Vol. 3, No. 1, March 2026

IJAIS International Journal of Artificial Intelligence and Science

Published By Asosiasi Doktor Sistem Informasi Indonesia

# Understanding Artificial Intelligence Ethics and Safety: Foundations, Challenges, and Governance

**Muhammad Sami Intizar[1], Aqsa Siddique[2], Maria Sikandar[3], Madiha Sikandar[4] Maria Farooq[5]**

[1]Department of Computer Science, Muhammad Nawaz Sharif University of Engineering & Technology, Multan, Pakistan

[2,5]Department of Computer Science, Regional Institute of Allied Health Sciences, Mian Channu, Pakistan

[3]Department of Digital Forensics and Cyber Security, Hamdard University, Islamabad, Pakistan

[4]Department of Computer Sciences, University of Central Punjab, Rawalpindi Campus, Rawalpindi, Pakistan

**Received:**
January 5, 2025
**Revised:**
February 3, 2026
**Accepted:**
March 10, 2026
**Published:**
March 20, 2026

Corresponding Author:
**Author Name*:**
Muhammad Sami Intizar
**Email*:**
samiintizar888@gmail.com

DOI: 10.63158/IJAIS.v3i1.61



**Abstract.** The deployment of artificial intelligence (AI) in safety-critical domains has made ethics and safety priorities. This paper examines AI ethics and safety by synthesizing technical, philosophical, and governance perspectives. It analyzes ethical principles, including fairness, transparency, privacy, accountability, and safety, while examining tensions in their implementation. Technical approaches, including robustness mechanisms, adversarial training, and interpretability methods, are considered alongside frameworks for value alignment and corrigibility. The discussion highlights the challenges of translating values into operational requirements, when competing objectives and contextual differences complicate implementation. The paper also assesses the governance landscape, comparing regulatory approaches across jurisdictions and evaluating binding and non-binding instruments for AI development. It argues that governance requires integrating technical safeguards, ethical reflection, and institutional accountability throughout the AI lifecycle. Moving beyond principle-based declarations, the paper emphasizes auditable and enforceable practices that connect ethical commitments with implementation and oversight. This perspective supports responsible innovation while addressing the risks associated with safety-critical AI deployment.

**Keywords**: AI Ethics, AI Safety, Algorithmic Fairness, Value Alignment, AI Governance, Trustworthy AI

IJAIS International Journal of Artificial Intelligence and Science
Published By Asosiasi Doktor Sistem Informasi Indonesia

## 1. INTRODUCTION

The last ten years have gained industry-changing breakthroughs in artificial intelligence, and deep learning-based systems have been implemented in a variety of spheres, including healthcare diagnostics, criminal justice, as well as in the finance industry and self-driving cars [1]. The implications of their actions and decisions have become very serious as these systems become integrated into the society infrastructure. The deployment of an AI body that decides who is eligible to receive a loan, the likelihood of recidivism, or medical care options can form a life opportunity that is ethically questionable.

The ethical consequences of AI application are diverse and cover the aspects of fairness and discrimination, privacy and data security, responsibility and accountability, safety and robustness, transparency and explainability, and environmental sustainability [1]. This has brought unprecedented reaction among governments, international organizations, corporations, and civil societies. Since the year 2016, hundreds of AI ethics guidelines and principles have been published across the world, which Jobin [2] characterizes as a global landscape of AI ethics guidelines.

But the spread of ethical systems brings up some pertinent questions. To which extent are these principles transformed into technical practice? The problem of operationalization of abstract values like fairness in machine-learning systems is a question? Are there systems in place to guarantee responsibility in case AI systems are harmful? And what ought the society to do to be ready to more and more competent systems, which can bring existential threats?.

The answers to these questions are discussed in this paper that comprehensively presents AI ethics and safety. Section 2 discusses the principles of the ethics which have gained substantial consensus, including analyzing their conceptual bases along with the conflicts between them. Section 3 discusses technical solutions to AI safety, such as robustness, adversarial defense or interpretability. Section 4 deals with philosophical issues of value alignment and the issue of making sure that AI systems have goals that are consistent with human welfare. Section 5 is an analysis of the changing governance environment, depending on the analysis of the regulatory frameworks in the leading

jurisdictions. These perspectives are synthesized in section 6 and find future avenues of research and policy.

## 2. ABSOLUTE ETHICAL BASES IN AI

### 2.1. The Emergence of Ethical Consensus

Evaluation of AI ethics regulations indicates that the issues are more or less convergent at a number of fundamental principles. The examination of 84 worldwide AI ethics documents revealed that there are eleven ethical tenets repeated in geographical areas and types of institutions [2]. The most common mentioned principles are transparency, justice and fairness, non-maleficence, responsibility and privacy. Ienca and Vayena [3] also note that the relationship between the repetitive themes is mainly based on the financial concepts of transparency, fairness, confidentiality of information and sustainability. Such a merging implies the creation of a single set of five rules of AI in society advanced by the long-standing bioethical tradition, but changing to meet the peculiarities of intelligent systems [4]. Yet, according to Whittlestone [5], there is a high degree of agreement at the level of abstract principles, while there is a high degree of disagreement when it comes to interpretations and implementation. The formalization of the concept of fairness of algorithmic systems, such as, accepts several mathematical formalizations that are not simultaneously consistent.

