Input for the US Action Plan to Advance America's AI Leadership.

The Human Factors and Ergonomics Society is pleased to provide the following input to help advance America's leadership in the development and implementation of artificial intelligence (AI) technology. The successful adoption of AI is highly dependent on its ability to gain the trust of human users and their ability to use it to successfully support their decision making and performance of tasks in various fields of application.

The development of technologies that are user-centered is not only good for people, it is also a highly successful business approach. A cross-industry study by McKinsey & Company [1] found that companies with a strong user-centered design approach to their products out-performed their industry competitors by as much as 2 to 1. These companies repeatedly create successful products by adopting an inter-disciplinary approach to design that centers around understanding user needs, designing products that emphasize usability, and prioritizing a systematic approach to user testing. This is what the human factors and ergonomics profession has been successfully doing for over 80 years.

The field of Human Factors science has conducted extensive research on how people interact with AI and other automated systems for over 40 years. This research base has created a significant trove of information on how to design AI to work effectively with people and to avoid the types of errors that can undermine people's confidence in AI. Based on this research base, we offer the following Guardrails for AI Development to advance America's leadership in the field.

Recommended Guardrails for AI

AI Shall Provide Explicit Labeling

Systems that use AI to perform tasks or provide recommendations must be labeled as being provided by a computational system. If the system provides or integrates information from data sources (e.g., web sites), the source of information should be specifically provided so that users can determine its reliability or trustworthiness[2]. The provenance of information should be transparent (e.g., AI system, peer reviewed reference, individual opinion)[3]. This guardrail is needed to support the ability of people to make appropriate "opt out" decisions with respect to AI information.

Recommendation 1: All AI outputs (e.g., generative language, videos, audio, images, recommendations) must be labeled as the product of a computer system and the source of information used for its outputs must be specifically identified.

AI Shall Not be Used to Commit or Promote Fraud

Generative AI systems are being used to create text, photographic images and video that may be inaccurate or misleading in terms of representing factual events or information. In these cases, developers intentionally alter or create images, text, and video to create false information.

Recommendation 2: All AI output that alters or creates text, images or video in order to communicate factually inaccurate events or information must be explicitly and prominently labeled as "fiction" or "fake". Violations of this rule will be legally considered as fraud.

AI Shall Avoid and Expose Bias

The challenge of AI bias has received considerable attention, with the potential of creating or perpetuating biases against certain groups of people. These biases often are introduced due to limited or statistically biased training sets (i.e., limited representativeness of problems) that create biases towards certain sets of conclusions [4-6], as well as artifacts that creep into the development of the AI algorithms [7]. AI biases lead the AI to perform more poorly or inaccurately in situations that are different than what it has been trained on. As a more general case, bias can be considered any use of an AI system in situations outside of its training [8], i.e., an over-generalization that occurs when AI trained to operate in certain conditions is applied in other conditions.

People are often expected to be able to compensate for AI shortcomings, like bias, by substituting their own knowledge and judgement in cases in which the AI may be deficient. Paradoxically, however, AI makes it very difficult to do so. First, these biases tend to be hidden due the opaque nature of machine learning techniques used to create AI. Even the developers of AI systems may not know what biases have inadvertently been introduced in the learning process. Furthermore, the users of AI systems are generally a different set of people than the developers of the AI, and therefore are even less likely to understand the limitations of its training or what situations it should be limited to.

Secondly, humans do not form their decisions independently from AI, but are directly influenced by the recommendations or assessments from the AI [9]. People tend to anchor on the recommendation of the AI system, and then gather information to agree or disagree with it, creating confirmation bias[8; 10]. AI biases therefore can directly compound human biases in the decision process, reducing the reliability of the joint human-AI system[8]. Further, the impact of the AI biases can vary depending on the format and framing of the AI system's recommendations [11-14]. Rather than overcoming human decision bias, AI can make it worse through the well-established human process of anchoring and confirmation bias.

