Introduction to The Special Section on Safe AI

Mykola Pechenizkiy
Department of Computer Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
m.pechenizkiy@tue.nl

Stiven S. Dias
Embraer S.A.
R&D Team
Av. Brigadeiro Faria Lima, 2170
São José dos Campos, Brazil
stiven.dias@embraer.com.br

ABSTRACT

Alignment with human values and compliance with ethical and regulatory standards has become a major concern as cutting-edge generative AI-based solutions are becoming ubiquitous and have been increasingly adopted as a productivity tool by many institutions. Despite their remarkable results across several fields, there are still objections regarding embedding AI models, even the vanilla ones, into safe-critical systems, specially, across healthcare, aeronautical and nuclear industries, as many AI methods lack explainability and robustness, and are still prone to adversarial and cyber attacks. The Safe AI field addresses these issues through the proposal of design and development procedures aimed at learning assurance and model alignment, the investigation of interpretable methods with explainable outputs, and the proper treatment of uncertainty. We summarize first the outcomes of a recently launched workshop on Safe AI and present then five selected papers in this special section representing a crosscut of different application areas for AI safety.

1. INTRODUCTION

The advent of generative AI has unlocked many applications, promoting more convenience in everyday tasks and leveraging productivity in labor activities. Conversely, it is widely recognized that generative models may hallucinate and manifest unfairness induced by, e.g., unwanted biases present in the training data. Ultimately, they may fail to the point they fully lack alignment with human values and fundamental ethical principles.

Deep models in general have also gained much attention in the industry, as they are able to perform many challenging tasks with human-like and even super-human performance. However, the use of AI models in safety-critical applications are still controversial, since it is hard to audit them and demonstrate they are trustworthy, behaving consistently and reliably as expected by their stakeholders.

There are multiple initiatives towards AI safety. It is notable, that over the past few years, an international network of AI Safety Institutes has emerged. Starting with UK AI Safety Summit in November 2023, where the concept of AI Safety Institute was introduced through the Bletchley Declaration, the network started with the AI Safety Institute (AISI)[1] (later renamed into AI Security Institute) it grew to include the North American Center for AI Standards and Innovation (CAISI)[2], the European Union AI Office[3] and national, e.g., Spanish Agency for the Supervision of Artificial Intelligence[4] institutes. As of today, the network includes variants of AISI in India, Singapore, South Korea, Japan, and Israel. More countries announce similarly intended initiatives.

Government organizations aim to bridge the knowledge gap between rapidly evolving AI technology and public sector understanding by providing technical expertise to inform policy decisions. Thus, the CAISI in the US focuses on working with NIST organizations to develop guidelines and best practices to measure and improve the security of AI systems, and to develop voluntary standards. The EU AI Office supports development of AI safety standards and Codes of Practice, particularly in support of the EU AI Act. Besides government-backed AI safety institutes, independent non-profits – e.g., Center for AI Safety (CAIS)[5] –, and international collaborations – notably, the International AI Safety Report [6] – have emerged. Many leading AI companies – e.g., Anthropic, Google AI / Google DeepMind, OpenAI and Meta AI – have their own internal AI safety teams and initiatives.

Outline. We discuss the landscape of Safe AI applications and taxonomy in Section 2. Section 3 highlights a recently launched initiative on AI safety, the 1^{st} Workshop on Safe AI at UAI 2025 conference. In the sequel, in Section 4, we introduce five selected articles to represent the state-of-the-art in the field. Finally, we conclude the introduction in Section 5.

2. SAFE AI LANDSCAPE

Safe AI is the field focused on minimizing the risks associated with the operation of future AI systems. A comprehensive risk assessment, in turn, raises different concerns spanning from engineering and security aspects through ethical and governance considerations. In any case, the ultimate goal is to ensure that increasingly capable AI systems operate reliably, aligned with human values, and do not cause harm to the society if they unintentionally fail or in case they are deliberately misused.

