

# Responsibly Harnessing the Power of AI

Chantell Murphy<sup>i</sup> and Jonathan Barr<sup>ii</sup>

## Abstract:

Artificial Intelligence (AI) based applications are on the cusp of offering the public and private sectors tools of tremendous potential that will likely transform the world in much the same way that previous technological revolutions have. To harness these tools, a vast and rapidly increasing assortment of ethical AI guidelines and principles are being developed to manage the myriad of risks these tools pose. Documents, frameworks, standards, and regulations have been developed by private industry, research institutions, governments and nongovernmental organizations, international standards bodies and more. This work details representative approaches to ethically developing, managing, and operating current AI technologies. The various approaches are analyzed for commonality, divergence, and implementation strategy to help develop approaches for managing AI tools in the nuclear landscape. This work will support the creation of a common vocabulary and a deeper understanding of the ethical/ responsible AI landscape and advance the considerations required for structured frameworks in the international safeguards domain.

**Keywords:** nuclear safeguards, artificial intelligence, ethical AI, structured frameworks.

## Introduction

Artificial Intelligence, or AI, refers to technologies that leverage computers and algorithms to analyze a given environment and make deductions and/or take actions based on that environment, and have been developed and evolving for many decades. These “learn by example” technologies and applications include pattern recognition, predictive analytics, computer vision, natural language understanding, and speech recognition, among others. Many AI tools used today employ a combination of “handcrafted knowledge” systems, where computer scientists capture specialized knowledge in rules that the system applies to situations of interest, and more sophisticated large-scale statistical Machine Learning (ML) that enables engineers to create models that can be trained to specific problem domains if given exemplar data or simulated interactions. Learning from data, these systems are designed to solve specific tasks and achieve particular goals with competencies that, in some respects, parallel the cognitive processes of humans: perceiving, reasoning, learning, communicating, deciding, and acting [1]. These human-like capabilities are embedded in everyday objects like smartphones, vacuums, and cars. Smartphones feature virtual assistants like Siri and Alexa, photo tagging applications through social media platforms like Instagram and Facebook, and phones can be unlocked using facial recognition security. At a larger scale, AI is helping predict pandemic outbreaks, monitor traffic, speed up drug and therapeutic discovery, and automating routine office functions [2]. The technology is changing rapidly, increasing effectiveness, capacity and capabilities. Big tech companies and States are at the forefront of development and use because of the high infrastructure and resource costs needed for large systems using high performance computing

and massive supercomputers, but that is rapidly changing as well. Inexpensive cloud computing and open-source applications and development tools help democratize this powerful technology by expanding AI-enabled capabilities across the world.

The automation and efficiency gain brought by the AI-based or automation-based capabilities are also bringing unintended and unethical consequences by, for example, offering government agencies and corporations more powerful ways to collect and process information, track individuals' behavior and movements, circumvent regulations, and act on the basis of computer-generated analyses [2]. Effective ethical AI frameworks can help stakeholders identify critical needs in key ethical principles and values; develop strategies to address those needs; and ensure the appropriate mitigation efforts remain effective over time. The right ethical AI framework will help stakeholders build trust in the AI system.

The field of Ethical AI goes beyond legal regulations and system reliability controls by prioritizing and respecting fundamental human and societal values. Similar to the environmental studies concept of *urgent governance* [3], which describes the difference between auditing for system reliability versus auditing for societal harm, the processes would evaluate and optimize for social good and values rather than typical performance metrics such as accuracy or profit. For example, hydropower is listed as a reliable clean energy source, powering 31.5% of total U.S. renewable electricity generation and about 6.3% of total U.S. electricity generation [4]. However hydropower dams also destroy wildlife habitat, accelerate drought and desertification, and displace local and indigenous communities. The ultimate goal is to design and deploy systems that maximize benefits to humans and reduce harm, and it can be complicated to get it right.

Defining social values and benefits differ across cultures, communities, and time. Each group must prioritize their set of principles that align most with the values of their culture, State, or organizations. International organizations like the United Nations and the International Atomic Energy Agency (IAEA) must take all cultures and perspectives into account. To address such needs, the Institute of Electrical and Electronics Engineers (IEEE) Global Initiative's Ethically Aligned Design explored two thousand years' worth of established ethics systems, including Buddhist, Ubuntu, and Shinto-inspired ethics, and converged around the general principles of human rights, well-being, accountability, transparency, and awareness of misuse [5]. Defining the appropriate set of principles is the first step in establishing an ethical AI foundation, the next step is to put those principles into practice. AI technologies have the potential to enrich the human experience and to provide enormous benefits to society if they are designed, deployed, and used ethically [2].

