Open Access Library Journal
2025, Volume 12, e14231
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Regulating AI: A Comprehensive Review of Strategies for the Ethical and Safe Use

Hong Yu

College of Communication and Information Engineering, Chongqing College of Mobile Communication, Chongqing, China
Email: 1029592383@qq.com

How to cite this paper: Yu, H. (2025) Regulating AI: A Comprehensive Review of Strategies for the Ethical and Safe Use. *Open Access Library Journal*, **12**: e14231.
https://doi.org/10.4236/oalib.1114231

Received: September 6, 2025
Accepted: October 7, 2025
Published: October 10, 2025

Abstract

Artificial intelligence (AI) technologies are progressing rapidly, presenting opportunities and intricate ethical and legal issues. This evaluation delineates modern methodologies and classifications of AI governance to facilitate its secure and beneficial implementation. Ethical considerations must be integrated into AI frameworks emphasizing transparency, accountability, and fairness. The paper also addresses the imperative of financing AI safety research to mitigate dangers, especially those associated with bias and unemployment. Ultimately, despite the urgency to deploy a model, it is imperative to solve numerous issues associated with large-scale implementation, necessitating thorough testing and validation before utilizing an AI system. The development of AI is subject to regulation by regulatory authorities that will maintain ethical standards and address public concerns. Moreover, promoting transparency and public awareness is a crucial element in effective AI governance. The paper outlines a strategy for future research to improve regulatory mechanisms to ensure AI algorithms promote ethical conduct while reducing obstacles to innovation and societal welfare. The paper presents a plan for future research to enhance regulatory instruments for maintaining AI algorithms that drive ethical behaviour and minimize barriers to innovation and society's well-being.

Subject Areas

Artificial Intelligence

Keywords

Artificial Intelligence, Human, Technology, Ethics, Policy

1. Introduction

The challenges of the proliferation of artificial intelligence (AI) in society require

a nuanced approach to regulation that balances innovation and corporate responsibility. One prominent strategy would be the development of responsible AI governance frameworks that prioritize transparency, accountability, and fairness, emphasizing that these principles will play a crucial role in addressing the ethical challenges associated with AI implementation. However, the rapid evolution of AI technologies offers exceptional opportunities and significant threats to different fields, such as the healthcare industry, vehicles, financial infrastructures, and educational systems. The potential of AI to transform this could be massive because it increases efficiency, accuracy, and on-the-ground access in these fields. However, with AI becoming a part of everyday life, important ethical, social, and safety questions must be addressed in the development and widespread application of the technology. As intelligent machines become ubiquitous among us, addressing issues such as privacy leakage, discrimination, unemployment, or security risks demands that we develop players of some sort for putting machine morality projects like friendly AI on a legal footing [1]. More conscious and concentrated regulations are required for the healthcare industry; these regulations should, in our opinion, safeguard patient safety, promote innovation, and address ethical concerns [2]. Although the number of massive health records and data that AI systems can process has the potential to transform public health significantly, it also highlights the significance of ethical principles like equity, bias, privacy, security, safety, transparency, confidentiality, accountability, social justice, and autonomy [3]. In the era of the AI-driven Fourth Industrial Revolution, a justice system is required that allows innovation and protects fundamental human rights and a freedom-based approach, as in the EU compared to the remaining regions, such as the US or China [4]. Regulations on how AI should be regulated have been suggested, including stringent testing and validation for safety research, supervision by regulators, and greater transparency. Education of the public about the implications and promotion of human-AI collaboration is also important to direct AI development toward positive societal outcomes. By tackling these multidimensional problems with a cocktail of ethical principles, detailed regulations, and proper monitoring, we can maximize the benefits of AI and minimize the risks of stateless, responsible, and fair adoption in our increasingly interwoven lives.

