create and implement processes and systems that monitor and verify functionality and effectiveness throughout the life of the system [17]. Clarity of the needed outcome for the human is required for effective evaluation. AI Engineering requires a shift from evaluating only system outputs (e.g., precision, accuracy, recall) to evaluating system outcomes (i.e., did the system provide the intended benefit? did the system respond appropriately given the situation?). Defining standard methods and processes for evaluating system outcomes is a significant need to advance the discipline [28].

As teams maintain AI systems—and as those systems evolve over time—teams should frequently test them with users to determine if they meet desired outcomes and align with stakeholder needs. Established human-centered evaluation methods such as usability studies can support testing how users interact with systems. Open questions exist for how to create and maintain the capacity for both humans and machines to conduct AI system monitoring and share easily interpretable results. Humans must be able to determine if the system is responding as expected, or if the system has been compromised in some way, and the source of the issue (e.g., data shifts, new bias, misuse, and/or abuse). Discussion exists around who is accountable for oversight and the differing (or overlapping) roles of development teams and third-party auditors in oversight. To build upon the interdependence between humans and machines, AI systems can be designed to flag shifts in behavior over time, as well as provide information and transparency to support partnership in critical oversight.

Human-Centered AI

Successful development and deployment of human-centered AI that incorporates systems context, drives human-machine teaming, and is guided by critical oversight requires a workforce that understands AI systems and their limits [29]. Adopting technology ethics is merely an initial step towards making AI systems robust and secure. AI systems are surrounded by risks—from unintentional misuse and from adversaries—and "making sure that the teams building these systems are able to imagine and then mitigate issues is profoundly important" [30].

There is broad acknowledgement that the DoD needs to pursue multiple strategies to cultivate a diverse, capable, and AI-fluent workforce. The National Security Commission on AI recommends that the DoD expand the digital workforce by creating new structures to house highly skilled specialists, increasing recruitment and training efforts, and shifting organizational culture to ensure people can engage in impactful work [31]. As AI Engineering focuses on whole-system design and not just on capability development, the AI Engineering workforce must be prepared with the mindset and skillset needed to support the entire AI life cycle. In addition to machine learning and data science specialists, the AI Engineering workforce will need system architects, product managers, social scientists, ethicists, and more.

The complex challenges brought forth by adoption and integration of human-centered AI systems will demand teaming across roles and organizational boundaries. Teaming in organizations is the "engine of organizational learning...a way of working that brings people together to generate new ideas, find answers and solve problems. But people have to learn to team; it doesn't come naturally" [11]. To achieve mission outcomes, AI teams throughout the DoD will need to draw upon available diversity, including different ways of seeing problems informed by team members' cultural backgrounds and experiences as well as different ways of solving problems. Enabling AI Engineering teams to effectively collaborate will require shifts in how DoD organizations are structured and incentivized.

Overall, the opportunity for development of human-centered AI within the DoD is significant. Creating human-centered AI systems requires a continuous focus on the humans who are creating, using, and affected by the system. By conducting research to understand the users, AI system designers and developers can begin to identify the benefits (and risks) of AI systems, as well as other contextual factors that influence mission goals. Through research, teams that are developing AI systems can not only identify the needs of the people who will be the primary users, but also forge a path to building an honest and respectful AI system for teaming with those users.

Enabling people to be more effective and productive in their work, to find information they would otherwise be unaware of, and to draw connections between data are just a few of the many advantages that human-centered AI systems will provide.

References

[1] Defense Innovation Board, "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense," 2019. [Online]. Available: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

[2] A. Van Pelt and J. Hey, "Using TRIZ and human-centered design for consumer product development," *Procedia Eng.*, vol. 9, pp. 688–693, 2011, doi: 10.1016/j.proeng.2011.03.156.

[3] P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery, "Human-aligned artificial intelligence is a multiobjective problem," *Ethics Inf. Technol.*, vol. 20, no. 1, pp. 27–40, Mar. 2018, doi: 10.1007/s10676-017-9440-6.

[4] M. Mitchell, "Why AI is Harder Than We Think," *ArXiv210412871 Cs*, Apr. 2021, Accessed: May 14, 2021. [Online]. Available: http://arxiv.org/abs/2104.12871

[5] M. Palaniappan, D. Malik, D. Hadfield-Menell, A. Dragan, and S. Russell, "Efficient Cooperative Inverse Reinforcement Learning," *Proc ICML Workshop Reliab. Mach. Learn. Wild*, pp. 1–5, 2017.

[6] R. Abebe and K. Goldner, "Mechanism Design for Social Good," *AI Matters*, vol. 4, no. 3, pp. 27–34, 2018, doi: 10.1145/3284751.328476.

[7] S. L. Beckman and M. Barry, "Innovation as a Learning Process: Embedding Design Thinking," *Calif. Manage. Rev.*, vol. 50, no. 1, pp. 25–56, Oct. 2007, doi: 10.2307/41166415.