[1] https://www.aisi.gov.uk/

[2] https://www.nist.gov/caisi

[3] https://digital-strategy.ec.europa.eu/en/policies/ai-office

[4] https://aesia.digital.gob.es/en/es

[5] https://safe.ai/

Figure 1 highlights the main AI safety principles that future-proof AI systems should follow. Alignment ensures that AI systems' outcomes contribute positively to human goals, avoiding unintended harm to society. Robustness ensures that AI systems perform consistently as designed in face of unforeseen or potentially harmful inputs. Transparency ensures that AI systems are understandable and auditable by humans, which is crucial for building trust and accountability. Accountability ensures that there are mechanisms to assign responsibility for the outcomes of AI systems.

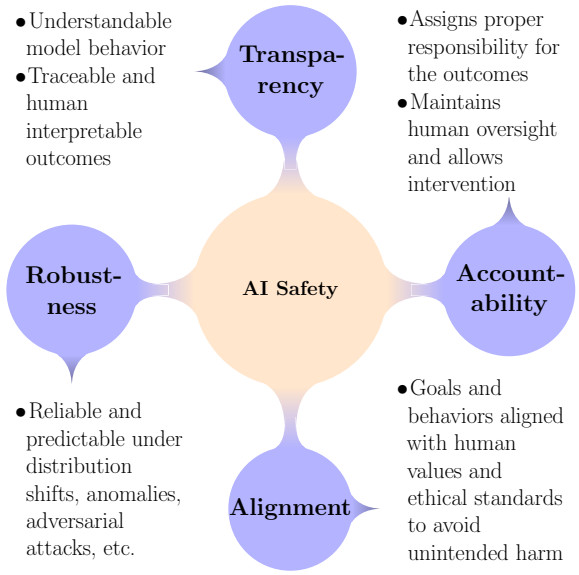

Figure 1: AI safety principles that future-proof AI systems should follow.

Much of progress in modern Machine Learning (ML) is attributed to testing multiple ideas and favoring most competitive through benchmarking. In the context of Safe AI, technical and normative concept and requirements are often translated into oversimplified intrinsic metrics. Thus, current fairness-aware machine learning benchmarking practices may result in suboptimal outcomes, undesirable side-effects, or misleading conclusions. Complementary evaluation frameworks [30,1] are advocated to address such limitations. Lastly, it is worth mentioning that verification of AI systems safety through formal system analysis has received increasing attention, see e.g. [27]. However, the practical relevance of the results obtained so far is not yet fully evident.
Major research areas within AI safety include:

- **Robustness and reliability** to ensure that AI systems will generalize satisfactorily, behaving dependably and consistently over time, given arbitrary inputs, including out of distribution (OOD) inputs, anomalies, adversarial attacks, distribution shifts, and concept drifts [32]. Controlling the accuracy and uncertainty trade-off can be useful in decision making [15]. Decidability and halting guarantees are also a concern [29].
- **Explainability and interpretability** to understand AI system (mis)behavior and its underlying working mechanisms, ensuring that the (causal) reasoning behind specific inferences are understandable to human beings [33].
- **Ethical alignment** to make sure that AI systems' goals and behavior are aligned with human values and intentions. Concerns range from non-discrimination [22] and fairness-awareness [10] to the raise of rogue AI agents [9].
- **Preventing misuse** is considered from the perspective of hypothetical superintelligence [2] to guidelines for improving the safety, security, and trustworthiness of dual-use existing and future foundation models [37] to prevent AI systems based on it from being weaponized to harm society.
- **Governance and assurance** are applied to all stages of the AI engineering lifecycle: developing, testing, verification & validation, deployment, and runtime monitoring. Governance frameworks encompass regulations, methods, procedures, and technological mechanisms to ensure that an organization's development and deployment of AI technologies align with its strategies, principles, and goals [28]. On the other hand, assurance processes ensure that AI technologies produce outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, yielding intelligent systems that are ethical in the context of their deployment, unbiased in their learning, and fair to their users [5].

