COZEN O'CONNOR CYBER LAW MONITOR
Following current trends in cyber, privacy, and data security law

NIST Issues New Artificial Intelligence Risk Management Framework

Christopher Dodson | cdodson@cozen.com

The National Institute of Standards and Technology (NIST) recently released version 1.0 of its Artificial Intelligence Risk Management Framework. The framework is available at https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, and a full set of supporting documents is available at https://www.nist.gov/itl/ai-risk-management-framework.

There is an emerging consensus that AI systems present a significantly different risk profile than conventional information technology systems. While there is currently no legal requirement to use a risk management framework when developing AI systems, there are a growing number of proposals that would require the use of a risk management framework or offer a safe harbor from certain types of liability if one is used.

The framework identifies 6 factors for mitigating risk and evaluating the trustworthiness of an artificial intelligence (AI) system.

Validity and Reliability

The results of the AI system must be close to true values. The system must be robust and maintain its performance under a variety of circumstances. Robustness requires not only that the system perform correctly under expected uses, but also that it should perform in ways that minimize potential harms to people if it is operating in an unexpected setting. A deployed system should be subject to ongoing testing to confirm the system is performing as intended. AI risk management efforts may need to include human intervention in cases where the AI system cannot detect or correct errors.

Safety

AI systems should not endanger human interests, including life, health, property, or the environment. Safety is generally enhanced through factors such as (i) responsible design and development; (ii) providing clear information to system implementers on responsible use of the system; and (iii) providing explanations and documentation of risks based on empirical evidence of incidents. Different types of safety risks may require AI risk management approaches tailored to context and the severity of potential risks.

Addressing safety considerations should begin as early as possible in the AI lifecycle to prevent conditions that could render a system dangerous. Other practical approaches for AI safety may use rigorous simulations, in-domain testing, and real-time monitoring, and include building in the ability to shut down, modify, or allow for human intervention into systems that deviate from expected functionality.

AI safety risk management approaches should take cues from safety guidelines in fields such as transportation and healthcare, and align with existing sector- or application-specific guidelines.

Security and Resilience

Security and resilience are related but distinct characteristics.

AI systems are resilient if they can maintain their function or, if necessary, degrade safely, if faced with adverse events or unexpected changes in their environment or use. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure. Common security concerns include data poisoning (the intentional manipulation of information used in the training data), and the exfiltration of models and training data through AI system endpoints. Other considerations include the provenance of training data (such as whether the developer has sufficient rights to use the training data), and the ability to withstand misuse or unexpected or adversarial use of the system.

Accountability and Transparency

Transparency is the extent to which information about an AI system and its output is available. Meaningful transparency provides access to appropriate information tailored to the role or knowledge of individuals interacting with the AI system. By promoting higher levels of understanding, transparency increases confidence in the AI system. When the potential negative consequences of an AI system are severe, such as when life and liberty are at stake, AI developers should consider proportionally increasing their transparency and accountability practices.

Maintaining information about the provenance of training data and supporting attribution of the AI system's decisions to subsets of training data can assist with providing transparency and accountability.

Explainability and Interpretability

Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes. Perceptions of negative risk often arise from not having the ability to make sense of, or contextualize, system output appropriately. Therefore, explainable and interpretable AI systems offer information that will help end users understand the potential impact of an AI system.

Risk from lack of explainability may be managed by describing how AI systems function, with descriptions tailored to individual differences such as the user's role, knowledge, and skill level.

Explainable systems can be debugged and monitored more easily, and they lend themselves to more thorough documentation, audit, and governance. Risks to interpretability often can be addressed by communicating a description of why an AI system made a particular prediction or recommendation.

Transparency, explainability, and interpretability are distinct characteristics that support each other. Transparency can answer the question of "what happened" in the system. Explainability can answer the question of "how" a decision was made in the system. Interpretability can answer the question of "why" a decision was made by the system and its meaning or context to the user.

Privacy

Privacy values, such as anonymity, confidentiality, and control, should guide decisions for AI system design, development, and deployment. Privacy-related risks may influence security, bias, and transparency and come with tradeoffs with these other characteristics. Like safety and security, specific technical features of an AI system may promote or reduce privacy. Design choices and data minimizing methods like de-identification and aggregation can enhance privacy in AI systems. Under certain conditions, such as data sparsity, privacy enhancing techniques can result in a loss in accuracy, affecting decisions about fairness and other values in certain domains.

Fairness with Harmful Bias Managed

NIST describes fairness in AI as including concerns for equality and equity by addressing issues such as harmful bias and discrimination. NIST notes, however, that standards of fairness are difficult to define because perceptions of fairness differ among cultures and may shift depending on application. Systems in which harmful biases are mitigated are not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals with disabilities or affected by the digital divide or may exacerbate existing disparities or systemic biases.

Bias is broader than demographic balance and data representativeness. NIST identifies three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of discriminatory intent.

Systemic bias can be present in AI datasets, the organizational practices and processes across the AI lifecycle, and the broader society that uses AI systems. Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples. Human-cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human-cognitive biases are inevitably present, even if unintentional, in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI. Indeed, the Federal Trade Commission (FTC) recently highlighted concerns about bias in a report, available at https://www.ftc.gov/reports/combatting-online-harms-through-innovation. The FTC's report includes an analysis of why AI tools may produce unfair or biased results. It also includes examples where, in the FTC's view, use of AI tools has resulted in discrimination against protected classes of people or blocked content in ways that negatively impacted freedom of expression.

There is increasing consensus that AI developers should use a risk management framework when building their AI systems. This will help developers comply with the evolving regulatory frameworks and mitigate the risk of potential lawsuits. NIST has offered a reasonable version 1 framework for developers to consider.