In that AI biases are generally invisible, unknown by both developers and users of systems, and they can affect human decision-making quite surreptitiously, their negative effects can be insidious. Therefore, people will be often unable to detect and compensate for these biases (by choosing when to use the system or interjecting corrections, for example). Work is being done to improve the transparency of AI biases. [15; 16] This guardrail supports OSTP principle #2 (protection from algorithmic discrimination) and principle #5 (support for opt out decisions).

Recommendation 3: Biases in AI systems, resulting in disparate impacts on people, should be exposed and eliminated. Any known limitation of the applicability of an AI system to a set of conditions or circumstances must be made transparent to the users of the AI.

Developers of AI Systems Must be Liable for Their Products

AI systems are being proposed as systems that will improve upon human performance or reduce human error in a wide variety of applications. However, over 40 years of research data show that such systems often introduce new types of errors and problems for human performance because people are expected to compensate for AI limits[17-21]. Additionally, human users of the technology are frequently unaware of these effects, particularly exactly *when* the system is unable to perform properly, and are thus unable to compensate for its deficiencies if required to do so. It is therefore inappropriate to hold human users accountable for the performance of AI systems when they often have limited understanding of its capabilities and limitations within specific situational contexts of use.

Recommendation 4: Developers of AI systems shall assume liability for the performance of their systems.

Additional AI Guardrails for Safety Critical Applications

AI is being proposed for many applications that directly impact the safety and well-being of people, including (but not limited to) driving, flying, healthcare, power systems, and military operations. For any use of AI in a safety-critical application, extra guardrails are required in order to protect people from poor performance or unintended consequences in the use of these systems. Although AI systems are often promoted as improving safety by eliminating human error, little data exists to support such claims. In fact, failures of AI can introduce new types of errors [22] with significant safety implications. A number of more specific guardrails are needed in safety critical systems where AI is implemented.

AI Shall be Explainable

AI systems must be equipped with explainability features that allow people who interact with it to understand the system's capabilities and limitations for performance (including what factors it does or does not consider in its assessments). Because AI can be both opaque and changeable, developing and implementing effective AI explanation systems is important for helping people to develop accurate mental models of the AI. The benefits of AI explainability have been demonstrated in several studies [23-25]. AI explanations need to consider the capabilities of the human receiver (e.g., expertise, bandwidth, prior knowledge and assumptions) as well as provide effective methods for explanation delivery (e.g., they need to be both causative and contextual) [26; 27].

Recommendation 5: AI systems must be able to explain the rationale for its actions or outputs in an understandable format for the people using the system. AI explanations should provide an explanation of why it makes particular recommendations or takes actions in each case (including relevant contextual features), tailored to the needs of the user.

AI Shall be Transparent

In order to be useful, people must trust the output of the AI when it is correct, and must know when to reject that output when it is incorrect or inappropriate for the situation.[2] They must also be aware of when the AI system is not functioning properly so that they avoid reliance on unreliable data and can take corrective actions. In addition to AI explanations (which tend to be retrospective and involve general capabilities), this requires *AI transparency* which involves presenting real-time information to users on the level of reliability of the AI for the current situation at hand[8; 26]. Transparency means that users should be provided with information on how well the AI is working, its assessment of the current situation, current mode, the reliability of the underlying data or sensors that feed the AI, and its level of confidence in any assessments or recommendations that it makes.[9; 26] Providing AI transparency has been shown to significantly reduce poor performance outcomes when people work with AI systems. [28; 29]

Further, users need to understand the capabilities and limitations of AI for addressing different types of situations and classes of data within the current and upcoming context.[9] AI transparency is important for not just understanding the overall reliability and robustness of the system in general, but for allowing people to properly calibrate their trust in real-time[30-32]. AI that provides just-in-time information with the intention of serving as a decision support tool must be transparent about the capabilities, confidence and variables considered within the AI model. [26; 33]

Recommendation 6: AI systems must be transparent to users during use, providing information on the ability of the AI to handle the current and upcoming situations, its current mode and situation assessment, the reliability of the underlying data or sensors that feed the AI, and its level of confidence in any assessments or recommendations that it makes. Transparency regarding accidents and incidents must also be provided through data sharing to relevant government agencies.