The AI alignment and AI safety fields are inter-twisted. As argued by Ji et al. in [21], the former constitutes a significant portion of the later concerns. They share common AI risks from model misalignment and, as indicated in Figure 2, they encompass thus common goals / principles: Robustness, Interpretability, Controllability, and Ethicality (RICE). Much attention has been directed to the alignment of Large Language Models (LLMs) [34] and their vulnerability to jailbreaking [11]. However, in general, the so-called AI alignment cycle – comprising backward and forward alignment procedures – also serves as a framework to ensure that a broad spectrum of AI systems adhere to human intentions and values [24].
Explainable AI (XAI) has been a topic of extensive research for Deep Neural Networks (DNN) [38,20] and, more narrowly, for Natural Language Processing (NLP) [12]. The XAI community has been striving to explain models as black-boxes [13], driving less effort though toward designing white-box models. Despite yielding explainable outputs, linked to the model inputs by human interpretable causal chains, white-box or glass-box models are often not as powerful and fail to achieve state-of-the-art performance when compared to black-box ones [17]. On the other hand, as argued by Rudin in [31], instead of perpetuating the bad practice of explaining black-box models that can potentially harm the society, the path forward for Safe AI should be the development of inherently interpretable models. Finally, application-wise, it can be argued that the XAI goal is to ensure that the generation of the AI system's outputs is auditable [7] and, ultimately, liable [42].
AI governance and assurance [5] have been gaining increasing attention from the industry, as strictly regulated business activities require auditable development and deployment processes. In this sense, regulatory efforts like the European Union Aviation Safety Agency (EASA)'s AI roadmap [18,19], guiding the safe and trustworthy development and implementation of AI technologies in aviation, are crucial to promote

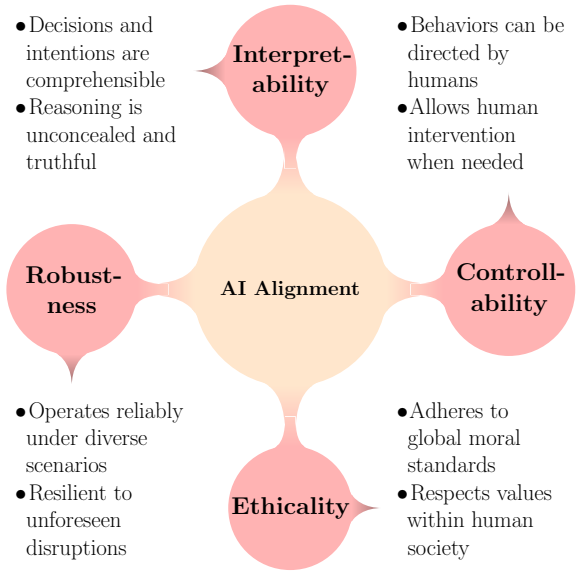

Figure 2: RICE principles that an aligned AI system should possess as introduced by Ji et al. in [21].

trustworthy AI in the industry. However, it is paramount that these recommendations and guidelines are also addressed by foundational research within the AI safety community.
Finally, the AI trustworthiness and AI safety domains overlap in many ways. However, the former is focused on ensuring that currently available AI systems are ethical, fair, transparent and reliable for today's society, whereas the later follows a risk-centric approach focused on preventing harmful, unintended, or catastrophic outcomes from increasingly capable models, pursuing thus longer timescales and an expanded risk framing. As stressed by Li et al. in their survey [23], AI trustworthiness must be systematically pursued throughout all life-cycle of an AI-based system. Figure 3 depicts the different aspects of trustworthy AI ranging from technical to ethical requirements. Li et al. also argued that an holistic assessment of these aspects is required to ensure the accountability of real-world AI systems, making sure they comply with all requirements and regulations. It is worth noting that performance (accuracy) requirements are important ingredients and trades-off with several requirements, but, from the authors' perspective, ethical requirements play a central role on the system life-cycle. In particular, they put accountability goals / aspects at the heart of the holistic assessment, raising thus the relevance of auditability, traceability and responsibility requirements.