## Need for Ethical Guidelines and Oversight

AI is uniquely suited to being a transformative force having profound impact across various societal domains. Because of its capabilities, AI is quickly becoming the next great power pursuit; world leaders are entering a phase of strategic AI competition, "The technologies will be the foundation of the innovation economy and a source of enormous power for countries that harness them. AI will fuel competition between governments and companies racing to field it. And it will be employed by nation-states to pursue their strategic ambitions" [2].

The anticipated power and capability bestowed upon AI technologies require careful consideration on how the tools are designed, built, operated, and understood; several concerns stem from how AI systems work and how they are developed. Because of the opaque nature of AI (e.g. machine learning, deep learning) and the preference for usability over interpretability it can be challenging to understand why a decision was made by the system, if it is correct, and how to verify it [5]. Additionally, the most powerful and pervasive AI tools have been and will likely continue to be developed by a very small set of private sector actors whose oversight, motivations, and goals for these systems may not always be transparent. Even in ideal circumstances, the developers of these systems might not understand how the underlying model works or understand what the system is fully capable of, including the myriad of ways it can be used once released. There is a desire to ensure AI is built, used, and retired in ways that uphold and prioritize ethical principles. While there are many framings to discuss ethical principles, this paper will focus on some of the more universal ethical considerations of transparency, bias, privacy, and accountability. The focus is not to create an exhaustive list of concerns, but rather to describe these ethical concepts and provide illustrative examples for context.

A fundamental concern is that AI operation must be appropriately transparent to the variety of stakeholders associated with a system. Transparency for AI (including the concepts of traceability, explicability, and interpretability) refers to the ability to understand how and why a system made a particular decision or took a particular action [5]. Financial “flash crashes” are events associated with a lack in transparency for why a decision is made and how one system impacts others. One such example is the British flash crash of 2016 [6]. On October 7, 2016, AI based trading algorithms began making financial decisions that dropped the value of the British pound (compared to the US dollar) 6% in two minutes. These financial algorithms are designed to examine a variety of sources, such as news releases and social media, in near real-time order to quickly interpret information and make quick financial decisions in order to outperform competitors. It is speculated in this instance that an algorithm reacted to comments made by the French President Francois Hollande regarding the United Kingdom paying a heavy price for pursuing Brexit.

The complexity of machine learning algorithms also pose a challenge around bias instantiated in the systems’ algorithms. Algorithmic bias can arise from decisions made in the system design process or biases present in system training and can result in systematic prejudice during the operation of the system. Algorithmic bias can occur if the training data is not representative of the situations and groups that a system will be required to operate in and with, respectively. Facial recognition algorithms are a representative example of AI tools that have historically been associated with this type algorithmic bias. The seminal “Gender Shades” project revealed substantial disparities in the accuracy of classifying darker-skinned females (with error rates of up to 34.7%), lighter-skinned females, darker-skinned males, and lighter-skinned males (0.8%) in three different gender classification systems [7]. In 2016 Buolamwini and Gebru warned “someone could be wrongfully accused of a crime based on erroneous but confident misidentification of the perpetrator from security video footage analysis”, which happened to Robert Williams [8], Michael Oliver [9], Nijeer Parks [10], and Randall Reid [11]. Alternatively, a system can intentionally or unintentionally perpetuate and reinforce biases if it is trained on data that has captured human discriminatory practices. Famously, an online retailer created a hiring software that discriminated against women because it was trained on their historical hiring data, which reflected a highly male dominated tech industry. The

company recognized the error, tried and failed to fix it, and ultimately had to discard the software [12].