AI Systems Shall be Tested with Human Users

The development of user interfaces that allow people to interact effectively with AI technologies and properly understand any performance issues requires testing of the technology in a wide variety of realistic situational contexts with a representative set of human users,[8] following informed consent and ethics [34].

The design of AI must avoid known human performance issues and provide effective mechanisms for human oversight and intervention. AI systems implemented in safety critical applications (e.g., driving, flying, power systems, healthcare) should be required to demonstrate equivalent or improved safety (as compared to manual operations), across both situations where it is reliable and those where it is not (i.e., safety must be established in automation failure conditions that involve resumption of control or over-ride by human operators). In cases of AI failure, or in situations that it cannot handle, safe transition to human control within the time available to allow accident avoidance is required. Safe transition time should take into account human-decision making and execution time, as well as time required to overcome human vigilance deficits induced by automation complacency[35] and lowered levels of task engagement[17; 36].

> *Recommendation 7: AI systems used in safety-critical applications must undergo testing in realistic conditions of use with representative users. The ability of human operators to detect AI performance deficiencies and safely assume control of operations within the time available to avoid accidents must be demonstrated, taking into account potential states of low human vigilance and distractions with competing tasks.*

AI Shall Provide Safety Alerts

AI systems introduce the need for additional information on displays and capabilities to support user interaction and decision-making[26; 37].

> *Recommendation 8: The user interface for AI systems must provide salient and timely alerts to operators when manual interventions are required to maintain safety, or when transition from automated to manual operation is required.*

AI Shall be Fail Safe

AI systems should include provisions for safe fallback states when the automation fails to perform correctly for any reason[38; 39]. Effective information displays and control override options for operators should be incorporated in the design and development of fallback strategies. Systems employing AI should not require the human operator to perform beyond human performance limits. When the AI is operating with uncertainty, the AI should operate in a less risky manner.

> *Recommendation 9: AI systems that are employed in any operation that has the potential for harm must be designed to revert to a safe state in conditions in which the system fails to perform properly for the situation.*

Training Shall be Provided for Users of AI Systems

The developers of AI systems should be required to provide user training on the capabilities, limitations and behaviors of its technology (including the range of operational conditions the AI systems can and cannot handle) so that operators obtain an accurate mental model required for effective oversight and interaction with them[8]. The effectiveness of the training format and content should be evidence-based to show successful outcomes with naïve operators. New training should be provided on any AI updates that are made over the course of the system's lifetime so that the AI's behavior remains predictable to the operator[40; 41]. Periodic updates to AI software (which may be provided over the internet on a frequent basis) can dramatically affect how the AI performs, affecting the human operator's understanding of AI actions and capabilities. Steps should be taken to require follow-on training for updates that affect AI behaviors and control.

> *Recommendation 10: Effective training on the capabilities, limitations and behaviors of AI systems must be provided to system operators by developers. Training updates are required each time the AI software is updated*
>
> .

Autonomous AI Systems Shall be Validated and Certified

In cases where the AI system is designed to operate independently (e.g., an autonomous vehicle without a human driver, or an autonomous air vehicle without a human operator), the AI system must go through validation testing to demonstrate a level of safety equivalent to or exceeding that of experienced and unimpaired human operators. A certification process should be implemented for such systems to establish testing requirements and review validation testing data to determine that high levels of system safety have been demonstrated prior to approving the use of these systems for operating in safety critical applications.[42; 43]

> *Recommendation 11: AI systems that operate autonomously must pass a certification process based on validation testing data that demonstrates safety performance that meets or exceeds that of experienced, unimpaired humans in realistic operational conditions, including hazard states.*

Summary

While AI can provide useful products and services, the potential for negative impacts on human performance are significant. Only by mitigating these problems through the establishment of effective guardrails can the benefits of AI be realized, and negative outcomes minimized. Particularly for systems that have safety impacts, it is critical that AI systems be designed and implemented to work effectively for human users of the AI, and that AI applications are objectively tested though a detailed certification process. This process is critical for moving AI forward into successful adoption to meet America's goals for leadership in the field.