3. 1ST WORKSHOP ON SAFE AI@UAI 2025

In this section, we summarize the 1^{st} Workshop on Safe AI held on July 25^{th}, as part of the 41^{st} Conference on Uncertainty in Artificial Intelligence (UAI 2025) in Rio de Janeiro, Brazil.
The workshop was conceived to foster technical discussions on cutting-edge safety techniques, including uncertainty quantification, adversarial robustness, and socio-technical discussions including AI alignment with human values, fairness and non-crimination among other important topics. Broadly, the workshop also aimed at discussing techniques that facilitate design and development of the future-proof AI.
The workshop's program included two plenary keynote talks, thirteen technical contributions selected by the program committee and presented by authors in the oral and poster sessions, and a closing discussion[6]. The keynote speakers covered topics related to AI uncertainty and AI governance as highlighted below:

- **Privacy-aware Bayesian Networks**. Professor Cassio de Campos introduced credal models as a practical solution for balancing privacy and utility. Credal models represent a set of standard precise models by having set-based parameter specifications. Then, he discussed how credal models can disguise the original model, thereby reducing the probability of successful attacks, while achieving meaningful inferential results. The versatility of the idea was illustrated through an experimental evaluation, comparing it with noise-based approaches. Lastly, the talk also stressed that proper treatment of model uncertainty is required for safe AI.
- **AI Security - Standards, STEADY and LASR**. In this talk, Professor Paul Miller provided an overview of the European Telecommunications Standards Institute (ETSI) approach to developing standards for AI security through security-by-design. Then, he discussed how him and his collaborators have been applying an AI safety and assurance methodology to a specific use-case related to autonomous ground vehicles. To summarize, the talk highlighted the need for AI governance frameworks to ensure that AI-based systems meet their purpose and specifications.

The technical papers presented at the workshop covered four thematic areas including different areas of XAI and uncertainty, AI fairness and alignment with human goals, AI robustness to adversarial attacks, and AI robustness in scenarios with AI agents.

XAI and uncertainty:

- **DT-sampler: A SAT-based Decision Tree Ensemble** by *Xiaotian Xue, Chao Huang, Koji Tsuda, Diptesh Das*, published in this special section [39].
- **SpaCE-VAE: Sparse and Confident Explanations using Variational Autoencoders** by *Alexander Liu, Sibylle Hess*, published in this special section [25].
- **Explaining Deep Learning Matching of Hand-Drawn Binary Symbols: A Visual Analysis with Grad-CAM on Cattle Brands** by *Leandra Soares, Marcos Medeiros, Aldo Díaz-Salazar, Edmundo Hoyle*, published in this special section [35].
- **A Non-Parametric Bayesian Approach Towards Online Sequence Learning** by *Stiven Schwanz Dias, Marcelo Gomes da Silva Bruno, Alberto Ferreira De Souza, Thiago Oliveira-Santos and Claudine Badue*, also published in this special section [16].

AI fairness and alignment with human goals:

[6]The full program and preprints of the papers not included in this special section can be found at `https://sites.google.com/view/safeai2025/`