Creating ethical AI frameworks represents an important step in ensuring that human values and desires are programmed into our AIs, keeping to the principles of trustworthy, accountable, and fair behaviour on behalf of our AI systems. These frameworks are a compass for developers, policymakers, and stakeholders in weaving through the intricate landscape of moral issues such as bias, discrimination, and privacy. For example, the Trustworthy AI guidelines of the European Union highlight privacy and data governance legal requirements as well as technical robustness that practitioners see in software engineering management practices as a risk requirement or quality attribute [5]. Ethics for AI in practice although, as highlighted by the experiences of researchers and engineers at Australia's CSIRO who need to design responsible AI systems, a gulf separates high-

level ethical principles from pragmatic methodologies [6], tensions and trade-offs between various principles such as privacy protection, reliability assurance transparency, and fairness. In addition, although there is consensus regarding the value input as markers for behaviour (indeed a critical part of engineering ethical AI), both computational frameworks and actual deployments require simple but effective means to ensure that AI systems are suitably aligned with human values–a challenging open problem on which this paper sheds light by suggesting an informal conception of values inherent in social sciences [7]. Notwithstanding the proliferation of frameworks, they mainly exist at the level of requirements elicitation in the software development life cycle (SDLC), and it means other phases are either less supported or not so thoroughly described for practitioners, as well as lacking full tool coverage on them [8]. Hence, having comprehensive frameworks in place to draw the line about ethicality, which covers all phases of SDLC and also focuses on involving both technical and non-technical stakeholders, is vital so that AI can be developed, keeping humanity at its core. These practices make it easier to cascade from ethical principles to practice, which will contribute to establishing a more responsible and trustworthy AI ecosystem.

Investing in AI safety research is a critical component of responsible AI regulation, as it addresses the unpredictable behaviour and vulnerabilities inherent in AI systems, particularly those utilizing machine learning and neural networks. Advanced AI models, or "frontier AI," can possess dangerous capabilities that pose severe risks to public safety, necessitating robust regulatory frameworks to manage these risks effectively [9]. The rapid adoption of large language models has heightened excitement and concern, underscoring the need for a sociotechnical approach to AI safety beyond the prevailing technical agenda [10]. Ensuring AI's ethical, trustworthy, and legal deployment requires comprehensive lifecycle audits and the development of compliance mechanisms to mitigate potential negative impacts on individuals, society, and the environment [11]. Historical patterns in high-tech regulation reveal that incidents often drive regulatory advancements, suggesting that a strategy for collecting and analyzing AI incident data is crucial for improving our understanding and regulation of AI technologies [12]. Furthermore, as AI transforms government operations, it is essential to connect emerging knowledge about internal agency practices with longstanding lessons about organizational behaviour and legal constraints to achieve meaningful accountability and prevent harmful outcomes such as job displacement [13] and the misuse of autonomous weapons [14]. These insights highlight the importance of AI safety research in developing methods to identify, measure, and address potential flaws and biases, thereby preventing unintended consequences and ensuring the responsible advancement of AI technologies.

It is essential to have solid testing and validation processes for AI systems to act reliably and safely in practical situations. This requires technical verification and verification against the law and established guidelines. The things that make it challenging to regulate the deployment of AI algorithms are increasing their

capabilities and continuously advancing them as a part of organic development [15], requiring a balance between safety assurance and innovation. This highlights the importance of governance and rigorous testing protocols in properly developing robust models [16]. Most often, ethical demands such as privacy and data governance are typified under legal requirements but require a more holistic approach that also considers technical solidity, safety, and the welfare of society [5]. AI application needs to be trustworthy, and therefore, practical assessment approaches are crucially needed that allow checking if an AI system adheres to high-quality demands on the one hand, but at least be protected against novel emerging dangers like bias or unfair respect of humans [17]. The speed of AI development, driven by this Fourth Industrial Revolution, poses both opportunities and threats, crystallizing the demand for a regulatory system that secures both innovation and credibility [4]. We need dedicated bodies to regulate the development to enforce ethical standards, monitor the applications, investigate potential violations, and ensure compliance with regulatory requirements. It establishes accountability for developers and users, which will help increase the overall trustworthiness of AI technologies.

