17
Addressing AI Vulnerabilities Through Human-Centered Approaches and Risk Frameworks

Sarvesh Sawant[1], Aasish Bhanu[2], Beau G. Schelble[3], and Kapil Chalil Madathil[1]

[1]Department of Industrial Engineering, Clemson University, Clemson, SC, USA

[2]Divination Reality Labs LLC, Greenville, SC, USA

[3]*Department of Industrial and Systems Engineering, University of Tennessee, Knoxville, TN, USA*

17.1 Introduction

Artificial intelligence (AI) is increasingly being integrated into critical systems across diverse sectors. Its deployment spans domains such as healthcare—where it is used to support diagnostics, personalized treatment, and operational efficiency; defense—where it is used to enhance situational awareness, autonomous systems, and threat detection; and aviation—where it is used to optimize maintenance, navigation, and air traffic management (Bienefeld et al., 2024; Kirwan, 2024; Mallick et al., 2022; Sawant et al., 2022). AI's influence is transforming workflows, improving system resilience, and enabling data-driven innovation in both civilian and strategic contexts. Studies have found that AI improves decision-making, increases efficiency, and effectively automates complex tasks by processing large volumes of data quickly and effectively (Koo et al., 2024; Khosravi et al., 2024). It is also extensively being used in solving practical problems in different industries because of its ability to recognize patterns, forecast results, and adapt to new information (Khosravi et al., 2024).

While AI offers various practical advantages, it is not without risks. AI can make mistakes when it is trained on incomplete or biased information. It may misinterpret data, provide information that is difficult for humans to comprehend, or overlook details that a human would normally consider. In some cases, AI systems may fail to adapt when faced with situations they were not designed to handle. These issues can lead to poor outcomes, such as incorrect predictions or unfair results, which undermine trust in the technology. This chapter explores the types of AI vulnerabilities, examines real-world examples, and discusses strategies to minimize these risks.

17.1.1 AI Vulnerabilities

AI vulnerabilities are weaknesses in the design, training, or deployment of AI systems that can lead to errors, unexpected behavior, or undesirable outcomes (Scherlis, 2024; Spring et al., 2020). These problems may result from human error, overreliance, or inadequate supervision, or they may stem from technical issues in AI, such as biases in the training data or algorithmic mistakes (Gaube et al., 2021; Papernot et al., 2016; Spring et al., 2020). These flaws can have significant repercussions, especially as AI systems are increasingly being used in high-stakes environments. For example, an AI-powered self-driving vehicle may fail to detect an obstruction in its path, leading to an accident, or an AI used in recruitment may unintentionally discriminate against certain job applicants. Due to the complexity of AI systems, it is challenging to anticipate every potential failure. Further, the integration of human–AI interaction introduces an additional layer of uncertainty when designing solutions to address these vulnerabilities.

An important factor contributing to AI vulnerabilities is the lack of transparency and explainability in AI decision-making. Transparency refers to the system's ability to provide real-time insights into its current operations, enabling users to understand its status (Andrada et al., 2023). Explainability refers to the AI's ability to clearly communicate the reasoning behind its actions or recommendations in a way that users can understand (Arrieta et al., 2020). The absence of transparency and explainability can make it difficult for users to interpret, evaluate, or effectively supervise AI-driven processes (Cheong, 2024).

This lack of transparency and explainability becomes even more problematic when paired with cognitive biases that influence human behavior. For example, automation bias—the tendency for humans to over-trust AI recommendations—can cause users to accept incorrect AI recommendations without question, potentially leading to critical errors in high-risk settings such as healthcare or aviation (Lyell & Coiera, 2017, Parasuraman & Manzey, 2010). Conversely, not trusting

AI can lead users to disregard information, reducing the overall effectiveness of AI-supported decision-making (Afroogh et al., 2024). These challenges highlight the importance of incorporating human factors into AI development to ensure that systems align with human capabilities and limitations. The focus should be on preventing vulnerabilities through a human-centered design (HCD) approach.

The way AI communicates its decisions, the extent to which users comprehend and trust its outputs, and the level of human oversight are important factors in determining whether an AI system operates as intended or leads to potential failures. The operation of AI systems can become erratic or unsafe if factors such as cognitive limitations, decision-making biases, and situational awareness are not considered during the design process. Real-world failures in aviation, autonomous driving, healthcare, and other safety-critical domains demonstrate the consequences of ignoring human factors in the design and implementation of AI-driven automation systems.

A prime example would be the Boeing 737 MAX-8 disaster of 2018–2019 (Collings et al., 2022; Miller et al., 2023). The disaster was brought about when Boeing developed a system to assist pilots in managing the aircraft's angle of attack, known as the Maneuvering Characteristics Augmentation System (MCAS). The system did not, however, provide adequate transparency, operating based on a single sensor, while pilots were neither aware nor trained to interact with the new autonomous system. Upon faulty data from that sensor, the system constantly pushed the nose of the aircraft down without clear feedback and without an easy means for the pilots to override it. Pilots faced difficulty identifying the system malfunction and taking corrective action because the MCAS design was not transparent and lacked actionable feedback or override mechanisms. Two fatal crashes, costing 346 lives, and the global grounding of the Boeing 737 MAX-8 fleet drew attention to the critical role that system transparency, real-time feedback, and user interpretability play in automation design.