About HFES

With over 3,000 members, the Human Factors and Ergonomics Society (HFES) is the world's largest nonprofit association for human factors and ergonomics professionals. HFES members include psychologists, engineers and other professionals who have a common interest in working to develop safe, effective, and practical human use of technology, particularly in challenging settings.

References

1. Sheppard, B., Kouyoumjian, G., Sarrazin, H., & Dore, F. (2018). The business value of design. McKinsey and Company. Retrieved from https://www.mckinsey.com/capabilities/mckinsey-design/our-insights/the-business-value-of-design, https://www.mckinsey.com/capabilities/mckinsey-design/our-insights/the-business-value-of-design

2. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.
3. Kale, A., Nguyen, T., Harris Jr, F. C., Li, C., Zhang, J., & Ma, X. (2022). Provenance documentation to enable explainable and trustworthy AI: A literature review. Data Intelligence, 1-41.
4. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine, 178(11), 1544-1547.
5. Daugherty, P. R., & Wilson, H. J. (2018). Human+ machine: Reimagining work in the age of AI: Harvard Business Press.
6. West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
7. Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749.
8. National Academies of Sciences Engineering and Medicine. (2021). Human-AI teaming: State-of-the-art and research needs. Washington, DC: National Academies Press.
9. Endsley, M. R., & Jones, D. G. (2012). Designing for situation awareness: An approach to human-centered design (2nd ed.). London: Taylor & Francis.
10. Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge, UK: Cambridge University Press.
11. Endsley, M. R., & Kiris, E. O. (1994). Information presentation for expert systems in future fighter aircraft. International Journal of Aviation Psychology, 4(4), 333-348.
12. Banbury, S., Selcon, S., Endsley, M., Gorton, T., & Tatlock, K. (1998). Being certain about uncertainty: How the representation of system reliability affects pilot decision making. Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (pp. 36-41). Santa Monica, CA: Human Factors and Ergonomics Society.
13. Friesen, D., Borst, C., Pavel, M., Masarati, P., & Mulder, M. (2021). Design and Evaluation of a Constraint-Based Helicopter Display to Support Safe Path Planning. Proceedings of the Nitros Safety Workshop (pp. 9-11).
14. Selcon, S. J. (1990). Decision support in the cockpit: Probably a good thing? Proceedings of the Human Factors Society 34th Annual Meeting (pp. 46-50). Santa Monica, CA: Human Factors Society.
15. Kiyasseh, D., Laca, J., Haque, T. F., Otiato, M., Miles, B. J., Wagner, C., . . . Hung, A. J. (2023). Human visual explanations mitigate bias in AI-based assessment of surgeon skills. NPJ Digital Medicine, 6(1), 54.
16. Mazijn, C., Prunkl, C., Algaba, A., Danckaert, J., & Ginis, V. (2022). LUCID: Exposing algorithmic bias through inverse design. arXiv preprint arXiv:2208.12786.
17. Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. Human Factors, 59(1), 5-27.
18. Funk, K., Lyall, B., & Riley, V. (2000). A comparative analysis of flightdecks with varying levels of automation. Final Report prepared for the FAA Chief Scientific and Technical Advisor for Human Factors, 1-17.
19. Hancock, P. A. (2019). Some pitfalls in the promises of automated and autonomous vehicles. Ergonomics, 62(4), 479-495.
20. Strauch, B. (2017). The automation-by-expertise-by-training interaction: Why automation-related accidents continue to occur in sociotechnical systems. Human factors, 59(2), 204-228.
21. Endsley, M. R. (2023). Ironies of artificial intelligence. Ergonomics.
22. Cummings, M. (2021). Rethinking the maturity of artificial intelligence in safety-critical settings. AI Magazine, 42(1), 6-15.
23. Bass, E. J., Baumgart, L. A., & Shepley, K. K. (2013). The effect of information analysis automation display content on human judgment performance in noisy environments. Journal of Cognitive Engineering and Decision Making, 7(1), 49-65.
24. Oduor, K. F., & Wiebe, E. N. (2008). The effects of automated decision algorithm modality and transparency on reported trust and task performance. Proceedings of the Proceedings of the Human Factors and Ergonomics Society Annual Meeting (pp. 302-306). Los Angeles, CA: Sage.

25. Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., & Gombolay, M. (2021). The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. Advances in Neural Information Processing Systems, 34, 610-623.

26. Endsley, M. R. (2023). Supporting human-AI teams: Transparency, explainability, and situation awareness. Computers in Human Behavior, 140, 107574.

27. Sanneman, L., & Shah, J. A. (2022). The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. International Journal of Human–Computer Interaction, 1-17. doi: 10.1080/10447318.2022.2081282

28. Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2022). Engineering psychology and human performance (5th ed.). New York: Routledge.

29. van de Merwe, K., Mallam, S., & Nazir, S. (2022). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. Human Factors, 00187208221077804.

30. Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. C. (2017). Using agent transparency to support situation awareness of the autonomous squad member. Cognitive Systems Research, 46(December), 13-25.

31. Panganiban, A. R., Matthews, G., & Long, M. D. (2020). Transparency in autonomous teammates: intention to support as teaming information. Journal of Cognitive Engineering and Decision Making, 14(2), 174-190.

32. Stowers, K., Kasdaglis, N., Rupp, M., Chen, J. Y. C., Barber, D., & Barnes, M. (2017). Insights into human-agent teaming: Intelligent agent transparency and uncertainty Advances in Human Factors in Robots and Unmanned Systems (pp. 149-160): Springer.

33. Klein, G. A., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intelligent Systems(November/December), 91-95.

34. U. S. Department of Health and Human Services. (2023). Office for Human Research Protections: Regulations, Policy & Guidance, from https://www.hhs.gov/ohrp/regulations-and-policy/index.html

35. Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381-410.

36. Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. Human Factors, 56(3), 476-488.

37. Federal Aviation Administration Human Factors Team. (1996). The interfaces between flightcrews and modern flight deck systems Washington, DC: FAA.

38. Saleh, J. H., Marais, K. B., & Favaro, F. M. (2014). System safety principles: A multidisciplinary engineering perspective. Journal of Loss Prevention in the Process Industries, 29, 283-294.

39. Steimers, A., & Bömer, T. (2021). Sources of risk and design principles of trustworthy artificial intelligence. Proceedings of the International Conference on Human-Computer Interaction (pp. 239-251). Springer.

40. Casner, S. M., & Hutchins, E. L. (2019). What do we tell the drivers? Toward minimum driver training standards for partially automated cars. Journal of Cognitive Engineering and Decision Making, 13(2), 55-66.

41. Endsley, M. R. (2017). Autonomous driving systems: A preliminary naturalistic study of the Tesla Model S. Journal of Cognitive Engineering and Decision Making, 11(3), 225-238.

42. Koopman, P., Ferrell, U., Fratrik, F., & Wagner, M. (2019). A safety standard approach for fully autonomous vehicles. Proceedings of the Computer Safety, Reliability, and Security: SAFECOMP 2019 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Turku, Finland, September 10, 2019, Proceedings 38 (pp. 326-332). Springer.

43. Koopman, P., Hierons, R., Khastgir, S., Clark, J., Fisher, M., Alexander, R., . . . Torr, P. (2019). Certification of highly automated vehicles for use on uk roads: Creating an industry-wide framework for safety.