Because so many AI systems are easily referred to as "black boxes" that conceal how and why decisions were made, transparency—when paired with an explainability tool by which users can better understand the model's decision-making processes—is essential in driving user trust. Recently, Explainable Artificial Intelligence (XAI) has gained increasing importance in addressing these challenges by providing transparency and interpretability through methods such as saliency maps, attention mechanisms or rule-based explanations, and model-agnostic approaches [18]. In safety-critical domains, such as air traffic control or even self-driving cars), the need for explainability is critical to AI systems that are practical and efficient and will only be trusted when they can explain their responses [19]. Supporting research also suggests that these questions differ per user group (e.g., developer or end users) and must be adapted accordingly for context, domain expertise, and cognitive resources [19]. Learning to express integrity in AI explanations, including appreciation of accountability for decision-related honesty, might also improve human users' subjectively trustworthy sense [20]. Although it falls short of requiring the application of XAI techniques, this provision in the proposed EU AI Act addresses some technical limitations and ongoing scientific research on explainability for human oversight at least [21]. Promoting informed decision-making and critical thinking also requires raising public awareness about AI, demystifying it for the common citizenry—who often need to be more informed or more accurate information, which leads them to mere speculation regarding its full implementation into reality by sophisticated data-driven agencies. Education can help make the public aware of what AI is and its benefits and risks in a more balanced way so that it can be used more responsibly [21].

However, recent research in AI governance laid the foundation for ethical and policy principles that have been very influential on current conversations regarding

regulation and safety. Moreover, the AI4People recommends beneficence, non-maleficence, autonomy, justice, and explicability as core values that are crucial for nurturing a "good AI society" [22]. Among them, IEEE's Ethically Aligned Design is an exemplar that advocates for the embedding of human well-being within technical standards [23]. Moreover, the comparative analyses of the international guidelines converge on principles such as transparency, accountability, and fairness, with interpretations and uses of them differing substantially [24]. In addition, early theoretical work on algorithmic decision-making has highlighted the ethical dimensions of these technologies and reiterated the importance of a comprehensive approach to AI governance that balances innovation with concerns for ethics [22] [25]. This integrative amalgamation of principles and frameworks offers a critical lens for operationalizing the messiness of AI ethics.

Human-AI collaboration must be fostered to provoke the best ideas while restraining the worst ones. Begging the question: AI is not another human capacity, and nor has it started degradation; rather, it is an assistant skill set in partnership with humans, i.e., Human-AI Teaming (HAT) instead [26]. This method optimally utilizes the capabilities of both humans and AIs, allowing for more robust and dynamic interaction in various domains. However, this partnership ought to be systemized because it may clash between the variance of views and interpretation, with potentially drastic consequences if avoided [27]. In order to guarantee that AI systems maintain themselves by human values and contribute effectively to overall conformance, it is vital to think of a Human-Centred AI (HCAI) method. Such as user empowerment, ethical concerns, and approaches to more humanistic design, which provide better user experiences and bring trust in the users [28]. Integrating ethical virtues such as fairness, transparency, accountability, and privacy preservation in developing AI can yield human rights-abiding systems without bias, benefiting people at large while contributing to global societal progress [28]. It is argued that this proposed conceptual framework of human-AI joint cognitive systems (HAIJCS) could be a practical solution to integrate HAT into the new paradigm so that AI systems can effectively act as teammates but are under control and supervision by us humans again, in line with their designs along principles originated from [26]. Based on Erik Hollnagel and David Woods's joint cognitive systems theory, Mica Endsley's situation awareness cognitive engineering theory, and the agent theory widely used in AI/CS communities, we propose a conceptual framework of joint cognitive systems to represent HAT (**Figure 1**) [29]. By promoting interdisciplinary collaboration and collective decision-making, we can unleash the power of AI to open up a future that more closely meets our shared human goals and values than any achieved before, one in which AI technologies will benefit humanity as a whole.