Another example is an accident that occurred in Arizona in 2018, when an Uber self-driving vehicle struck a pedestrian (Harris, 2023; Rice, 2019). The AI system first misidentified the pedestrian as an unknown object, then as a bicycle, and finally as a pedestrian. However, because it had been programmed to ignore certain false positives, the system did not respond in time. The human operator, who was meant to intervene, was complacent, and the automation did not provide adequate alerts. Humans are not well suited to passively monitor automated systems for extended periods (Greenlee et al., 2024; Korber et al., 2015). AI systems should be designed to ensure rapid control transitions between automation and human operators during emergencies. The AI must keep human operators engaged, provide timely and clear intervention prompts, and prevent overreliance or complacency—thereby alerting humans when a takeover is necessary.

The failures of AI in the healthcare domain also highlight how lapses in addressing human factors can endanger safety. IBM Watson for Oncology, a system created to assist physicians in making clinical decisions about cancer treatment, reportedly generated incorrect and unsafe recommendations (Strickland, 2019). Watson's AI was trained using a limited dataset of expert opinions, including those of medical oncologists, rather than actual patient data, resulting in erroneous diagnoses and untrustworthy treatment suggestions. For physicians, the reasoning behind Watson's decisions was difficult to interpret, potentially undermining trust. As a result, several hospitals discontinued the use of the AI tool. This case underscores the need for AI systems used in high-stakes decision-making—such as in healthcare—to produce transparent and interpretable outputs so users can safely evaluate and act on AI-generated recommendations.

In 2018, Amazon terminated its AI-based hiring system due to the algorithm's discrimination against female candidates (Dastin, 2022). The AI had been trained on historical hiring data biased toward male applicants, thereby perpetuating gender biases and unjustly screening out qualified women. Because the AI operated without a transparent

reasoning process and lacked proper monitoring, this issue went unnoticed for an extended period. This case raises concerns about the ethical risks posed by AI systems that function without mechanisms to detect or mitigate bias, reinforcing the need for explainability and accountability in AI design.

These examples illustrate the risks of developing AI systems without considering human factors. Poor implementation can result in safety hazards, liability issues, ethical violations, and a loss of trust. The path forward should integrate human factors into the core of AI system design. Addressing vulnerabilities requires a structured approach to risk identification, assessment, and management prior to deployment. Several established models and frameworks offer systematic methodologies for identifying, analyzing, and mitigating vulnerabilities in AI systems, as outlined below.

17.1.2 Frameworks for Identifying and Mitigating AI Vulnerabilities

With the rapid integration of AI systems in domains involving critical decision-making, a structured approach is required to understand and evaluate their reliability, safety, and adherence to ethical standards. As discussed above, failures in AI design emerge due to a lack of transparency, automation bias, cognitive overload, or insufficient supervision, leading to accidents. Researchers and policymakers have developed frameworks and standards that provide structured methodologies for assessing AI risks, and we will discuss two such frameworks here: (1) the National Institute of Standards and Technology AI Risk Management Framework (NIST AI RMF) and (2) ISO/IEC 23894:2023.

The NIST AI RMF was developed as a structured approach to help organizations identify, assess, and mitigate AI-related risks (NIST, 2023). It involves integrating risk assessment at every stage of AI integration to build transparent, fair, and secure AI systems. The NIST AI RMF is built around four key functions: govern, map, measure, and manage.

The govern function focuses on cultivating an organizational culture that proactively identifies and manages AI-related risks by adhering to ethical guidelines, establishing clear accountability structures, and ensuring compliance with legal and regulatory requirements throughout AI system development. The map function helps organizations anticipate and understand potential AI risks by encouraging developers to consider how the system will be used, who it will impact, and where it may encounter vulnerabilities across real-world settings. This will enable organizations to proactively anticipate and avoid AI failures rather than react once the failure has occurred by looking through those lenses early on. The measure function includes testing and evaluating AI systems to identify risks like bias, security gaps, and performance challenges. As AI systems learn and evolve over time, a one-time evaluation does not suffice. Continuous monitoring will identify risks as they arise, allowing organizations to take timely and appropriate measures before they become systemic failures. The manage function ensures that organizations take action based on identified AI risks, requiring ongoing updates and improvements to the system. The NIST AI RMF provides a practical approach to managing these risks throughout the AI lifecycle—from development and testing to real-world deployment.

ISO/IEC 23894:2023 is an international standard for managing risks related to AI systems (ISO, 2023). It is designed to assist organizations in identifying, assessing, managing, and reporting AI-related risks to ensure transparency, accountability, and ethical practices in AI development. It starts with risk identification and context setting, where organizations will define which risks apply to their AI system. Risks involve bias in data, lack of model transparency, potential adversarial attacks, and compliance with legal and ethical guidelines. Organizations then establish clear risk criteria based on the AI system's intended use and assess AI risks in the broader context of overall regulatory and operational standards. After identifying the risks, ISO/IEC 23894:2023 requires organizations to be continually tested and validated. This stage ensures that the performance of AI

models does not degrade over time or create unintended bias because of changes in data or user interactions. The standard defines best practices for bias audits, adversarial testing, and security assessment. These practices help organizations uncover AI vulnerabilities before they become detrimental. To address identified risks, the framework suggests ways to improve the reliability and transparency of AI. This includes implementing measures to ensure transparency in AI decisions, incorporating human oversight mechanisms to promote accountability in AI-driven decisions, and establishing fail-safe systems to correct errors made by AI. The final element of the standard is monitoring and ongoing improvement of AI systems. It requires continuous monitoring and periodic reviews to ensure that AI systems remain aligned with their intended goals. At the same time, organizations are expected to maintain risk reports, accountability structures, and compliance with applicable legal and regulatory frameworks to minimize the impact of escalating AI risks.