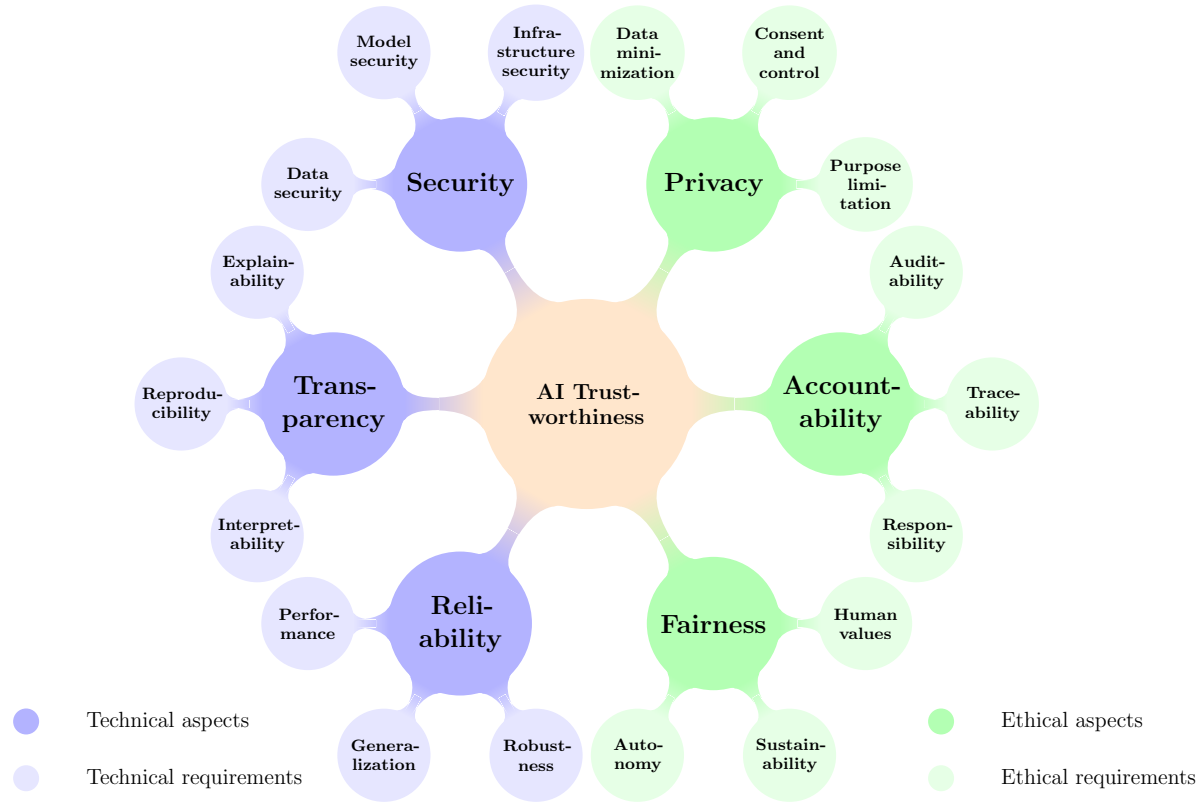

Figure 3: Non-exhaustive breakdown of AI trustworthiness aspects into requirements that existing AI systems should meet.

- **Robust Data Merging Strategies for Aerial Object Detection of Flying Objects** by *Luca Plaster, Aldo Díaz-Salazar*, the manuscript available on the workshop website as a full paper.
- **Alignment Metrics for the Evaluation of Observed Autonomous Systems** by *Tommaso Mannucci, Wouter Arink, Johan van den Heuvel*, also available on the workshop website as an extended abstract.

AI robustness to adversarial attacks:

- **ExpProof: Operationalizing Explanations for Confidential Models with ZKPs** by *Chhavi Yadav, Evan Laufer, Dan Boneh, Kamalika Chaudhuri*, available as a preprint on arXiv [40].
- **Influence Attributions can be Systematically Altered by Model Manipulation** by *Chhavi Yadav*, also available as a preprint on arXiv [41].
- **Riemannian Manifold Learning for Stackelberg Games with Neural Flow Representations** by *Larkin Liu, Kashif Rasul, Yutong Chao, Jalal Etesami*, an extended abstract available on the workshop website, and a preprint of the full paper on arXiv [26].

AI robustness in scenarios with AI agents:

- **An Analysis of Robustness of Non-Lipschitz Networks** by *Maria-Florina Balcan, Avrim Blum, Dravyansh Sharma, Hongyang Zhang*, published in the Journal of Machine Learning Research, volume 24, in 2023 [4].
- **The Value of Recall in Extensive-Form Games** by *Ratip Emin Berker, Emanuel Tewolde, Ioannis Anagnostides, Tuomas Sandholm, Vincent Conitzer*, published in the proceedings of the 39^{th} Conference on Artificial Intelligence (AAAI-2025) [8].
- **HASARD: A Benchmark for Vision-Based Safe Reinforcement Learning in Embodied Agents** by *Tristan Tomilin, Meng Fang, Mykola Pechenizkiy*, published in the proceedings of the 13^th International Conference on Learning Representations (ICLR 2025) [36].
- **Decision Making under Imperfect Recall: Algorithms and Benchmarks** by *Emanuel Tewolde, Brian Zhang, Ioannis Anagnostides, Tuomas Sandholm, Vincent Conitzer*, preprint available on the workshop website.
- **SafeFlowNet: Safe Control with Generative Flow Networks** by *Yucheng Yang, Tianyi Zhou, Meng Fang, Mykola Pechenizkiy*, preprint available on the workshop website.