Privacy is a major ethical concern for AI systems due to how these systems are trained and the powerful insights that they are capable of providing. Data collection of personal information is a powerful source of information for training AI systems to support making accurate predictions and recommendations. Collecting and storing personal information is not a new phenomenon, and has historically required measures be put in place to mitigate misuse, unauthorized access, and data breaches. Beyond traditional data privacy concerns, AI systems can analyze behavior to profile an individual to make predictions and decisions of that individual without their knowledge or consent. For example in 2012, a major US retailer used predictive analytics to identify various classes of customers for targeted advertising [13]. A teenage customer was categorized as pregnant based on shopping patterns and pregnancy related coupons were mailed to her family home, effectively notifying her parents of the pregnancy. As previously mentioned, the applications of AI tools might change as the capabilities of the system are explored and put to new uses. This can mean that an individual may have consented to providing information to one system, but the model built from that system might have value elsewhere and be repurposed without the individual's awareness. This returns the discussion to facial recognition, which can be used to accomplish many functions (e.g. searching for friends in photos, biometrics for phones, etc). Interestingly, half of American adults have photos in a facial recognition network used by law enforcement without the individuals' consent or awareness [11]. Finally, there is concern that AI systems that have taken precautions to de-identify or anonymize personal information may be susceptible to algorithms that reverse these processes to access the model and the underpinning personal information.

The last of these ethical principles is accountability. Accountability refers to the responsibility of the various AI stakeholders at the different stages of the system's lifecycle (development, deployment, operation, retirement). Accountability considers legal and regulatory obligations, mechanisms to detect, track and manage issues or errors, and ensuring responsible development and use. Without accountability individuals, community, and society writ large are left feeling helpless and subject to the whims of decision makers without any recourse. An example is the case of an AI system that imposed stricter jail sentences on Black defendants. This system is widely used to assess the risk of recidivism for defendants in pretrial hearings. In 2016, ProPublica investigated how the system was being used in Broward County, Florida [14]. Their analysis revealed that even though the system predicted recidivism equally well for white and black defendants, it made different kinds of systematic mistakes for the two populations. The system was more likely to mistakenly predict that black defendants were high-risk, while making the opposite type of mistake for white defendants. Despite this finding, the developers insisted the algorithm operated as expected, and no changes were made to the system. The lack of standard processes for mitigating algorithmic bias allowed the company to define their requirements for fairness, and the lack of mechanisms for holding stakeholders accountable means that the system is still allowed to be in use without any consequences to the developers or the courts.

Similar to the maturation of other disruptive technologies, AI capabilities have outpaced the risk mitigation strategies that aim to allow for the productive use of these tools while minimizing the harm they are capable of. However, there are many promising approaches being pursued that can support industry, developers, AI adopters, end users, and the public in ensuring that these powerful AI tools can be safely and ethically developed and operated.

## Mitigation Approaches and Perspectives

The previous section details the rationale behind a rising concern for the societal implications of AI systems. This in turn has inspired a wave of ethical AI principles and guidance documents. In 2019, Jobin et al identified 84 published documents containing ethical principles or guidelines for AI [15]. As the concept of ethical AI has matured, certain ethical principles related to AI have begun to coalesce. Harvard’s Berkman Klein Center for Internet and Society studied initiatives across cultures and countries, authored by governments and intergovernmental organizations, companies, professional associations, advocacy groups and multi-stakeholder initiatives, analyzing and comparing thirty-six prominent AI principles documents [16]. While researchers observed variation across different dimensions, they were surprised to find consensus emerge around: *Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and non-discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values* (Table 1). This table provides an overview of the range of considerations that have consensus and would be included in a comprehensive ethical framework.

**Table 1.** Ethical AI themes and associated ethical principles in 36 ethical AI documents [16].

<p><b>Privacy</b></p> <ul style="list-style-type: none"> <li>• Consent</li> <li>• Control over the use of data</li> <li>• Ability to restrict processing</li> <li>• Right to rectification</li> <li>• Right to erasure</li> <li>• Privacy by design</li> <li>• Recommends data protection laws</li> </ul>	<p><b>Accountability</b></p> <ul style="list-style-type: none"> <li>• Verifiability and replicability</li> <li>• Impact assessments</li> <li>• Environmental responsibility</li> <li>• Evaluation and auditing requirement</li> <li>• Creation of a monitoring body</li> <li>• Ability to appeal</li> <li>• Remedy for automated decision</li> <li>• Liability and legal responsibility</li> <li>• Recommends adoption of new regulations</li> <li>• Accountability per se</li> </ul>	<p><b>Safety and Security</b></p> <ul style="list-style-type: none"> <li>• Security by design</li> <li>• Predictability</li> </ul>	<p><b>Transparency and Explainability</b></p> <ul style="list-style-type: none"> <li>• Open source data and algorithms</li> <li>• Open government procurement</li> <li>• Right to information</li> <li>• Notification when AI makes a decision about an individual</li> <li>• Notification when interacting with AI</li> <li>• Regular reporting</li> </ul>
<p><b>Fairness and Non-discrimination</b></p> <ul style="list-style-type: none"> <li>• Prevention or reduction of bias</li> <li>• Representative and high quality data</li> <li>• Fairness</li> <li>• Equality</li> <li>• Inclusiveness in impact</li> <li>• Inclusiveness in design</li> </ul>	<p><b>Human control of technology</b></p> <ul style="list-style-type: none"> <li>• Human review of automated decision</li> <li>• Ability to opt out of automated decisions</li> </ul>	<p><b>Professional responsibility</b></p> <ul style="list-style-type: none"> <li>• Accuracy</li> <li>• Responsible design</li> <li>• Consideration of long-term effects</li> <li>• Multi-stakeholder collaboration</li> <li>• Scientific integrity</li> </ul>	<p><b>Promotion of human values</b></p> <ul style="list-style-type: none"> <li>• Human values and human flourishing</li> <li>• Access to technology</li> <li>• Leveraged to benefit society</li> </ul>