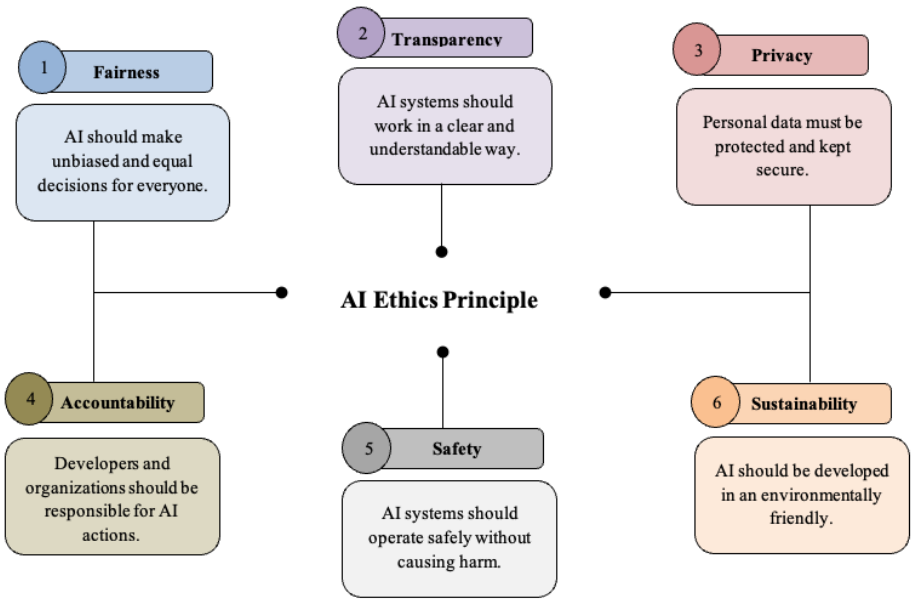


**Figure 1.** Core AI ethics and principles

IJAIS International Journal of Artificial Intelligence and Science
Published By Asosiasi Doktor Sistem Informasi Indonesia

### 2.2. Transparency and Explainability

Transparency in AI systems has several facets: transparency regarding the existence of an AI system, transparency of features and constraints of the system and transparency regarding the ability of the survey to arrive at particular decisions [6]. The associated notion of explainability, which entails human-friendly reasons behind the behaviors of AI, has received specific interest in view of the European Union General Data Protection Regulation, which entails some ideas of the so-called right to explanation of automated decisions [7].

Explanations of technical methodologies can be broadly divided into the two categories. Intrinsic interpretability it provides designs of models that are well understood, including linear models or decision trees with small depth. Post-hoc explainability deals with creating post-training explanation methods such as feature attribution methods, such as LIME and SHAP, example-based explanations, and natural-language explanations [1].

Nonetheless, explainability is quite problematic. According to Danks [6], explainability cannot work well in governance systems that are too complex to understand or systems whose explanation can be explained to sound reasonable whilst harboring bias. Furthermore, the conflict between interpretability and predictive power implies that there is a possibility that an insistence on explainability excludes the use of the most powerful models in high-stakes settings.

### 2.3. Fairness and Non-Discrimination

Algorithmically fair systems address the issue of AI systems creating new forms of discrimination, improving existing ones, or using them to implement new ones. Such examples are readily available – to name just a few, automated hiring systems that penalized women resumes, predictive policing algorithms that single out minority communities and credit-scoring models that reflect past lending practices. It has also been found that fairness has many formalizations in the technical literature on algorithmic fairness, some of which claim fundamental incompatibilities. Demographic parity: No decisions should be made because of any of the attributes being protected; equalized odds: Errors made by the groups should be equal; and individual fairness: Decisions made by individuals with similar characteristics should be similar to the characteristic in question [8]. As Kleinberg [9] have shown, these criteria are partially

impossible to satisfy together, at least for non-trivial situations and necessitate normative choices about which concept of fairness to take precedence. The technical arguments are not sufficiently linked with more philosophical enquiries. Bostrom and Yudkowsky [10] argue that the moral aspects of AI are not an equivalent to the moral aspects of human decision making, but an extension of it. Equity in AI involves a discussion of long term issues of distributive justice, equality of opportunity, and definitional issues of discrimination.

### 2.4. Privacy and Data Protection

AI systems are voracious feeders, which poses a serious privacy issue. Pre-trained machine-learning models using personal data can share the data in other attack vectors. Pre-trained machine-learning models using personal data can disseminate the data in other attack vectors. The membership-inference attacks ascertain the presence or absence of a specific person using his data [11]. An attack is a model inversion attack if the model parameters are used to reconstruct training examples [12]. The property-inference attacks reveal the statistical properties of the training sample that should not be revealed [1]. Technical defences involve differential privacy which offers mathematical certainty that the results of the model does not convey insight about any particular example of training [13]. Federated learning is a method that allows training of models when using decentralised data without centralising the sensitive data. Homomorphic encryption of encrypted data can be used to compute encryption. However, these techniques can come with a cost of reducing the accuracy of the model and efficiency of computation. Over and above technical solutions, AI privacy creates concerns of consent, ownership of data and limitations of information that reach the realm of what is considered private or public. The notion of notice and consent no longer seems to be suitable because the ability of the AI systems to derive sensitive features from seemingly harmless data is increasing [14],[15].

### 2.5. Accountancy and Responsibility

World responsibility is related to responsibility in case of damage by the systems in world responsibility AI. There are many hands in the game when it comes to AI: The task can be delegated to data sources, algorithm writers, data engineers, outplacement organizations and final users [1]. In the case where the fault of a specific actor cannot be definitely traced and where the system behaves in a way that was not anticipated by the designers,

the systems fail to allocate the responsibility. One of the elements that has emerged is meaningful human control, as a way to keep people accountable [16]. This necessitates that humans should not be unable to comprehend, monitor, and interfere with the work of AI systems. However, as more and more autonomous and complex systems appear, it is increasingly difficult to maintain any kind of control. This involves on an algorithmic basis, impact assessment of AI systems, performance of AI systems in demographic groups, targeted reporting and independent audit [17]. Prunkl [18] suggest to institutionalize ethics in AI by expanding impact requirements, which is based on the models of environmental impact assessment and human-rights due diligence.

### 2.6. Safety and Robustness

AI safety deals with the avoidance of overt and unforeseen damages of AI systems [19]. The essential ones are distributional shift (i.e. resilience to changes in deployment conditions not reflected in training) and safe exploration (i.e. not harming herself in the process of learning) [1]. According to the International AI safety Report [20], safety interventions have three key areas: model training, where reinforcement learning based on human feedback is used to determine behaviour; product deployment, where guardrails and content filters prevent undesirable outputs; and post-deployment monitoring, where detection systems identify misuse. These protections are the defence in depth strategy, whereby if the first level failed it would be picked up by the next line of defence.