There is a need for clear regulatory structures to ensure the responsible and safe use of AI, especially in business sectors where the lack thereof has slowed adoption [30]. As demonstrated by the European Union's example, regulations

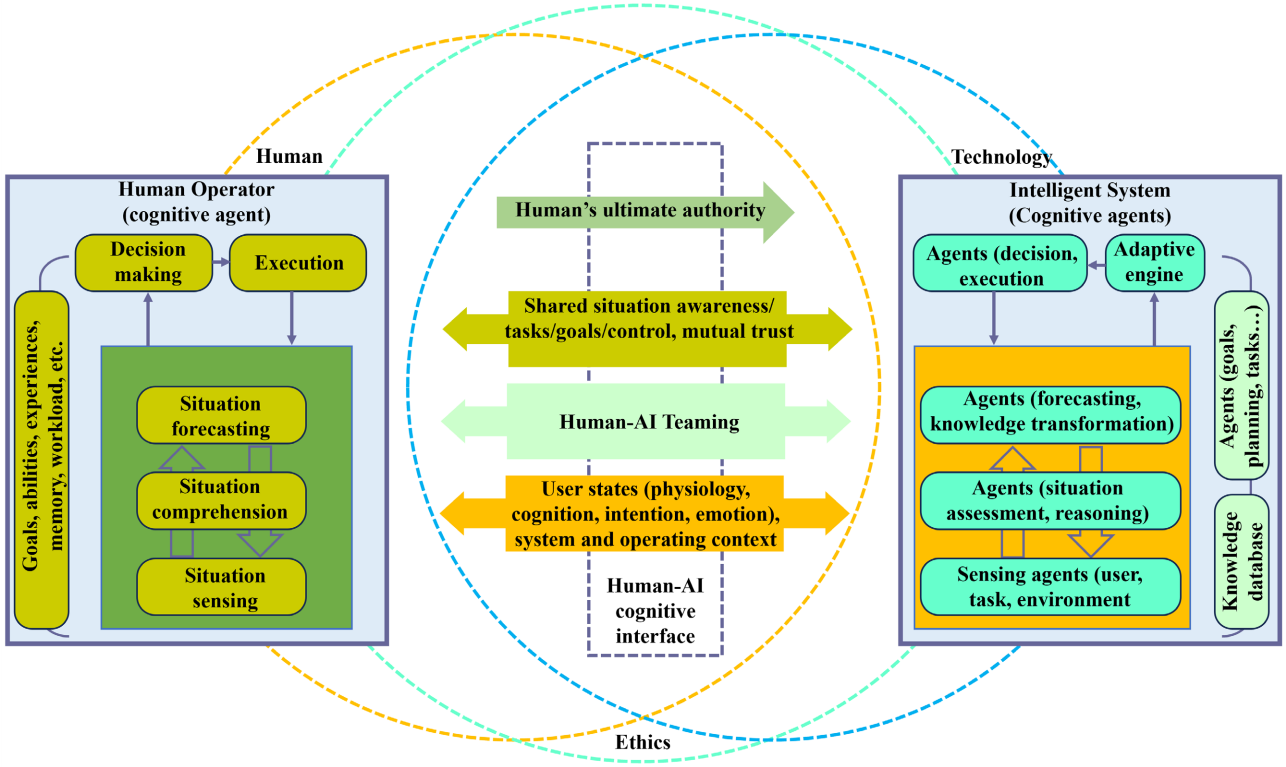

Figure 1. The conceptual framework of human-AI joint cognitive systems (HAIJCS), redrawn from [29].

will thus need to be specific and strike a balance between innovation freedom and ethical considerations in light of these shifts are characterized under what is now known as the Fourth Industrial Revolution [4]. Designing and deploying ethically sound, reliable, accountable AI technologies at scale necessitates fitting practices across the lifecycle of these systems with new governance tools to span operational gaps [11]. This shift of AI into discretion-heavy policy spaces in government applications, argues [14], necessitates a nuanced understanding of organizational behaviour and law at once that is fit to demand meaningful accountability without impeding further innovation. The fast uptake of AI in the healthcare sector has disrupted the entire industry perspective and made it harder for us to draw up suitable guidelines. A few experts suggest that we need a granular set of regulations tailored differently to accommodate unique challenges, patient safety and innovation [2]. In this review paper, we break down each of these strategies and analyze its significance followed by how it gets implemented drawing from case studies across sectors tackling diverse challenges to stitching together the different regulatory responses in AI to offer an end-to-end view on regulation.