The workshop promoted an opportunity to i) leverage the AI safety community, ii) promote the field among young researchers, iii) prospect which topics are being most investigated by the community and iv) which topics should be further explored by it. From our perspective, explainability and resilience to adversarial attacks are being the focus of many research efforts. On the other hand, as suggested by the position paper by Dr. Miller, there is room for the community to further investigate the AI governance ecosystem,

since heavily regulated industries such as the aeronautical one must demonstrate to certification authorities that complex systems, e.g., an aircraft, comply with strict norms and standards throughout the entire product lifespan.
Lastly, the workshop included full papers and extended abstracts. Some of the full papers were selected and polished to this special section as described in the following compendium.

4. CONTRIBUTED ARTICLES

This special section includes one position paper related to AI governance and assurance, based on the invited workshop keynote, along with four full papers presenting technical contributions:

Assuring the Case: A Safety Engineering Approach to AI-Enabled Systems [14] introduces an assurance framework for ML-based safety-critical systems. Specifically, the authors follow a safety engineering approach and employ a graphical, six-stage process for assurance case construction, culminating in the instantiation of the assurance argument. They demonstrate the assurance process considering a deep CNN image classifier for an autonomous vehicle application.

DT-sampler: A SAT-based Decision Tree Ensemble [39] considers the problem of creating a decision tree (DT) ensemble taking into account the size and accuracy of the sampled trees. The authors combine a SAT-based decision tree sampler to generate small-size DTs with the conformal prediction framework to post-calibrate the sampled trees. Their approach unlocks the path for more-explainable decision trees. Lastly, they demonstrate that their approach is more statistically efficient and provides stable predictions compared to random forests across several real-world benchmark datasets.

SpaCE-VAE: Sparse and Confident Explanations using Variational Autoencoders [25] introduces an extension of the Vector Quantized Variational Autoencoder (VQ-VAE) to generate sparse explanations for image classification with DNNs, the SpaCE-VAE. The authors evaluated the proposed method against three other approaches that identify pixels that contribute negatively to the predicted class: DeepLIFT, Integrated Gradients, and Feature Ablations. Quantitative assessment using several examples with explanations was favorable to SpaCE-VAE, whereas qualitative (visual) inspection of the results indicates that SpaCE-VAE can provide insightful explanations.

Explaining Deep Learning Matching of Hand-Drawn Binary Symbols: A Visual Analysis with Grad-CAM on Cattle Brands [35] presents a case study in explainable AI, where Grad-CAM is used to understand the classification of a particular family of images. The authors investigate the application of Grad-CAM to analyze VGG-16 activations on binary cattle brand images. Careful experimental assessment exposed expected CNN limitations: rotation sensitivity and fine-grained feature under-activation.

A Non-Parametric Bayesian Approach Towards Online Sequence Learning [16] introduces a grid-based filter which is intrinsically designed to solve the online sequence learning problem and deal with uncertainty in time-series observations. The authors derive interpretable Bayesian procedures to both recursively predict observations and incrementally build a non-parametric, probabilistic model of the underlying temporal phenomenon. Moreover, they propose an one-shot learning, memory-based implementation of the filter which allows one to fully trace back inferences against associative input-output pairs stored within a weightless neural network, thus unlocking the potential for auditable (explainable) inferences. Lastly, they explore the potential of their filter using toy anomaly detection tasks. Despite presenting still limited results, the paper introduces a promising approach to the core of AI safety: accountability through the design of auditable methods with traceable outputs.