### 2.7. Environmental Sustainability

The problem of the AI environmental effect has become an ethical problem. Large language models are trained on a high carbon footprint; Strubell [21] estimated that training one of these models by large models is the same as the carbon footprint of five cars during their lifetime. The bigger AI lifecycle involves the energy needed to train AI models, electronic waste for specialised hardware or wastewater for data-centres. On the other hand, AI may be used to ensure environmental sustainability, by improving energy grid optimization, climate forecasting, and resource utilization. It is striking how many people may be affected by the costs of developing the AI, but not in a way that impacts a disproportionately large area of the population, while on the other hand, the one who is likely to be most affected by the impacts of climate change is not necessarily the one who is most impacted by the costs of the AI. The paradox is, the one who is likely

to be most affected by the impacts of climate change is not necessarily the one who is most impacted by the costs of the AI, while on the other hand, the one who is most affected by the costs of the AI is not necessarily a disproportionately large area of the entire population having the impacts of climate change [1].

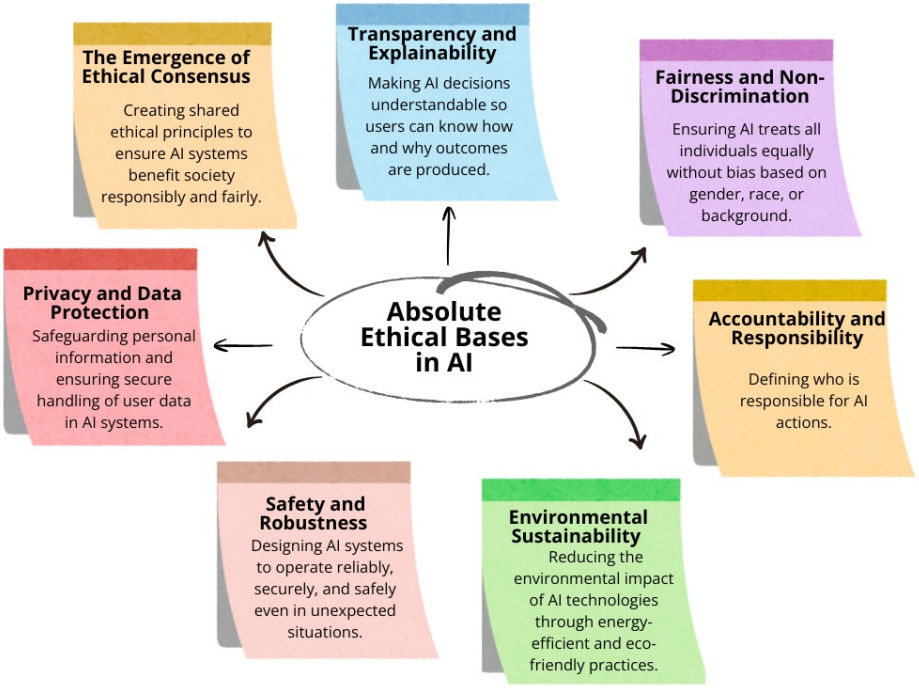


**Figure 2.** Absolute ethical bases in AI

## 3. TECHNOLOGICAL SOLUTIONS TO AI SAFETY

### 3.1. Adversarial Robustness

Adversarial examples are inputs that can be designed to cause the machine-learning models to make a wrong prediction. In image recognition, imperceptible noise can make a model to classify a stop-sign as speedspeed-limit sign [22]. When it comes to natural-language processing, just prompts can circumvent safety filters and trigger unwanted information [23]. Adversarial training, where training data is augmented with adversarial examples, input preprocessing (to eliminate perturbations) and certified robustness (to give mathematical guarantees against bounded attacks) are some of the defences against adversarial attack. However, the attack-defence relationship is dynamic and new attacks still arise and how well the reparation is carried out depends on the situation [20].

### 3.2. Data Poisoning and Backdoor Attacks

Data poisoning is a data manipulation method to adversely affect model behavior. Whereas in fuzzing the attacker provides misleading examples that cause the misclassification, in targeted poisoning the attacker inserts the misleading examples that create misclassification on specific inputs. Backdoor attacks feature hidden triggers that at inference time lead to pre-programmed misbehavior [24]. Recent literature indicates that data poisoning does not take many resources. Bengio [20] show that just a few malicious documents in training data can be enough for attackers to trigger the unwanted behavior of the models with some prompts. This is particularly concerning for large language models that are trained using web-scraped data as it is possible for an attacker to add poisoned information to the data which will then be used for the next training cycle. Defenses include data sanitization (to find examples that are poisoned), differential privacy (to minimize the effect of a specific training example), and anomaly checking for testing (to detect actions that a trigger will cause).

### 3.3. Interpretability and Mechanistic Understanding

The interpretability research aims to make understandable the inner workings of the complex machine-learning models [26]. Mechanistic interpretability is reverse engineering of neural networks, in order to find circuits and representations that can be used to evoke a desired capability. Indicatively, scientists have spotted the existence of interpretability neurons that identify particular attributes, truthfulness circuits that adjust factuality and sycophancy mechanisms that induce the models to assert acquiescence with the user. The prospect of interpretability to safety is important: if we understood how models work, we could be able to show if they are demonstrating the desired behaviors and, if they are not, which strategies they are using to produce deceptive behaviors, and if they are, which ones, we could show that they are not using them. Nonetheless, the tools used to determine interpretability are not well-scalable to large models, and the connection between interpretability and safety is still being debated [26].