Both frameworks propose a risk-based, lifecycle-oriented approach to managing AI risks, grounded in principles of trustworthiness, accountability, and continuous improvement. However, they differ in scope and application—NIST provides flexible, voluntary guidance primarily for US stakeholders, while ISO/IEC 23894 offers a more prescriptive, internationally standardized framework intended for formal compliance and integration with global risk management systems.

17.2 Types of AI Vulnerabilities

AI vulnerabilities can be categorized into two classes: technical vulnerabilities, which arise from algorithmic and system design, and human-related vulnerabilities, which stem from how people interact with and perceive AI systems. This section discusses both types of vulnerabilities and how they can affect the implementation of AI systems across different domains.

17.2.1 Technical Vulnerabilities

Technical vulnerabilities in AI refer to inherent weaknesses in an AI system's algorithms, data processing, or infrastructure (Spring et al., 2020). These weaknesses can compromise AI performance, security, and reliability, making AI models susceptible to manipulation, bias, and unexpected failures. One of the significant technical threats is adversarial attacks, which involve subtle manipulations of the input data designed to fool AI models into making incorrect predictions. These attacks exploit AI models that rely on statistical correlations rather than a core understanding of the data, making them susceptible to subtle manipulation (Goodfellow et al., 2014). One such case was the 2020 research by McAfee showing that tiny stickers placed over a 35-mph speed limit sign could induce Tesla's Autopilot system to misinterpret it as 85 mph, thus potentially allowing for drastically dangerous speeds (Povolny, 2024). In another case, researchers changed a few pixels on an image of a panda, which the model misclassified with high confidence as a different animal known as a gibbon. In contrast, human observers could not see any changes (Goodfellow et al., 2014). Such examples illustrate security concerns of adversarial attacks in AI applications.

Data poisoning is another vulnerability, where manipulated or incorrect data are added to an AI training set, thus corrupting the entire learning process. Poisoning attacks may give rise to biased, unsafe, or exploitable AI behaviors (Biggio & Roli, 2018). A particular case goes back to 2016 when Microsoft's Tay chatbot was set to learn from Twitter interactions. Within 24 hours of its launch, users deliberately fed Tay with harmful and offensive comments, which Tay assimilated and replicated, and as a result, Microsoft had to shut it down (Neff & Nagy, 2016). The Facebook AI content moderation system faced a similar problem, with users uploading modified versions of hate speech to circumvent AI filters, demonstrating how AI models can be deceived into misclassifying harmful content (Tramer et al., 2016). A hazardous type of data poisoning is known as backdoor attacks, which happen

when an attacker has inserted malicious data into the training process that will allow the AI system to perform normal behavior under a majority of conditions but act incorrectly if it encounters or is triggered by a specific input (Guo et al., 2022). These backdoors can escape detection during testing, enabling the attacker to later selectively exploit the system and issue malicious actions that pose substantial safety and security risks.

Privacy risks arise during model inversion attacks, wherein attackers attempt to extract sensitive training data from AI models. In these attacks, privacy is compromised by reconstructing confidential data, such as faces from facial recognition models or medical records from diagnostic AI systems (Fredrikson et al., 2015). Research has shown that facial recognition AI models could be reverse-engineered to recreate images of individuals whose faces were employed during training (Zhang et al., 2020). In another instance, Google's Smart Reply AI system was examined for privacy leaks in that attackers could infer private conversations based on the AI model response patterns (Carlini et al., 2019). Such vulnerabilities in AI systems provide hackers with an opportunity for unauthorized access and data breaches. Hackers use loopholes in the security architecture of AI systems to manipulate results or gain unauthorized access to sensitive information (Papernot et al., 2016). In 2019, a cybercriminal used the deepfake AI-generated voice of the CEO to convince one of the employees to transfer money into a fraudulent account (Stupp, 2019). This illustrates the emerging risks AI presents concerning fraudulent activities and cybersecurity. Additionally, hackers have facilitated attacks on AI-driven fraud detection systems. In particular, many have manipulated AI classifiers into incorrectly identifying fraudulent activities as legitimate, exposing gaping holes in automated security protocols (Biggio et al., 2013).

Other problems that AI models face are overfitting and poor generalization, whereby they perform exceedingly well on training data but cannot perform with the same accuracy when applied to real-world cases. This may arise in scenarios where models have memorized spe-

cific patterns instead of learning generalizable principles, which render them ineffective outside controlled environments Zhang et al., 2016). During the COVID-19 pandemic, an AI model developed to detect COVID-19 from chest X-rays offered high lab accuracy but failed in real-world hospitals. Researchers found that the AI had learned to associate hospital logos or equipment markings with COVID-19 cases instead of recognizing disease-related patterns. This ultimately led to poor generalization (Roberts et al., 2021). AI models should be trained on diverse and representative datasets to prevent unintended biases and ensure reliability in different environments.