While common themes around principles have emerged, the question of how to operationalize these principles remains a challenge. While not compulsory, well-developed and understood standards,

training, and policies help to ensure risks are mitigated and ethical concerns are managed. When standards, training, and policies are not sufficient, then stricter rules and regulations with enforcement mechanisms and consequences are created. The ethical commitments made so far have not been backed by strong oversight and accountability, and lack the teeth to hold organizations accountable without government regulation [17, 18]. AI statements provide top cover to organizations to demonstrate a commitment to ethical development, but unless ethical principles are embedded in the culture of an organization (which an ethical AI framework supports) and are actively enforced, they are unlikely to significantly impact decision-making [19].

Many of these principles can be considered *essentially contested concepts* – abstract concepts like fairness and dignity with different and potentially conflicting meanings that are interpreted through one’s beliefs, leading to different requirements in practice. States are likely to approach ethical AI principles through their own cultural lens and driven by their own interests; however, global deployment of AI technology will require a common approach and will drive the development of an acceptable path. According to Dr. Eric Horvitz, former Director of Microsoft Research, there is a challenge between the precision of developing AI and the broader interpretations of values. Horvitz notes that “if we get this right” there will be training, awareness, guidance, and reporting for AI development and use [20].

A thematic analysis conducted by Jobin et al analyzed current principles and guidelines for ethical AI to identify converging themes and determine where divergence is occurring [15]. Their thematic analysis revealed divergences in how the identified ethical principles are interpreted, handled, and how areas of concern are derived from each principle. For example, Jobin et al identified variations for the theme of transparency for “the interpretation, justification, domain of application, and mode of achievement” [15]. For the theme of justice, the private sector placed less emphasis on discrimination, and more on fairness and bias. Vastly different interpretations of justice existed across organizations where some viewed justice as pertaining to diversity, inclusion, and equality, whereas others viewed justice to be about the ability to appeal or challenge decisions, or the right to redress or remedy. Divergence also exists with public sector documents focusing more on AI impacts on employment and the need to address societal issues as compared to private sector documents.

Jobin et al identified substantive divergence among the main ethical principles relating to four major factors: the interpretation of ethical principles; rationale for inclusion; identification of domains/stakeholders; and implementation strategies. The authors note that these divergences, which affect the resolution of conflicts between ethical principles, may also impact the development of a global agenda for ethical AI [15].

### Implications and opportunities for the nuclear community

Organizations can define and incorporate ethical principles and values into their culture in a variety of ways. The mechanisms for turning principles into practice also vary, but historically there has been a demonstrated benefit for adopting frameworks aligned with global norms [5]. Consistency across organizations creates norms that can be measured and interpreted by the broader community. For example, global norms provides a system of understanding such that if a vender in Brazil sells their AI tool to a school in Mexico, both the developer and the user will have the same approach to identify,

alert, and mitigate issues that threaten transparency or result in algorithmic bias, etc. This approach simplifies the verification system to an agreed upon set of ethical metrics, as opposed to multiple sets that will need to be cross referenced, understood, and compared.