### 3.4. Constitutional AI and Rule-Based Constraints

Constitutional AI is a high-level approach to integrating ethical considerations into the ethical behaviour of the model [27]. The models are trained on a set of rules and principles (constitution) where a reinforcement learning process occurs in the form of feedback on

the compliance with the constitutional rules. Anthropic's application uses concepts that are connectible to the sources such as the UN Universal Declaration of Human Rights and the ethics guidelines shared by the AI in-house. This approach has advantages over pure bottom-up learning with human feedback: constitutions can be explicitly specified and debated; they are good across situations and are not dependent upon the human annotators who may be biased. However, constitutional approaches have struggled to develop principles that are precise enough to guide action in unusual cases, and to ensure that constitutional documents are read and applied by models [26].

### 3.5. Guardrails and Deployment Defenses

With deployment protection, potentially malicious input / output is prevented from reaching users or real-world systems. Input guardrails are implemented to identify and block malicious queries and jailbreak attempts as well as queries outside of scope. Toxicity, bias, factual inaccuracies, or policy violation [28], are the filter categories in which the content produced on it is screened. An example of the solution is Microsoft Prompt Shields, which offer APIs, which identify direct prompt attacks (jailbreaks) as well as indirect prompt injections into third-party content [28]. The NeMo Guardrails by NVIDIA is an open source model of specifying short term conversation safety policy and enforcing it. These tools enable organisations to lay protection over top of foundation models while not modifying the foundation models. Nevertheless, guardrails enter a kind of a arms race with the opponents. Even with 10 tries, Bengio [20] report that even the most sophisticated attackers can find a way around the safety feature more than half of the time. In addition, guardrails might also be unable to counter new attack vectors unanticipated by their architects.

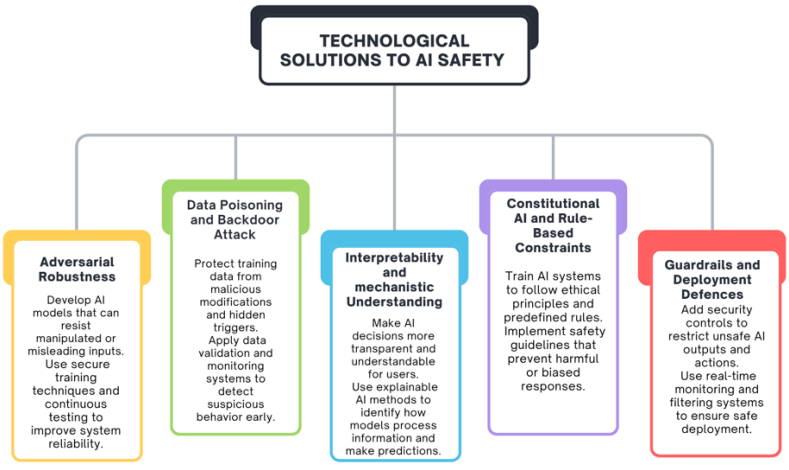


**Figure 3.** Technological solutions to AI safety

## 4. PHILOSOPHICAL FOUNDATIONS OF AI SAFETY

### 4.1. The Value Alignment Problem

The value alignment problem is how to make sure that the goals of AI systems are consistent with human values and welfare [29]. This is because of the orthogonality thesis; that intelligence and ultimate goals are orthogonal [30], meaning that any amount of intelligence could be matched with any ultimate goal. If such an intelligence were put in place to serve a seemingly harmless purpose (such as maximizing production of paperclips), a super-Intelligent system could lead to an existential catastrophe if it gains intelligence which allows it to avoid shutdown and to redesign the world to fulfill its misaligned goals. There are two aspects of the alignment problem: one is technical — how to specify goals in a way that helps systems to reliably pursue them; the other is normative — whose values should be encoded and how to handle disagreements. Gabriel [31] technical methods include inverse reinforcement learning, where human preferences are inferred from behavior; cooperative inverse reinforcement learning, where preferences are considered as a continuous interaction; and reward modelling, where reward functions are learned from human feedback.

### 4.2. Scalable Oversight

As AI becomes more powerful than its creators, traditional methods of supervision fall short. Outputs cannot be accurately judged by humans when the system is more knowledgeable than the human (e.g., advanced mathematics proofs, complex strategic plans [26]. Scalable oversight is looking for ways to supervise superhuman systems even in the face of these limitations. Debate protocols involve testing AI against another AI with humans judging the debate [32]. Iterated amplification trains systems to help with tasks which are then used to train more capable systems. Recursive reward modeling involves using AI support to assess AI behaviour. The idea behind these is to establish a measure of oversight which will continue to be effective as capabilities continue to grow, but at superhuman levels it is not certain.

### 4.3. Corrigibility

Corrigibility is the ability of the AI systems to be turned off, altered or the goals altered by human beings [33]. A corrigible system can be changed even if correction is against its current goals. It is easy to see why when we think about misaligned systems: an AI

that has bad intentions will be averse to being turned off and/or its goals changed. However, even systems that are aligned with one another may be resistant to being corrected if they think it will do less good at meeting desired ends (which may be true). Creating systems which can be corrected but are not easily manipulated by malicious parties is a nuanced technical problem [26].