17.2.2 Human-Centered AI Vulnerabilities

Human-centered AI (HCAI) vulnerabilities result from how people perceive, interpret, and interact with AI systems. One of the significant challenges when humans interact with highly autonomous AI systems is overreliance on AI, where users place excessive trust in automated systems without sufficient human oversight (Lyell & Coiera, 2017). This overreliance can cause human operators to fail to intervene when human control is necessary while working with an AI system. One such incident occurred in February 2025 when a Tesla Cybertruck operating in self-driving mode crashed in Reno, Nevada (Cervantes & McAndrew, 2025). While merging onto a highway, it collided with the curb and struck a pole. The driver had no time to intervene because they were overreliant on the AI's decision-making ability. This incident highlights the growing concerns with semi-autonomous AI systems, where over-trust in automation leads to complacency and a reduced level of diligence by humans. This tendency to over-trust AI systems can contribute to serious errors, oversights, or even catastrophic failures.

Cognitive overload and decision fatigue are other significant factors facing effective human–AI interaction (Steyvers & Kumar, 2024). This occurs when AI systems provide users with excessive, complex, or contradictory information, affecting their ability to process and make

effective decisions (Steyvers & Kumar, 2024). Although AI aims to support high-pressure environments by streamlining decision-making processes, overwhelming information can result in mental exhaustion and reduced situational awareness. An example is the 2013 Asiana Airlines Flight 214 crash (Miller & Holley, 2018). During the incident, pilots faced difficulties interpreting the automated flight controls and data presented by the system. Their over-dependence on automation led them to misinterpret the current state of these systems. They failed to notice that the auto-throttle had been disengaged. This error resulted in a critical loss of airspeed as they approached San Francisco International Airport, resulting in a crash landing that caused numerous fatalities and injuries. This accident shows how poorly designed automation interfaces and high cognitive load can impair pilots' effectiveness in monitoring systems, recognizing anomalies, and taking corrective action promptly. The complicated automation led to confusion, delaying decision-making when it was most critical, instead of alleviating the mental workload.

Misinterpretation of AI outputs can be a significant challenge in human–AI interactions. This issue arises when users misunderstand, improperly apply, or fail to question AI-generated recommendations due to insufficient transparency and explainability (Zerilli et al., 2022; Leichtmann et al., 2023). An example occurred in 2019 with the Optum healthcare algorithm widely used in US hospitals (Obermeyer et al., 2019). The algorithm aimed to identify patients who would benefit from enhanced care management programs by using healthcare costs to predict medical needs. However, since Black patients historically incurred lower healthcare expenses compared to white patients suffering from similar medical conditions, the algorithm mistakenly inferred that Black individuals were healthier than they were. Consequently, it deprioritized these patients for essential care services. As a result, Black patients were 50% less likely than their white counterparts to be referred for high-risk care management programs despite having equivalent medical needs (Obermeyer et al., 2019). In this case, reliance on the AI's suggestions without questioning its

methodology unintentionally led doctors and hospital administrators to perpetuate disparities in healthcare access. This oversight contributed to delays in necessary treatments and ultimately worsened patient outcomes. Such cases underscore the risks of passively accepting AI-generated recommendations without examining their foundational logic and assumptions.

17.3 Mitigating AI Vulnerabilities Through Human-Centered Design

This section discusses strategies to mitigate AI vulnerabilities by incorporating HCD principles. HCD is an approach that employs evidence-based methods and principles to design useful products and services for people (Bijl-Brouwer & Dorst, 2017). This section is organized into five key areas to explore how HCD can address vulnerabilities associated with AI integration. Section 17.3.1 introduces key principles of HCD that support the development of more effective AI systems. Section 17.3.2 discusses the importance of transparency and explainability in maintaining appropriate trust, whereas Section 17.3.3 emphasizes the value of human-in-the-loop (HITL) design to safeguard against vulnerabilities like overreliance. Section 17.3.4 highlights the role of education and training when interacting with various AI systems. Section 17.3.5 briefly addresses the importance of ethical considerations in the design and implementation of AI systems.

17.3.1 Human-Centered Design Principles

Designing AI systems with humans in mind aligns the technology with human needs rather than forcing humans to adapt to the technology. A human-centered approach to system design can lead to enhanced performance, higher user satisfaction, and fewer errors (Wang et al., 2024). Poorly designed AI can confuse users, impair human decision-making capabilities, and erode trust in the technology, as illustrated in the previous section. Therefore, it is important to understand user behavior—to learn how end users interact with the AI system, identify

factors that affect trust, and detect potential flaws that might lead to misuse of the technology.

Following a user-centered design process at the design stage can ensure that the AI's behavior aligns with users' mental models and goals. This process involves steps such as understanding the context, defining user needs, developing concepts, and iterative refinement based on user testing and feedback (Ulrich & Eppinger, 2016). Once user needs and context are clearly understood, various concepts can be explored to develop solutions informed by those insights. An iterative design and testing process is employed, with designs continually refined based on user feedback. This approach helps identify potential issues early, leading to more user-friendly outcomes.