This review paper aims to comprehensively and critically analyze the major approaches proposed for governing AI focusing on beneficial AI. It is also intended to be a complete groundwork in ethical frameworks, safety research and testing protocols, regulatory bodies, transparency practices, public education inclusion programs, and human-AI partnership initiatives or value alignment. The paper

attempts to do this by examining these strategies, their significance, and suggestions for AI developers and policymakers. In the longer term, it seeks to help shape a prudent and pro-humanist approach to AI policy by laying out an inclusive path toward human flourishing with technology.

2. Methodology

The current study uses a qualitative, comparative and descriptive methodology to analyze the ethical usage of AI in a few sectors. The study starts with an extensive literature review that rigorously evaluates academic articles, policy papers, and industry guidelines to uncover the principal ethical concerns, such as fairness, accountability, safety, and transparency of AI use. After that, a comparative policy analysis of global AI governance frameworks takes place, which includes a look at how various nations handle AI safety and ethics. This comparative lens allows for the recognition of similarities and differences in regulatory practices around the world. In addition, expert consultations with AI developers, ethicists, and policymakers enrich the literature and policy review with concrete insights into the lived experience of AI governance and implementation. This study proposes a conceptual framework for ethical AI practices that regulators and stakeholders can implement to ensure the responsible deployment of AI based on the aforementioned findings. This is followed by a comprehensive analysis of existing AI regulatory frameworks, pinpointing optimal practices and proposing enhancements for AI governance. The study integrates qualitative methodologies to produce refined and practical insights and recommendations for policymakers and practitioners to establish more robust, transparent, and accountable AI oversight frameworks.

We used a systematic method for literature search on Scopus, Web of Science, IEEE Xplore, ACM Digital Library, PubMed, SSRN, and Google Scholar. In search, we combined controlled vocabularies and free-text terms, for example, “artificial intelligence” and avoidable healthcare harm (governance or regulation or “risk management” or “ethics framework” or “safety” or “explainability” or “human-AI collaboration”). English language peer-reviewed or authoritative policy/standards that met eligibility requirements related to AI governance/risk/regulation. Excluded were performance studies that were exclusively technical in nature without implications for governance, non-scholarly commentaries, and duplicates. Screening was conducted in two phases (title/abstract, full-text) and with snowballing for key articles.

For law and policy contexts, incorporating multi-disciplinary perspectives into research strengthens the application of knowledge for evidence that limits bias and supports decision-making, such as the need to include different aspects of social science, from economics and sociology, into legal reforms that target Sustainable Development Goals (SDGs), creating holistic synthesis-solutions for complex issues [31]. Similarly, highlight the need to amalgamate global environmental knowledge to catalyze national actions and call for an intersectoral

collaboration to tackle concurrent environmental challenges [32]. Furthermore, it has been shown that expert views on gene drive technologies carry moral complexities that can help navigate responsible policy-making [33]. Earlier work introduces the difficulties faced in the policy arena in using transdisciplinary insights and highlights the possibility of contextual contingency [34]. These findings illustrate the added value of expert involvement and thematic analysis to complex and informed policy frameworks [35].

3. Developing Ethical AI Frameworks

Ethical AI frameworks are important to ensure that AI systems are designed and developed to respect societal values, but more is needed. These provide the fundamental principles to build trust and encourage responsible AI practice, i.e., fairness, transparency, accountability, or data privacy protection. The increase in AI importance, as far as Fisher is concerned, and its environmental impact have turned the legislative spotlight on ethical concerns with privacy issues following close behind-forcing primary legislation sooner than later done via an international cooperation mechanism [36]. In addition, integrating the ethical requirements with SW engineering practice at the management (middle and upper) level becomes a must. Privacy and data governance are usually the primary focus from a legal perspective, yet is also emphasized on other ethical aspects (e.g., technical robustness, safety, societal well-being) that ought to be an integral part of management practices employing frameworks like Agile portfolio management [5]. While various frameworks for Responsible AI (RAI) already exist, there currently needs to be a comprehensive framework that serves the needs of both technical and non-technical stakeholders in all stages of the Software Development Life Cycle from Ideation to Deployment. Currently, most of the frameworks in use only consider the Requirements Elicitation step and no other phases, emphasizing the necessity for inclusive guidelines [8]. In addition, the problem of value alignment (ensuring AI stays consistent with human values) highlights a necessity for developing provably beneficial AI: systems whose actions can be shown not only as useful but deployed in valuable ways according to some ethical framework. This needs the type, style, and formalization of a values definition or reasoning system recommended by those who say we need an interdisciplinary approach to AI ethics based on a social science-oriented ethical framework [7]. **Figure 2** illustrates the evolution of ethical AI frameworks, supporting policymakers, developers, and users to design principles-based systems roles. These complex challenges lay the foundation for ethical AI frameworks to help policymakers, developers, and users handle delicate moral issues when designing or employing various AI technologies.