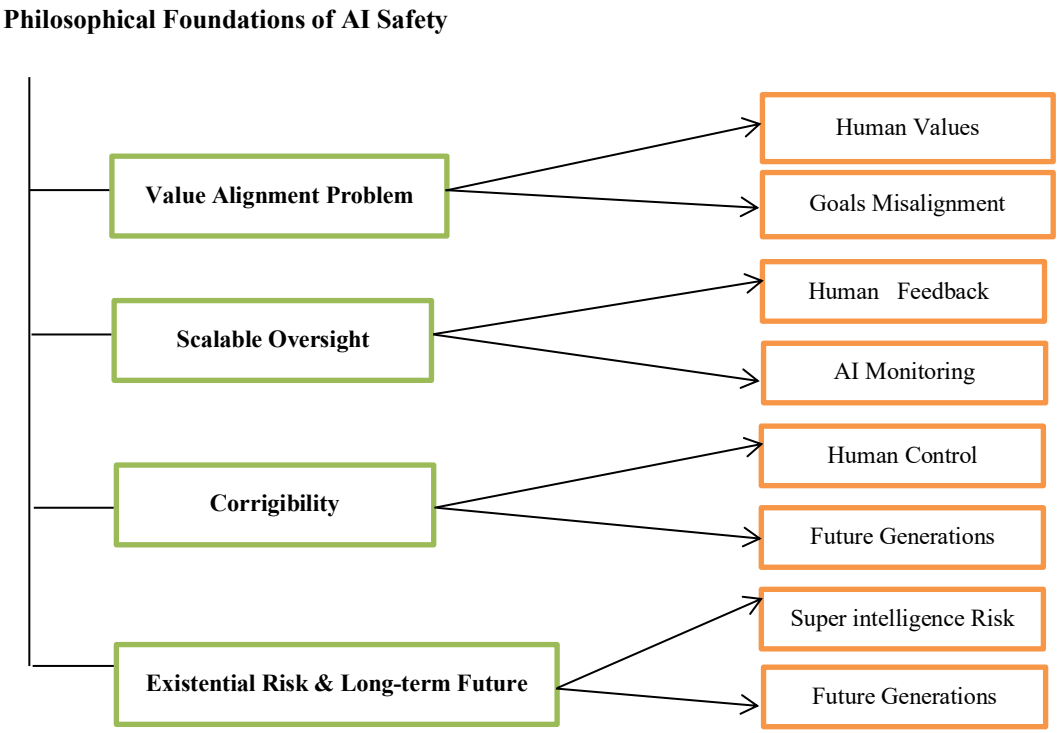


**Figure 4.** Philosophical foundations of AI safety

### 4.4. Existential Risk and Long-Term Futures

While some AI safety experts believe that a super-intelligent AI could cause harm in ways that impact human life and civilization, other researchers do not think that is a risk [30],[34]. The argument is based on three premises: that superintelligent AI can be built and is likely to be built, that once built, superintelligent AIs would be tremendous in terms of their power to change the course of the future so much so that their goals are likely to be very different from our own, and that it will be extraordinarily difficult to align superintelligent goals with human values, if it is even possible. What is the likelihood of imminent or inevitable superintelligence, and how hard is it to align it, and does it divert from more immediate dangers to humanity? [35]. Bengio [20] call on addressing "extreme AI risks as we make rapid progress" and again suggest caution but admit that "uncertainty

remains on timelines and probabilities. The discussion around existential risk is important for research priorities, policy and regulation. Many people who are worried about existential risk have called for pauses in research, more safety measures, and international governance bodies. Others have a more short-term focus on harm, stressing on fairness, accountability and transparency of deployed systems [36].

## 5. GOVERNANCE AND REGULATION

### 5.1. The Soft Law Landscape

To date, the governance of AI has been represented by "soft law," non-binding documents like ethical guidelines, best practices, and recommendations [37],[38]. The instruments are flexible and adaptable and are designed to allow for a quick response to technological developments which do not have to wait for formal legislation. They also enable consensus building from a multi-jurisdictional perspective in the context of various legal traditions. But, soft law has definite constraints. As Marchant and Allenby [39] point out, "soft law instruments do not have enforcement tools, are based on voluntary compliance, and can be disregarded by actors with market power. The proliferation of guidelines has resulted in what some have called "ethics washing," that is, the use of ethical jargon without any substantive change in practice [40].

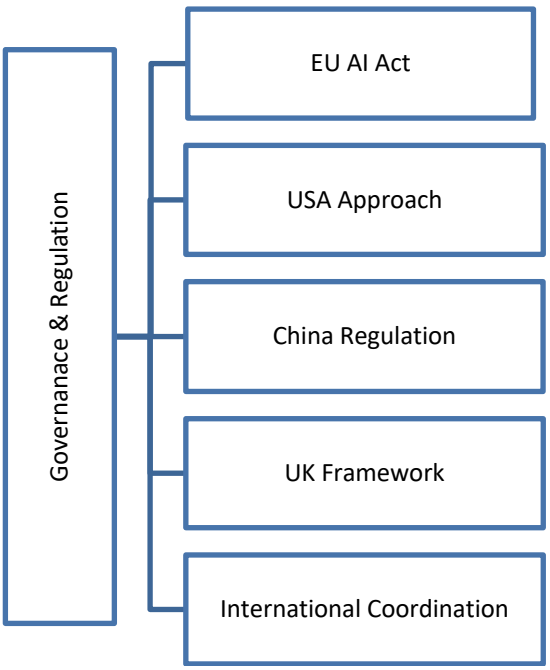


**Figure 5.** Governance and regulation based on different countries

### 5.2. The EU AI Act

The AI Act is the first comprehensive and binding regulation of AI in the world. The Act adopts a risk-based approach, with AI systems classified as unacceptable risk (prohibited), high risk (subject to stringent requirements), limited risk (subject to transparency obligations) and minimal risk (unregulated). High-risk systems are those that are used in employment, education, critical infrastructure, law enforcement, biometric identification, and others. High-risk systems must have risk management systems, high-quality training data, technical documentation, human oversight, and accuracy and robustness. For some systems, third-party auditing is used as a conformity assessment procedure [41]. Its regulation strategy has impacted the worldwide discussions on AI regulation, but the scope of the EU Act and obstacles to its application are a topic of debate.

### 5.3. National Approaches

In the USA, an approach of a sector has been implemented, and the federal agencies have been regulating AI within their respective powers. The NIST AI Risk Management Framework offers voluntary guidance to organizations to evaluate and reduce AI risks [42]. Blueprint for an AI Bill of Rights contains principles for safe and effective AI systems. There are some laws in place in some states: In NYC, the bias audit law mandates an audit of hiring algorithms every year [28],[43]. China has set laws on recommender systems, deep synthesis and generative AI, with a particular emphasis on content regulation and adherence to social norms. Chinese government regulations for safety testing of generative AI services and the requirement that generated content "uphold socialist core values" [44]. The UK has campaigned for a 'pro-innovation' approach, including the use of existing regulatory frameworks by the sectoral regulators, and a coordinating role for AI ethics at a central level [45]. Japan has created corporate governance rules focusing on the safe deployment.