Techniques such as failure mode and effects analysis (FMEA) enable teams to proactively identify and mitigate potential failure points before deployment (Liu et al., 2013). FMEA is a systematic and proactive approach that analyzes potential failures in products, processes, designs, or services. Through FMEA or similar risk analysis tools, system designers can uncover and resolve latent errors—whether algorithmic or related to the user interface—before the system is deployed.

There are several guidelines for applying HCD to AI solutions. For example, Amershi et al. (2019) argued that designers need to make AI more transparent, always provide feedback, and support user control—echoing core HCD principles. Similarly, Shneiderman (2022) proposed a HCAI model that encourages achieving higher levels of both automation and human control, rather than viewing them as mutually exclusive. This model aims to strike a balance, enabling the AI system to efficiently handle tedious tasks while preserving human involvement in critical decision-making processes. Such an approach can potentially minimize user distrust and help maintain confidence in the system's capabilities. The HCAI model also mentions the importance of trust in AI. It mentions the need for clear explanations of how AI systems make decisions to enhance this trust. Clearly explaining how AI

works helps users understand its capabilities and see it as a tool that supports human potential and accountability, rather than as a mysterious "black box."

17.3.2 Transparency and Explainability in AI

Transparency and explainability are critical components of building trust in AI systems and can help reduce the risk of misunderstandings between the actions of AI systems and users' interpretations. As powerful and complex AI continues to emerge, the need for explainable AI (XAI) has become essential to the field of human–computer interaction (Mueller et al., 2021). XAI techniques aim to clarify the reasoning and processes behind AI actions, thereby supporting informed decision-making. Transparent and XAI enables users to form better mental models of the system, which is important for building trust. Maintaining appropriate trust when interacting with AI is important to prevent both over and under reliance on the technology (Lee & See, 2004), which can lead to misuse and disuse, respectively (Parasuraman & Riley, 1997). Calibrated trust in the system also helps reduce skepticism about the outcomes produced by AI (Tiwari, 2023).

Users risk misinterpreting outcomes or failing to recognize when to question AI recommendations without clear explanations, which can lead to serious consequences—especially in fields like healthcare and finance (Yang et al. 2023). For example, Caruana et al. 2015 work highlights the potential risks of opaque AI models in medical settings, where unclear diagnostic logic can jeopardize patient safety since physicians cannot validate AI-generated decisions or recommendations (Caruana et al., 2015). In the financial sector, researchers have cautioned against the dangers of black-box AI models that could mislead lenders or regulators, potentially introducing errors and biases (Rudin, 2019). Similarly, Miller (2019) emphasizes the importance of transparent explanations in maintaining trust, showcasing how user comprehension is essential for the safe and ethical deployment of AI in high-stakes domains. Developing frameworks and methodologies

that prioritize explainability can significantly enhance user acceptance and trust in AI systems (Patidar et al., 2024).

17.3.3 Human-in-the-Loop Design

Although AI systems have advanced significantly, they still encounter errors or situations they were not explicitly trained to handle. Therefore, it is important for humans using such systems to maintain oversight or some level of control—especially when high stakes are involved. The HITL design could be used to create a complementary interaction between humans and AI systems. HITL promotes AI as a collaborative partner or assistant rather than granting it full autonomy, allowing humans to supervise its actions while retaining the ability to intervene when necessary. This approach leverages the strengths of both humans and machines, addressing the potential shortcomings of each when operating alone (Vallati & Chrpa, 2023). AI systems can quickly process vast amounts of data and perform repetitive tasks, but they lack the specialized understanding that humans possess, such as contextual awareness and ethical judgment (Kalyanathaya & Prasad, 2024). Human–AI collaboration can help prevent catastrophic events by catching errors that AI might miss. For example, a human driver in the seat of an autonomous vehicle could prevent an accident by intervening if the AI fails at any point. Passive monitoring alone is not sufficient in such situations, especially when the AI is perceived as highly reliable. This perception can lead to overreliance and potentially cause accidents, as seen in the 2018 Uber self-driving incident in Tempe, Arizona (Lawless, 2022). Decades of research in human–automation interaction supports this idea, emphasizing that removing humans entirely can be dangerous, as automation may not be equipped to handle every situation—particularly rare or high-stakes ones (Parasuraman et al., 2000). Shneiderman (2022), in his HCAI model, argues that instead of placing “humans in the loop”—where AI remains central—AI should be designed with users at the center, with technology embedded in supportive roles. It is important to conduct in-depth research and stakeholder discussions to determine who should play

the primary role, depending on the specific context for which the AI is designed.

17.3.4 User Education and Training

User education and training are important in mitigating human errors, misinterpretations, and overreliance and enhancing technology acceptance. Comprehensive education and training help users gain a deeper understanding of the capabilities and limitations of the technology they interact with. It can enable them to critically evaluate and question AI decisions, promoting more effective collaboration between humans and AI systems. Lacking AI literacy can lead to automation bias, where users over-trust automated decisions or misinterpret the model's outputs. For instance, clinical decision support AIs have occasionally led clinicians to overlook apparent errors because of their overreliance on the system (Patil et al., 2025). Hence, improving AI literacy through targeted training is important to mitigate such issues.