The normative approach might be challenging to adopt as some Governments are already starting to adopt ethical AI frameworks into their own structures. China is the first country to incorporate a responsible AI framework into its governance framework, *Governance Principles for a New Generation of Artificial Intelligence* [21, 22], the EU is actively developing a regulatory regime, *The Artificial Intelligence Act* [23], and the United States issued a non-binding *Blueprint for an AI Bill of Rights* that provides a framework for how government, technology companies, and citizens can work together to ensure more accountable AI [24]. Incorporating these frameworks into governance structures where AI technologies will be regulated in different ways will impact how AI tools can be developed, deployed, and used around the world.

However, adopting a normative approach to ethical AI frameworks could work well for international organizations, like the United Nations and the International Atomic Energy Agency, who must adopt a more global approach to ethics and principles. The European Commission High-Level Expert Group on Artificial Intelligence developed their own ethical AI framework, the *Ethics Guidelines for Trustworthy AI*, to support the vision of “ethical, secure and cutting-edge AI made in Europe”, but the guidelines are addressed to all AI stakeholders designing, developing, deploying, implementing, using or being affected by AI [25]. The IEEE Standards Association uses a global participation approach to develop applied ethics resources and offers standards, training and education, certification programs, and more, to empower stakeholders designing, developing, and using AI [26]. Such an egalitarian approach may be quite suitable for international organizations.

Concerns around the ethical themes of transparency, algorithmic bias, privacy, and accountability are relevant to the safeguards community. Ethical AI frameworks can help the IAEA navigate potential universal and domain specific ethical questions. Ethical transparency may arise when using AI technologies in Member State facilities; they may want to know that AI is being used and how it is being used. The system should also be transparent internally at the IAEA; is the model performing as anticipated? Algorithmic bias questions may arise when an AI system is trained on historic proliferation data and used to assess other States, or when algorithms are trained in States with a mature nuclear fuel cycle versus a nuclear newcomer, or those with less capabilities. Ethical AI frameworks can help navigate non-technical bias concerns by ensuring underrepresented cultural norms, interactions, and work structures are not omitted, miscategorized, or misunderstood in models and outcomes. Privacy issues may come into play when AI is used to analyze surveillance footage; should facial recognition technology be used, would people know it is being used, is it tracking and identifying all people entering a facility or given boundary? Finally, ethical AI frameworks can help to clearly identify the chain of accountability for a given AI technology and application. Accountability will identify responsible parties associated with AI operation in or for a facility and provide appropriate mechanisms for redress; properly developed accountability for AI can mitigate risks associated with deploying autonomous verification tools in facilities and creating clear lines of responsibility when something goes wrong (e.g., the facility added a new feature that the tool was not

trained on, the tool does not operate as expected, the operator was using the tool outside of intended parameters, etc).

Ethical AI frameworks can promote a culture around ethical practices by helping developers and users ensure algorithms, use cases, and data adhere to the values and principles of the nuclear community. Ethical AI frameworks help ensure potential consequences are identified and considered, so mitigation strategies can be developed preemptively. They help ensure transparency measures are put in place, such as documentation, audit trails, and information on the composition of training and benchmark datasets; additionally, transparency and accountability can also include mechanisms for consent and redress. Because of the rapid pace of development of AI technology, ethical decommissioning processes may help ensure the privacy of Member State data from cradle to grave; similar processes are not unique to AI and should already exist at the IAEA. And particularly for the IAEA, ethical AI provides consistency among high turnover research groups and analysts, and changing leadership styles [27]. Work continues unpacking the details of how the myriad of ethical frameworks can mitigate these concerns and determine how they perform across the globe. Questions remain about the adjustments companies and organizations are willing to make to meet ethical requirements and the types of gain they may make in return.

## References

- [1] Defence Advanced Research Projects Agency (DARPA), "A DARPA Perspective on Artificial Intelligence," [Online]. Available: <https://www.darpa.mil/attachments/AIFull.pdf>.
- [2] The National Security Commission on Artificial Intelligence, "National Security Commission on Artificial Intelligence Final Report," 2021. [Online]. Available: <https://reports.nscai.gov/final-report/>.
- [3] A. H. Lynch and S. Veland, *Urgency in the Anthropocene*, The MIT Press, 2018.
- [4] Department of Energy Office of Energy Efficiency & Renewable Energy, "Hydropower Basics," [Online]. Available: <https://www.energy.gov/eere/water/hydropower-basics>. [Accessed 19 April 2023].
- [5] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Version 2," IEEE, 2017.
- [6] Buttonwood, "Sterling takes a pounding," 7 October 2016. [Online]. Available: <https://www.economist.com/buttonwoods-notebook/2016/10/07/sterling-takes-a-pounding>. [Accessed 8 May 2023].
- [7] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of Machine Learning Research*, 2018.
- [8] J. Bhuiyan, "First man wrongfully arrested because of facial recognition testifies as California weighs new bills," 27 April 2023. [Online]. Available: <https://www.theguardian.com/us-news/2023/apr/27/california-police-facial-recognition-software>. [Accessed 24 April 2023].
- [9] E. Stokes, "Wrongful arrest exposes racial bias in facial recognition technology," 19 November 2020. [Online]. Available: <https://www.cbsnews.com/news/detroit-facial-recognition-surveillance-camera-racial-bias-crime/>. [Accessed 24 April 2023].