5. CONCLUDING REMARKS

We hope you will enjoy reading the papers on AI safety included in this special section.
In the introduction to this special section, we presented an overview of some of the developments in this fast evolving field. Despite being a relatively new field of research, the AI Safety community has been reaching a critical mass, with dedicated AI Safety Institutes [3] and a variety of academic events from technical governance of AI workshop series[7], aiming to provide analyses and tools to guide policy decisions and enhance policy implementation, to workshop on safety of generative AI[8] and safety-guaranteed LLMs[9], aiming to advancing technical and socio-technical research on critical topics. Nevertheless, we think that the core topics of knowledge representation, learning, and reasoning in the presence of *uncertainty* deserve drawing more attention from the UAI and sister communities, and we expect that these topics will become even more vivid in the future landscape of Safe AI research.

Acknowledgments

We thank all the authors who contributed to the 1^{st} Workshop on Safe AI at the UAI 2025 conference and to this special section, the program committee for helping review the manuscripts and shape the technical program, and all the participants for their contributions.

6. REFERENCES

[1] A. Ahmed, K. Klyman, Y. Zeng, S. Koyejo, and P. Liang. Speceval: Evaluating model adherence to behavior specifications. *arXiv preprint arXiv:2509.02464*, 2025.

[2] M. Alfonseca, M. Cebrian, A. Fernandez Anta, L. Coviello, A. Abeliuk, and I. Rahwan. Superintelligence cannot be contained: Lessons from computability theory. *J. Artif. Int. Res.*, 70:65–76, May 2021.

[3] R. Araujo, K. Fort, and O. Guest. Understanding the first wave of AI safety institutes: Characteristics, functions, and challenges. *arXiv preprint arXiv:2410.09219*, 2024.

[4] M.-F. Balcan, A. Blum, D. Sharma, and H. Zhang. An analysis of robustness of non-Lipschitz networks. *Journal of Machine Learning Research*, 24(98):1–43, 2023.

[7] `https://www.taig-icml.com/`
[8] `https://safegenaiworkshop.github.io/`
[9] `https://simons.berkeley.edu/workshops/safety-guaranteed-llms`

[5] F. A. Batarseh, L. Freeman, and C.-H. Huang. A survey on artificial intelligence assurance. *Journal of Big Data*, 8(1):60, 2021.

[6] Y. Bengio, S. Clare, C. Prunkl, M. Andriushchenko, B. Bucknall, P. Fox, N. Maslej, C. McGlynn, M. Murray, S. Rismani, et al. International ai safety report 2025: Second key update: Technical safeguards and risk management. *arXiv preprint arXiv:2511.19863*, 2025.

[7] C. Berghoff, B. Biggio, E. Brummel, V. Danos, T. Doms, H. Ehrich, T. Gantevoort, B. Hammer, J. Iden, S. Jacob, et al. Towards auditable AI systems. *Whitepaper. Bonn Berlin: Bundesamt für Sicherheit in der Informationstechnik, Fraunhofer-Institut für Nachrichtentechnik und Verband der TÜV eV*, 2021.

[8] R. E. Berker, E. Tewolde, I. Anagnostides, T. Sandholm, and V. Conitzer. The value of recall in extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13631–13640, 2025.

[9] N. Bostrom. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford, UK, 2014.

[10] S. Caton and C. Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), Apr. 2024.

[11] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr, H. Hassani, and E. Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[12] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. A survey of the state of explainable AI for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

[13] A. Das and P. Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[14] S. Davidson, O. Igene, and P. Miller. Assuring the case: A safety engineering approach to ai-enabled systems. *SIGKDD Expl.*, 27(2), 2025.

[15] Y. Deng, A. D. Bucchianico, and M. Pechenizkiy. Controlling the accuracy and uncertainty trade-off in rul prediction with a surrogate wiener propagation model. *Reliability Engineering System Safety*, 196:106727, 2020.

[16] S. S. Dias, M. G. da Silva Bruno, A. F. D. Souza, T. ago Oliveira-Santos, and C. Badue. A non-parametric bayesian approach towards on- line sequence learning. *SIGKDD Expl.*, 27(2), 2025.

[17] W. Ding, M. Abdel-Basset, H. Hawash, and A. M. Ali. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, 615:238–292, 2022.

[18] European Union Aviation Safety Agency. EASA concept paper: First usable guidance for level 1 machine learning applications. *EASA AI Roadmap*, 2021.