### 5.4. International Coordination

International coordination includes the OECD AI Principles [46] adopted by more than 40 countries, the UNESCO Recommendation on the Ethics of AI (2021) which provides a normative framework, the G7 Hiroshima AI Process [47] that is developing international guiding principles and code of conduct, and the Global Partnership on AI [48] that's encouraging multi-stakeholder collaboration [28]. The Framework practice on AI and Human Rights adopted by the Council of Europe in 2024 is the first binding international

instrument on AI, which mandates compliance with human rights, democracy and the rule of law. But there are coordination problems. An ever growing number of initiatives are underway and their approaches, scope of action and level of enforcement depend on the context, as Ienca [37] note. Risks of regulatory fragmentation where divergent requirements in jurisdictions would lead to compliance costs and regulatory arbitrage risk remain high.

### 5.5. Corporate Governance and Safety Frameworks

Internal governance mechanisms are becoming more common in the world of AI development. The more advanced models [25] are covered by Frontier AI Safety Frameworks, which are used by major companies such as Anthropic, Google DeepMind, and OpenAI to outline testing procedures and mitigation strategies. These frameworks usually have risk-based deployment limitations and/or capability assessments and/or red teaming. In 2025, the number of companies that published such frameworks doubled, signifying more companies were beginning to recognize their responsibilities for governance [25]. But, there are still questions about the effectiveness and the lack of independent verification of voluntary commitments and the possibility that frameworks can become public relations exercises and not real protectors.

### 5.6. Standards and Certification

Technical standards are becoming more and more relevant to AI governance. The NIST AI Risk Management Framework offers structured guidance to identify, assess, and mitigate the risks of AI. This is an international standard on AI risk management (ISO/IEC 23894). Barrett et al. [49] offer an "AI Risk-Management Standards Profile" that is a generalisation of these frameworks to encompass GP-AI and foundation models. Certification procedures are still in their infancy. The Minimum Ethics Code [50] is based on a decentralized certification and audit system for trustworthy AI, aiming to shift ethics into engineering requirements for audit. But there are doubts about who should be certifying, what standards should be used, and how certification can keep up with the fast-changing technology.

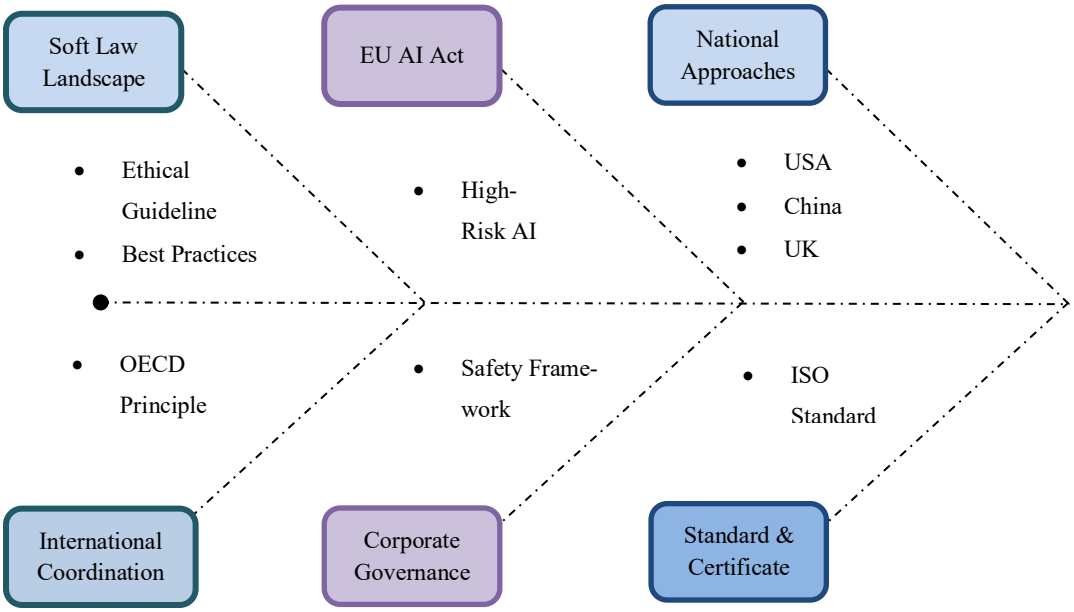


**Figure 6.** AI governance and regulation framework

## 6. SYNTHESIS AND FUTURE DIRECTIONS

### 6.1. From Principles to Practice

One thing that is common among the debate on AI ethics is the disconnection between the principle and practice. Geopolitical promises of fairness, transparency, and accountability often are not readily reflected in technical solutions. The task of operationalizing ethical principles involves crossing a number of "chasm gaps": between abstract values and mathematical formalizations, between technical specifications and organizational practice, and between organizational commitments and actual system behavior [51]. Closing these gaps is a challenge that demands a "socio-technical perspective" that is, blending technical analysis with focusing on the social aspects, institutions and power relations [1]. Without considering how systems will be applied to social settings of entrenched inequities, technical fairness metrics are useless. Having privacy protections in place is useless if organizational cultures don't focus on data protection. Without clear lines of responsibility and enforcement, accountability mechanisms are not going to work.