[8] H. Mariampolski, "The power of ethnography," *Mark. Res. Soc. J.*, vol. 41, no. 1, pp. 1–12, 1999.

[9] R. Gibbons, "Incentives in Organizations," *J. Econ. Perspect.*, vol. 12, no. 4, pp. 115–132, 1998.

[10] M. Johnson and A. Vera, "No AI Is an Island: The Case for Teaming Intelligence," *AI Mag.*, vol. 40, no. 1, pp. 16–28, Mar. 2019, doi: 10.1609/aimag.v40i1.2842.

[11] A. C. Edmondson, *Teaming: How organizations learn, innovate, and compete in the knowledge economy*. John Wiley & Sons, 2012.

[12] J. R. Katzenbach and D. K. Smith, *The Discipline of Teams*. Harvard Business Press, 2008.

[13] J. van Diggelen, M. Neerincx, M. Peeters, and J. M. Schraagen, "Developing Effective and Resilient Human-Agent Teamwork Using Team Design Patterns," *IEEE Intell. Syst.*, vol. 34, no. 2, pp. 15–24, Mar. 2019, doi: 10.1109/MIS.2018.2886671.

[14] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI—Two Frameworks for Ethical Design Practice," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 34–47, Mar. 2020, doi: 10.1109/TTS.2020.2974991.

[15] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI.," Berkman Klein Center for Internet & Society, 2020. [Online]. Available: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

[16] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," ArXiv Prepr. *ArXiv181204608*, 2018.

[17] M. Mitchell et al., "Model Cards for Model Reporting," *Proc. Conf. Fairness Account. Transpar.*, pp. 220–229, Jan. 2019, doi: 10.1145/3287560.3287596.

[18] C. Rudin and J. Radin, "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition," *Harv. Data Sci. Rev.*, vol. 1, no. 2, Nov. 2019, doi: 10.1162/99608f92.5a8a3a3d.

[19] C. J. Smith, "Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development," in AAAI FSS-19: *Artificial Intelligence in Government and Public Sector*, Arlington, Virginia, 2019, p. 6. [Online]. Available: https://arxiv.org/pdf/1910.03515.pdf

[20] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[21] N. Sambasivan and J. Holbrook, "Toward Responsible AI for the Next Billion Users," *Interactions*, vol. 26, no. 1, pp. 68–71, 2018.

[22] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, "Algorithmic Fairness," *AEA Pap. Proc.*, vol. 108, pp. 22–27, May 2018, doi: 10.1257/pandp.20181018.

[23] J. Torresen, "A Review of Future and Ethical Perspectives of Robotics and AI," *Front. Robot. AI*, vol. 4, p. 75, Jan. 2018, doi: 10.3389/frobt.2017.00075.

[24] G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer, 2018, pp. 19–36.

[25] Dept. of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," Feb. 2020. Accessed: Mar. 05, 2021. [Online]. Available: https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod%E2%80%90adopts%E2%80%90%20ethical%E2%80%90principles%E2%80%90for%E2%80%90artificial%E2%80%90intelligence/

[26] A. Edmondson, “Psychological Safety and Learning Behavior in Work Teams,” *Adm. Sci. Q.*, vol. 44, no. 2, p. 350, Jun. 1999, doi: 10.2307/2666999.

[27] T. Gebru et al., “Datasheets for datasets,” *ArXiv Prepr. ArXiv180309010*, 2018.

[28] J. Hernández-Orallo, “Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement,” *Artif. Intell. Rev.*, vol. 48, no. 3, pp. 397–447, Oct. 2017, doi: 10.1007/s10462-016-9505-7.

[29] N. D. Bastian, “Building the Army’s Artificial Intelligence Workforce,” *Cyber Def. Rev.*, vol. 5, no. 5, pp. 59–64, 2020.

[30] C. Smith, “Creating a Curious, Ethical, and Diverse AI Workforce,” *War on the Rocks*, Mar. 05, 2020. Accessed: Mar. 08, 2021. [Online]. Available: https://warontherocks.com/2020/03/creating-a-curious-ethical-and-diverse-ai-workforce/

[31] National Security Commission on AI, “Final Report,” 2021. [Online]. Available: https://www.nscai.gov/2021-final-report/

About the SEI

The Software Engineering Institute is a federally funded research and development center (FFRDC) that works with defense and government organizations, industry, and academia to advance the state of the art in software engineering and cybersecurity to benefit the public interest. Part of Carnegie Mellon University, the SEI is a national resource in pioneering emerging technologies, cybersecurity, software acquisition, and software lifecycle assurance.

Contact Us

CARNEGIE MELLON UNIVERSITY
SOFTWARE ENGINEERING INSTITUTE
4500 FIFTH AVENUE; PITTSBURGH, PA 15213-2612

sei.cmu.edu
412.268.5800 | 888.201.4479
info@sei.cmu.edu