- [10] K. Johnson, "How Wrongful Arrests Based on AI Derailed 3 Men's Lives," 7 March 2022. [Online]. Available: <https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/>. [Accessed 24 April 2023].
- [11] K. Hill and R. Mac, "'Thousands of Dollars for Something I Didn't Do'," 31 March 2023. [Online]. Available: <https://www.nytimes.com/2023/03/31/technology/facial-recognition-false-arrests.html>. [Accessed 24 April 2023].
- [12] J. Daston, "Amazon scraps secret AI recruiting tool that showed bias against women," 10 October 2018. [Online]. Available: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>. [Accessed 8 May 2023].
- [13] C. Duhigg, "How Companies Learn Your Secrets," 16 February 2012. [Online]. Available: <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>. [Accessed 24 April 2023].
- [14] J. Angwin, J. Larson, S. Mattu and L. Kirchner, "Machine Bias," 23 May 2016. [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. [Accessed 25 April 2023].
- [15] A. Jobin, M. Ienca and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, p. 389–399, 2019.
- [16] 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.
- [17] M. Whittaker, K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. Myers West, R. Richardson, J. Schultz and O. Schwartz, "AI Now Report 2018," AI Now Institute, New York City, 2018.
- [18] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, p. 501–507, 2019.
- [19] J. C. Newman, "Decision Points in AI Governance," UC Berkeley Center for Long-Term Cybersecurity, 2020.
- [20] National Security Commission on Artificial Intelligence, *Ethical and Responsible AI: International Standards*, National Security Commission on Artificial Intelligence, 2021.
- [21] C. Li and L. Yang, "Responsible AI: The Revolution in Governance Technology in China," in *2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)*, Dali, China, 2021.
- [22] Library of Congress, "China: AI Governance Principles Released," 9 September 2019. [Online]. Available: <https://www.loc.gov/item/global-legal-monitor/2019-09-09/china-ai-governance-principles-released/>. [Accessed 5 May 2023].
- [23] Future of Life Institute, "The Artificial Intelligence Act," [Online]. Available: <https://artificialintelligenceact.eu/the-act/>. [Accessed 5 May 2023].
- [24] The White House Office of Science and Technology Policy, "Blueprint for an AI Bill of Rights," October 2022. [Online]. Available: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>.
- [25] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy Artificial Intelligence," European Commission, 2019.

- [26] IEEE SA, "Autonomous and Intelligent Systems (AIS)," IEEE, [Online]. Available: <https://standards.ieee.org/initiatives/autonomous-intelligence-systems/>. [Accessed 5 May 2023].
- [27] C. Murphy and J. Barr, "Validating Ethical AI Frameworks for International Safeguards," Consolidated Nuclear Security, LLC, 2023.

#### DISCLAIMER

This work of authorship and those incorporated herein were prepared by Consolidated Nuclear Security, LLC (CNS) as accounts of work sponsored by an agency of the United States Government under Contract DE-NA-0001942. Neither the United States Government nor any agency thereof, nor CNS, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility to any non-governmental recipient hereof for the accuracy, completeness, use made, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency or contractor thereof, or by CNS. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency or contractor (other than the authors) thereof.

#### COPYRIGHT NOTICE

This document has been authored by Consolidated Nuclear Security, LLC, under Contract DE-NA-0001942 with the U.S. Department of Energy/National Nuclear Security Administration, or a subcontractor thereof. The United States Government retains and the publisher, by accepting the document for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this document, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, or allow others to do so, for United States Government purposes.

---

<sup>i</sup> Consolidated Nuclear Security, LLC. Y-12 National Security Complex.

<sup>ii</sup> Pacific Northwest National Laboratory