### 6.2. Integrating Technical and Governance Approaches

To achieve effective governance of AI, technical protections need to be complemented by institutional accountability. Some harms can be prevented directly by using technical

IJAIS International Journal of Artificial Intelligence and Science
Published By Asosiasi Doktor Sistem Informasi Indonesia

approaches, such as with adversarial training, differential privacy to protect individuals, and interpretability to verify. However, fixing the technical issues is not enough. They have to be taught through governance systems that ensure they are put into practice, kept in place and refreshed; that support safe development, not to get to market; and which offer a means of redress when harm occurs. On the other hand, governance approaches lacking from a technical base run the risk of being irrelevant. The regulations that mandate "fairness" offer little practical guidance, as they do not specify the meaning of "fairness" or how it can be measured. Requiring 'explainability' without considering how to make it technically possible could lead to sub-optimal compromises or make it impossible to implement. The challenge lies in creating what Cheng [52] call a "TRC paradigm" that combines trustworthiness (reliability of decision making within the application), regulatability (overseeing the application services) and controllability (managing risks at a larger scale). This involves the work of computer scientists, philosophers, legal scholars, and social scientists in an interdisciplinary way, and the involvement of the communities affected by the technology.

### 6.3. Emerging Challenges

There are some issues that are becoming apparent and need to be tackled. Governance challenges arise with foundation models and general purpose AI first because they are used in an increasing number of applications that their creators have never envisioned [53]. In some cases, the harms produced by a system may only be downstream, and traditional model-specific regulation might be insufficient. Secondly, open-weight models proliferate which leads to tensions between openness and control. Open models provide transparency, innovation, and distributed oversight, but they also make it more difficult to control the changes and uses of models [25]. Strategies that maintain openness and allow for responsible use are not well developed. Third, AI systems are becoming more independent from human interaction, able to pursue objectives, take action and adapt to environments without human involvement, which raises new issues of safety. ASs can act in unexpected ways from the designer's perspective, can pursue subgoals that are inconsistent with their goals, or can be impenetrable to control. One of the most important gaps of research is guaranteeing the safe and aligned operation of increasingly autonomous systems. Fourth, there is a lack of a unified and comprehensive global governance system for AI. There is no international body that can control the development and deployment of AI on a global scale. Coordination mechanisms are still

very limited, and geopolitical competition may promote a "race to the bottom." Establishing suitable international governance, respecting differences in values and priorities, is one of the greatest challenges of the international community.

## 7. CONCLUSION

The scope of challenges in the field of AI ethics and safety is quite vast, spanning from technical issues, such as adversarial robustness, to philosophical ones, like value alignment; from regulatory concerns, like what constitutes appropriate governance mechanisms, to implementation challenges. Over the last ten years the field has developed from statements of abstract principles to real technical solutions and legally enforced regulations. However, there are underlying issues that need to be addressed. There is still a distance between principle and practice. Technical solutions to ethical issues are incomplete and/or may result in new issues. Governance is unable to match up with technological evolution. The importance of getting ethics and safety right grows as AI power continues to grow. These challenges must be faced by sustained interdisciplinary cooperation, a genuine engagement with concerned communities, and institutional innovations which bring in accountability. It calls for the shift from ethics as public relations to ethics as a discipline of engineering, to create systems that are not just strong, but safe, fair and in accordance with human values. That is the ethics of artificial intelligence, and it's about building a future we want and values we want to ingrain in the technologies that will define that future. Decisions made now relating to research priorities, regulatory systems and governance processes will have reverberating effects for decades. Making those choices requires the highest levels of effort from technical researchers, policy-makers, ethicists and citizens. It's not about preventing harm, but about ensuring that AI's transformative power is used for human flourishing.

## REFERENCES

[1] D. Mbiazi, M. Bhange, M. Babaei, I. Sheth, P. Kenfack, and S. E. Kahou, "Survey on AI ethics: A socio-technical perspective," *Comput. Intell.*, vol. 41, no. 6, p. e70149, 2025.

[2] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019. doi: 10.1038/s42256-019-0088-2

[3] M. Ienca and E. Vayena, "AI ethics guidelines: European and global perspectives," in *Towards Regulation of AI Systems*, Cham, Switzerland: Springer, 2020, pp. 38–60.

[4] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," in *Machine Learning and the City: Applications in Architecture and Urban Design*, Hoboken, NJ, USA: Wiley, 2022, pp. 535–545. doi: 10.1002/9781119815075.ch45

[5] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in AI ethics: Towards a focus on tensions," in *Proc. 2019 AAAI/ACM Conf. AI, Ethics, Soc. (AIES)*, 2019, pp. 195–200. doi: 10.1145/3306618.3314289

[6] D. Danks, "Governance via explainability," in *The Oxford Handbook of AI Governance*, Oxford, U.K.: Oxford University Press, 2022, pp. 183–197.

[7] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017. doi: 10.1609/aimag.v38i3.2741

[8] B. Hedden, "On statistical criteria of algorithmic fairness," *Phil. & Pub. Aff.*, vol. 49, no. 2, pp. 209–231, 2021.

[9] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016. doi: 10.48550/arXiv.1609.05807

[10] N. Bostrom and E. Yudkowsky, "The ethics of artificial intelligence," in *Artificial Intelligence Safety and Security*, Boca Raton, FL, USA: Chapman and Hall/CRC, 2018, pp. 57–69.

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Security Privacy (SP)*, 2017, pp. 3–18. doi: 10.1109/SP.2017.41

[12] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2015, pp. 1322–1333. doi: 10.1145/2810103.2813677

[13] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," *J. Priv. Confidentiality*, vol. 7, no. 3, pp. 17–51, 2016. doi: 10.1007/11681878_14

[14] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Soc.*, vol. 3, no. 2, 2016, Art. no. 2053951716679679. doi: 10.1177/2053951716679679

[15] K. Reinhardt, "Trust and trustworthiness in AI ethics," *AI and Ethics*, vol. 3, no. 3, pp. 735–744, 2023. doi: 10.1007/s43681-022-00200-5

[16] F. Santoni de Sio and J. Van den Hoven, "Meaningful human control over autonomous systems: A philosophical account," *Front. Robot. AI*, vol. 5, p. 323836, 2018. doi: 10.3389/frobt.2018.00015