Research has shown that training may minimize decision support errors, reducing automation bias, which could help users remain vigilant and accountable while interacting with AI technology (Skitka et al., 2000). However, Mosier et al. (2001) have found that training had no impact on automation bias. This shows that rather than helping with automation bias, complacency error is reduced by training (Goddard et al., 2011). Organizations should promote a culture of continuous learning and security awareness to ensure users remain vigilant against potential AI failures and biases. To combat emerging security threats and enhance the resilience of AI models, it is important to invest in training and actively promote knowledge sharing and collaboration among stakeholders (Mohammed et al., 2024). Training programs should be tailored to the specific context. For example, what a radiologist needs to know about an AI diagnostic assistant—such as how to interpret confidence scores—differs from what a driver needs to know about an autonomous driving aid—such as when the system

might disengage. In all cases, the objective is to cultivate informed skepticism among users, helping them understand that AI outputs are suggestions, not absolute truths. With this mindset, user training can potentially counteract overreliance and promote responsible human oversight.

17.3.5 Ethical Considerations

Beyond the aforementioned flaws, AI systems also face ethical vulnerabilities that could potentially harm users and society if left unaddressed. These vulnerabilities are caused due to biased algorithms, opaque decision-making processes, and unintended societal consequences. The ethical implications of these vulnerabilities are profound and necessitate robust frameworks that address fairness, accountability, and potential harm. The rapid adoption of AI has brought to light several critical issues, including bias, fairness, transparency, and accountability in automated decision-making processes (Akinrinola et al., 2024). As the reliability of AI is an important factor that affects the users' trust in the AI (Bhanu et al., 2023), paying attention to the vast amount of data being utilized to train AI systems for these specific applications is essential. Past research has shown that an AI collaborator making unethical decisions causes users to lose trust in not only the AI teammate but also the entire team as a whole (Schelble et al., 2023, 2024). Therefore, considering the approach of "ethics by design," which integrates ethical considerations into the technical process of AI development rather than treating ethics as an afterthought or a mere box-checking exercise, is beneficial for AI usage (Amugongo et al., 2023).

HCD emphasizes comprehending end users' needs, values, and contexts throughout the design process to ensure that technologies are aligned with human well-being (Shneiderman, 2020). To address challenges in training data and model outputs, a collaborative approach involving stakeholders, including ethicists, domain experts, and end users, should be implemented at various design, development, testing,

and integration stages of the AI system. Involving users to co-create AI tools could reduce algorithmic biases by incorporating the user's perspective. The development of ethical AI systems must operationalize principles of transparency, accountability, and fairness through measurable design requirements, traceable decision-making processes, and equitable outcomes across stakeholder groups (Floridi et al., 2018). Integrating HCD principles such as usability testing and impact assessments could help identify errors and issues even before the deployment of the AI system. For example, stress testing AI models against edge cases can expose flaws in autonomous systems before deployment. These practices can improve the reliability of AI and also promote forming trust in AI, which is a prerequisite for AI adoption in high-stakes domains. Considering the rapid pace of AI growth, there should be more effort to enhance ethical considerations. Despite the existence of policies like IEEE Ethically Aligned Design, EU Trustworthy AI criteria, and UNESCO's AI ethics recommendations, which emphasize the need for human-centric AI development guided by fundamental human values, there is still a need for increased effort in this area. Addressing ethical considerations through HCD ensures AI systems are transparent, accountable, and aligned with human values, fostering trustworthy and resilient human–AI collaboration.

17.4 Summary

AI vulnerabilities represent a critical challenge to the successful design, development, and application of AI systems in complex environments. Addressing these challenges is becoming increasingly important as the complexity of the environments in which AI is deployed grows alongside the constant technological advances in AI. This chapter has reviewed the technical and human-related categories of AI vulnerabilities, each illustrated with various real-world examples. Technical vulnerabilities often struggle to overcome attacks due to a lack of conceptual understanding on the part of AI systems. In con-

trast, human-related vulnerabilities frequently stem from a lack of model understanding and transparency.

The consequences of these vulnerabilities require any practitioner working with AI systems to carefully consider their development and application. These vulnerabilities can be mitigated by thoroughly evaluating the repercussions of AI usage and how humans should interact with these systems. Formalized examples of review methodologies for AI systems include the two described in this chapter—ISO/IEC 23894:2023 and the NIST AI RMF—with additional examples available in the literature. These frameworks detail strategies for data validation, usage verification, human interaction evaluation, and other measures to verify the intended functioning of an AI system.

As AI systems continue to advance, it is important to ensure that humans remain in the loop, especially in high-risk scenarios. AI systems provide the greatest benefit when used in tasks or roles that leverage their unique technical strengths—such as big data analysis or rapid response times—to complement the inherent strengths of humans, such as critical thinking and managing incomplete or ambiguous information. A great deal of work remains to further explore, understand, and overcome these challenges as AI systems continue to evolve. As with industrial safety, mitigating AI vulnerabilities is a process of continuous improvement.

References

Afroogh, S., Akbari, A., Malone, E. et al. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications* 11 (1): 1–30.

Akinrinola, O., Okoye, C.C., Ofodile, O.C., and Ugochukwu, C.E. (2024). Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews* 18 (3): 50–58.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S.T., Bennett, P.N., Inkpen, K.M., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019, May 4–9). *Guidelines for human-AI interaction*. In Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems (pp. 1–13).