[19] European Union Aviation Safety Agency. EASA concept paper: Guidance for level 1 & 2 machine-learning applications. *EASA AI Roadmap*, 2024.

[20] D. Gunning and D. Aha. DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, 40(2):44–58, 2019.

[21] J. Ji, T. Qiu, B. Chen, J. Zhou, B. Zhang, D. Hong, H. Lou, K. Wang, Y. Duan, Z. He, L. Vierling, Z. Zhang, F. Zeng, J. Dai, X. Pan, H. Xu, A. O'Gara, K. Ng, B. Tse, J. Fu, S. Mcaleer, Y. Wang, M. Yang, Y. Liu, Y. Wang, S.-C. Zhu, Y. Guo, Y. Yang, and W. Gao. AI alignment: A contemporary survey. *ACM Comput. Surv.*, 58(5), Nov. 2025.

[22] F. Kamiran, T. Calders, and M. Pechenizkiy. Techniques for discrimination-free predictive models. In B. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky, editors, *Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases*, pages 223–239. 2013.

[23] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou. Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[24] X. Li, Q. Jiang, L. Jiang, S. Zhang, and S. Hu. The landscape of AI alignment: A comprehensive review of theories and methods. *International Journal of Pattern Recognition and Artificial Intelligence*, 2025.

[25] A. Liu and S. Hess. SpaCE-VAE: Sparse confident explanations using variational autoencoders. *SIGKDD Expl.*, 27(2), 2025.

[26] L. Liu, K. Rasul, Y. Chao, and J. Etesami. Riemannian manifold learning for stackelberg games with neural flow representations. *arXiv preprint arXiv:2502.05498*, 2025.

[27] M. Liu, C.-H. Lu, and M. Kwiatkowska. Exact verification of graph neural networks with incremental constraint solving. *arXiv preprint arXiv:2508.09320*, 2025.

[28] M. Mäntymäki, M. Minkkinen, T. Birkstedt, and M. Viljanen. Putting AI ethics into practice: The hourglass model of organizational AI governance. *arXiv preprint arXiv:2206.00335*, 2023.

[29] G. A. Melo, M. R. O. A. Máximo, N. Y. Soma, and P. A. L. Castro. Machines that halt resolve the undecidability of artificial intelligence alignment. *Scientific Reports*, 15(1):15591, May 4 2025.

[30] M. Pechenizkiy, H. Weerts, C. de Campos, Y. Sasaki, and J. Stoyanovich. From benchmarking to understanding fairml. In *ECAI 2025*, pages 38–45. IOS Press, 2025.

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

[32] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022.

[33] G. Schwalbe and B. Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101, 2024.

[34] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

[35] L. Soares, M. Medeiros, A. Dıaz-Salazar, and E. Hoyle. Explaining deep learning matching of hand-drawn binary symbols: A visual analysis with grad-cam on cattle brands. *SIGKDD Expl.*, 27(2), 2025.

[36] T. Tomilin, M. Fang, and M. Pechenizkiy. HASARD: A benchmark for vision-based safe reinforcement learning in embodied agents. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] U.S. AI Safety Institute. Nist AI 800-1: Second public draft — managing misuse risk for dual-use foundation models. Interagency or Internal Report NIST AI 800-1 (2nd Public Draft), National Institute of Standards and Technology (NIST), Gaithersburg, MD, January 2025. Second Public Draft.

[38] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019.

[39] X. Xue, C. Huang, K. Tsud, and D. Das. DT-sampler: A sat-based decision tree ensemble. *SIGKDD Expl.*, 27(2), 2025.

[40] C. Yadav, E. M. Laufer, D. Boneh, and K. Chaudhuri. Expproof: Operationalizing explanations for confidential models with zkps. *arXiv preprint arXiv:2502.03773*, 2025.

[41] C. Yadav, R. Wu, and K. Chaudhuri. Influence-based attributions can be manipulated. *arXiv preprint arXiv:2409.05208*, 2024.

[42] H. Zech. Liability for AI: public policy considerations. In *ERA forum*, volume 22, pages 147–158. Springer, 2021.