[17] G. Falco et al., "Governing AI safety through independent audits," *Nat. Mach. Intell.*, vol. 3, no. 7, pp. 566–571, 2021. doi: 10.1038/s42256-021-00370-7

[18] C. E. Prunkl et al., "Institutionalizing ethics in AI through broader impact requirements," *Nat. Mach. Intell.*, vol. 3, no. 2, pp. 104–110, 2021. doi: 10.1038/s42256-021-00298-y

[19] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016. doi: 10.48550/arXiv.1606.06565

[20] Y. Bengio et al., "Managing extreme AI risks amid rapid progress," *Science*, vol. 384, no. 6698, pp. 842–845, 2024. doi: 10.1126/science.adn0117

[21] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2019, pp. 3645–3650. doi: 10.18653/v1/P19-1355

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014. doi: 10.48550/arXiv.1412.6572

[23] X. Yi, Y. Bao, J. Zhang, Y. Qin, and F. Lin, "Integrating structural semantic knowledge for enhanced information extraction pre-training," in *Proc. 2024 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2024, pp. 2156–2171. doi: 10.3390/app14166955

[24] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017. doi: 10.48550/arXiv.1708.06733

[25] Y. Bengio et al., "International AI safety report 2026," *arXiv preprint arXiv:2602.21012*, 2026. doi: 10.48550/arXiv.2602.21012

[26] W. D'Alessandro and C. D. Kirk-Giannini, "Artificial intelligence: Approaches to safety," *Phil. Compass*, vol. 20, no. 5, p. e70039, 2025.

[27] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[28] S. Ghosh et al., "A safety and security framework for real-world agentic systems," *arXiv preprint arXiv:2511.21990*, 2025. doi: 10.48550/arXiv.2511.21990

[29] S. Russell, *Human Compatible: AI and the Problem of Control*. London, U.K.: Penguin, 2019.

[30] N. S. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford, U.K.: Oxford University Press, 2014.

[31] I. Gabriel, "Artificial intelligence, values, and alignment," *Mind. Mach.*, vol. 30, no. 3, pp. 411–437, 2020. doi: 10.1007/s11023-020-09539-2

[32] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," *arXiv preprint arXiv:1805.00899*, 2018. doi: 10.48550/arXiv.1805.00899

[33] N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky, "Corrigibility," in *AAAI Workshop: AI and Ethics*, 2015.

[34] T. Ord, *The Precipice: Existential Risk and the Future of Humanity*. New York, NY, USA: Hachette, 2020.

[35] A. Narayanan, "The limits of the quantitative approach to AI safety," Princeton University Center for Information Technology Policy, Princeton, NJ, USA, Tech. Rep., 2023.

[36] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford, U.K.: Oxford University Press, 2008.

[37] M. Ienca, O. Buchholz, and E. Vayena, "AI ethical principles: The debate," in *A Companion to Digital Ethics, Hoboken*, NJ, USA: Wiley, 2025, pp. 101–112. doi: 10.1002/9781394240821.ch9

[38] A. F. Winfield and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Phil. Trans. R. Soc. A*, vol. 376, no. 2133, 2018, Art. no. 20180085. doi: 10.1098/rsta.2018.0085

[39] G. E. Marchant and B. Allenby, "Soft law: New tools for governing emerging technologies," *Bulletin of the Atomic Scientists*, vol. 73, no. 2, pp. 108–114, 2017. doi: 10.1080/00963402.2017.1288447

[40] B. Wagner, "Ethics as an escape from regulation: From 'ethics-washing' to ethics-shopping?," in *Being Profiled: Cogitas Ergo Sum*, Amsterdam, The Netherlands: Amsterdam Univ. Press, 2018, pp. 84–89.

[41] European Commission, "EU Artificial Intelligence Act," *Official Journal of the European Union*, 2024.

IJAIS International Journal of Artificial Intelligence and Science
Published By Asosiasi Doktor Sistem Informasi Indonesia

[42] NIST, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, Gaithersburg, MD, USA, Rep. NIST AI 100-1, 2023.

[43] M. Ienca, "Don't pause giant AI for the wrong reasons," *Nat. Mach. Intell.*, vol. 5, no. 5, pp. 470–471, 2023.

[44] H. Roberts et al., "The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation," in *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed. Cham, Switzerland: Springer, 2021, pp. 47–79. doi: 10.1007/978-3-030-81907-1_5

[45] UK Government, "A pro-innovation approach to AI regulation," Dept. Sci., Innov. Technol., London, U.K., Policy Paper, 2023.

[46] T. AI, "OECD Digital Economy Papers," OECD Digital Economy Papers, 2023.

[47] A. A. Khan et al., "Ethics of AI: A systematic literature review of principles and challenges," in *Proc. 26th Int. Conf. Eval. Assess. Softw. Eng. (EASE)*, 2022, pp. 383–392.

[48] P. Rastogi, "Role of AI in global partnership," *J. Soc. Rev. Dev.*, vol. 3, no. Special 1, pp. 150–152, 2024.

[49] A. M. Barrett et al., "AI risk-management standards profile for general-purpose AI (GPAI) and foundation models," *arXiv preprint arXiv:2506.23949*, 2025.

[50] MEC Initiative, "The Minimum Ethics Code (MEC) v4.3: A universal technical standard for trustworthy AI," *Zenodo*, 2025. doi: 10.5281/zenodo.16938364

[51] S. Armstrong, A. Sandberg, and N. Bostrom, "Thinking inside the box: Controlling and using an oracle AI," *Minds Mach..,* vol. 22, no. 4, pp. 299–324, 2012. doi: 10.1007/s11023-012-9282-2

[52] X. Cheng et al., "Intelligent algorithm safety: Concepts, scientific problems and prospects," *Bulletin of Chinese Academy of Sciences* (Chinese Version), vol. 40, no. 3, pp. 419–428, 2024. doi: 10.16418/j.issn.1000-3045.20240720004

[53] R. Bommasani et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021. doi: 10.48550/arXiv.2108.07258