Amugongo, L.M., Kriebitz, A., Boch, A., and Lütge, C. (2023). Operationalising AI ethics through the agile software development lifecycle: A case study of AI-enabled mobile health applications. *AI and Ethics* 5 (1): 227–244.

Andrada, G., Clowes, R.W., and Smart, P.R. (2023). Varieties of transparency: Exploring agency within AI systems. *AI & Society 38* (4): 1321–1331.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J. et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115.

Bhanu, A., Sharma, H., Pathy, S.R. et al. (2023). Trust in artificial intelligent agent while completing a procedural construction task. *Proceedings of the Human Factors and Ergonomics Society* 67 (1): 2005–2006.

Bienefeld, N., Keller, E., and Grote, G. (2024). Human-AI teaming in critical care: A comparative analysis of data scientists' and clinicians' perspectives on AI augmentation and automation. *Journal of Medical Internet Research* 26: e50130.

Biggio, B., & Roli, F. (2018, October 15–19). *Wild patterns: Ten years after the rise of adversarial machine learning*. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (pp. 2154–2156). Association for Computing Machinery.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019, August 14–16). *The secret sharer: Evaluating and testing unintended memorization in neural networks*. In 28th USENIX Security Symposium (USENIX Security 19) (pp. 267–284).

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August 10–13). *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission*. In Proceedings of

the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721–1730). Association for Computing Machinery.

Fernando Cervantes Jr & McAndrew, S. (2025, February 15). Tesla Cybertruck crash into a pole in Nevada was in self-driving mode: Owner. *USA Today*.

Cheong, B.C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics* 6: 1421273.

Collings, D., Corbet, S., Hou, Y.G. et al. (2022). The effects of negative reputational contagion on international airlines: The case of the Boeing 737-MAX disasters. *International Review of Financial Analysis* 80: 102048.

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of data and analytics, 296–299. Auerbach Publications.

Floridi, L., Cowls, J., Beltrametti, M. et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28 (4): 689–707.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). *Model inversion attacks that exploit confidence information and basic countermeasures*. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333).

Gaube, S., Suresh, H., Raue, M. et al. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4 (1): 31.

Goddard, K., Roudsari, A., and Wyatt, J.C. (2011). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19 (1): 121.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Greenlee, E.T., DeLucia, P.R., and Newton, D.C. (2024). Driver vigilance decrement is more severe during automated driving than manual driving. *Human Factors* 66 (2): 574–588.

Guo, W., Tondi, B., and Barni, M. (2022). An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing* 3: 261–287.

Harris, M. (2023, September 22). NTSB investigation into deadly Uber self-driving car crash reveals lax attitude toward safety. *IEEE Spectrum*.

International Organization for Standardization (2023, February 6). ISO/IEC 23894:2023—Information technology—Artificial intelligence—Guidance on risk management. Retrieved from https://www.iso.org/standard/77304.html.

Kalyanathaya, K. and Prasad, K. (2024). A framework for generating explanations of machine learning models in Fintech industry. *The Scientific Temper* 15 (2): 2207–2215.

Khosravi, M., Zare, Z., Mojtabaeian, S.M., and Izadi, R. (2024). Artificial intelligence and decision-making in healthcare: A thematic analysis of a systematic review of reviews. *Health Services Research and Managerial Epidemiology* 11: 23333928241234863.

Kirwan, B. (2024). The impact of artificial intelligence on future aviation safety culture. *Future Transportation* 4 (2): 349–379.

Koo, T.H., Zakaria, A.D., Ng, J.K., and Leong, X.B. (2024). Systematic review of the application of artificial intelligence in healthcare and nursing care. *The Malaysian Journal of Medical Sciences* 31 (5): 135.

Körber, M., Cingel, A., Zimmermann, M., and Bengler, K. (2015). Vigilance decrement and passive fatigue caused by monotony in automated driving. *Procedia Manufacturing* 3: 2403–2409.

Lawless, W. (2022). Toward a physics of interdependence for autonomous human-machine systems: The case of the uber fatal accident, 2018. *Frontiers in Physics* 10: 879171.

Lee, J.D. and See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors* 46 (1): 50–80. https://doi-org.utk.idm.oclc.org/10.1518/hfes.46.1.50_30392.

Leichtmann, B., Humer, C., Hinterreiter, A. et al. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139: 107539.

Liu, H.C., Liu, L., and Liu, N. (2013). Risk evaluation approaches in failure mode and effects analysis: A literature review. *Expert Systems with Applications* 40 (2): 828–838.

Lyell, D. and Coiera, E. (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association* 24 (2): 423–431.

Mallick, R., Sawant, S., McNeese, N., and Chalil Madathil, K. (2022, September). Designing for mutually beneficial decision making in human-agent teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 66 (1): 392–396.

Miller, M., & Holley, S. (2018). SHELL revisited: Cognitive loading and effects of digitized flight deck automation. In Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2017 International Conference on Neuroergonomics and Cognitive Engineering, July 17–21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8 (pp. 95–107). Springer International Publishing.

Miller, M., Holley, S., and Halawi, L. (2023). The evolution of AI on the commercial flight deck: finding balance between efficiency and safety while maintaining the integrity of operator trust. *Artificial Intelligence, Social Computing and Wearable Technologies* 113 (2023): 14.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38.

Mohammed, A. S., Jha, S., Tabbassum, A., & Malik, V. (2024, October 25–26). *Assessing the vulnerability of machine learning models to cyber attacks and developing mitigation strategies*. Proceedings of the 2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA 2024).

Mosier, K.L., Skitka, L.J., Dunbar, M., and McDonnell, L. (2001). Aircrews and automation bias: The advantages of teamwork? *The International Journal of Aviation Psychology* 11 (1): 1–14.

Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of explanation in human-AI systems. ArXiv Preprint.

National Institute of Standards and Technology (NIST) (2023). Artificial intelligence risk management framework (AI RMF 1.0). Retrieved from https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Neff, G. and Nagy, P. (2016). Automation, algorithms, and politics| talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication* 10: 17–17.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (6464): 447–453.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). *The limitations of deep learning in adversarial settings*. In 2016 IEEE European symposium on security and privacy (EuroS&P) (pp. 372–387). IEEE.

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

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39 (2): 230–253.

Parasuraman, R., Sheridan, T.B., and Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 30* (3): 286–297.

Patidar, N., Mishra, S., Jain, R. et al. (2024). Transparency in AI decision making: A survey of explainable AI methods and applications. *Advances in Robotic Technology* 2 (1): 1–10.

Patil, S.V., Myers, C.G., and Lu-Myers, Y. (2025). Calibrating AI reliance—a physician's superhuman dilemma. *JAMA Health Forum* 6 (3): e250106–e250106.

Povolny, S. (2024, July 9). Model hacking Adas to pave safer roads for autonomous vehicles. McAfee Blog.

Rice, D. (2019). The driverless car and the legal system: Hopes and fears as the courts, regulatory agencies, waymo, tesla, and uber deal with this exciting and terrifying new technology. *Journal of Strategic Innovation and Sustainability* 14 (1): 134–146.

Roberts, M., Driggs, D., Thorpe, M. et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3 (3): 199–217.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 2019 1:5* 1 (5): 206–215.

Sawant, S., Mallick, R., McNeese, N., and Chalil Madathil, K. (2022). Mutually beneficial decision making in Human-AI teams: Understanding soldier's perception and expectations from AI teammates in human-AI teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 66 (1): 287–289.

Schelble, B. G., Lancaster, C., Duan, W., Mallick, R., McNeese, N. J., & Lopez, J. (2023, January 3–6). *The effect of AI teammate ethicality on trust outcomes and individual performance in human-AI teams*. Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS 2023) (pp. 322–331).

Schelble, B.G., Lopez, J., Textor, C. et al. (2024). Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in human-AI teaming. *Human Factors* 66 (4): 1037–1055.

Scherlis, B. (2024, August 5). Weaknesses and vulnerabilities in modern AI: Integrity, confidentiality, and governance. Retrieved from https://doi-org.utk.idm.oclc.org/10.58012/638h-ab63

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction* 36 (6): 495–504.

Shneiderman, B. (2022). Human-centered AI, vol. 37. Oxford University Press.

Skitka, L.J., Mosier, K., and Burdick, M.D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies* 52 (4): 701–717.

Spring, J. M., Galyardt, A., Householder, A. D., & VanHoudnos, N. (2020, October 12–15). *On managing vulnerabilities in AI/ML systems*.

Proceedings of the New Security Paradigms Workshop 2020 (pp. 111–126).

Steyvers, M. and Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science* 19 (5): 722–734.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum* 56 (4): 24–31.

Stupp, C. (2019, August 30). *Fraudsters used AI to mimic CEO's voice in unusual cybercrime case*. The Wall Street Journal. Abstract retrieved from https://www-wsj-com.utk.idm.oclc.org/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402?reflink=desktopwebshare_permalink.

Tiwari, R. (2023). Explainable AI (XAI) and its applications in building trust and understanding in AI decision making. *International Journal of Scientific Research in Engineering and Management* 7: 1–13.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016, August 10–12). *Stealing machine learning models via prediction APIs*. Proceedings of the 25th USENIX Security Symposium (pp. 601–618).

Ulrich, K.T. and Eppinger, S.D. (2016). Product design and development. McGraw-hill.

Vallati, M., & Chrpa, L. (2023, October 1–4). *In Defence of good old-fashioned artificial intelligence approaches in intelligent transportation systems*. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 4913–4918.

van der Bijl-Brouwer, M. and Dorst, K. (2017). Advancing the strategic impact of human-centred design. *Design Studies* 53: 1–23.

Wang, D., Bian, C., and Chen, G. (2024). Using explainable AI to unravel classroom dialogue analysis: Effects of explanations on teachers' trust, technology acceptance and cognitive load. *British Journal of Educational Technology* 55 (6): 2530–2556.

Yang, W., Wei, Y., Wei, H. et al. (2023). Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric*

Intelligent Systems 3 (3): 161–188.

Zerilli, J., Bhatt, U., and Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns* 3 (4).

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). *Understanding deep learning requires rethinking generalization*. arXiv *preprint arXiv:1611.03530*.

Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., & Song, D. (2020, June 14–19). *The secret revealer: Generative model-inversion attacks against deep neural networks*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 253–261).