These models can be divided into regulatory, self-regulatory, and co-regulatory frameworks, each presenting unique aspects to regulate the multitudes of AI systems. The EU AI Act serves as a case study of a regulatory framework with an architecture for enforcement involving multiple institutional actors, from the

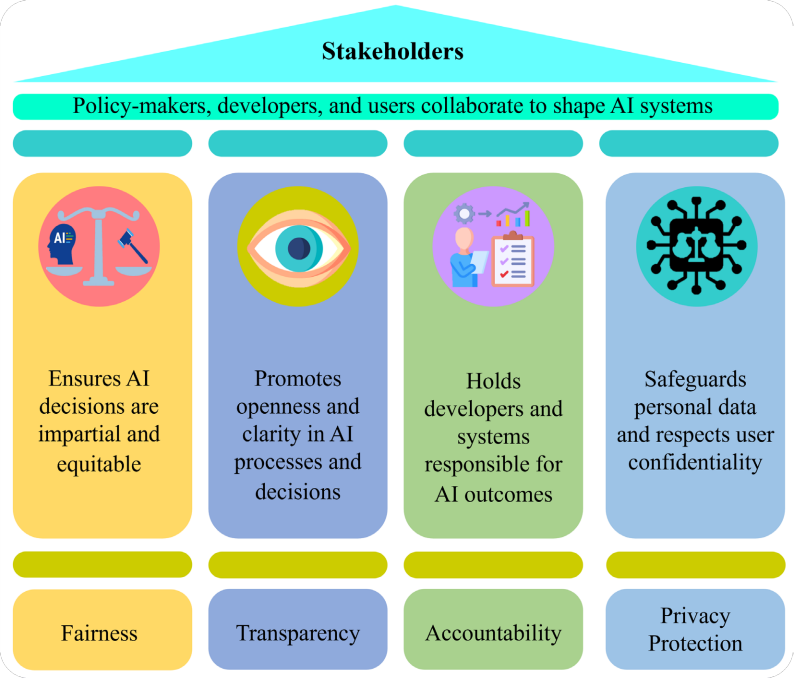

Figure 2. Towards building ethical AI together with stakeholders.

European Commission to the newly-established AI Office, where the enforcement of relevant (AI) laws is structured and executed across national and supranational areas [37]. As one example of a well-formed legal approach to regulating AI in the EU, the legal package in Europe aims to regulate aspects of AI in a way that addresses concerns while encouraging innovation [38]. On the other hand, self-regulatory frameworks are typically industry-driven initiatives that enable organizations to create their governance models emphasizing flexibility and innovation while addressing the risks of AI [39]. Such frameworks are critical in industries where the pace of technological development outstrips formal regulatory processes. Co-regulatory mechanisms combine the best features of regulatory and self-regulatory models, combining government oversight with industry involvement. This hybrid model is naturally significant in guaranteeing public security and human rights, but it also safeguards an environment of technological innovation [39]. Previous studies emphasize these frameworks' relevance at different governance levels, such as team and international, to appropriately mitigate AI risks and apply adequate governance practices [40]. These varied strategies are part of an international movement to create effective AI regulation consistent with social philosophy and technical development. This section leverages methods and explains normative architectures that would describe the responsible AI.

4. Investing in AI Safety Research

The most important thing we can be doing is investing in AI safety research and figuring out what dangerous failure modes these systems could have, especially those that use machine learning or reinforcement learning techniques. **Figure 3**

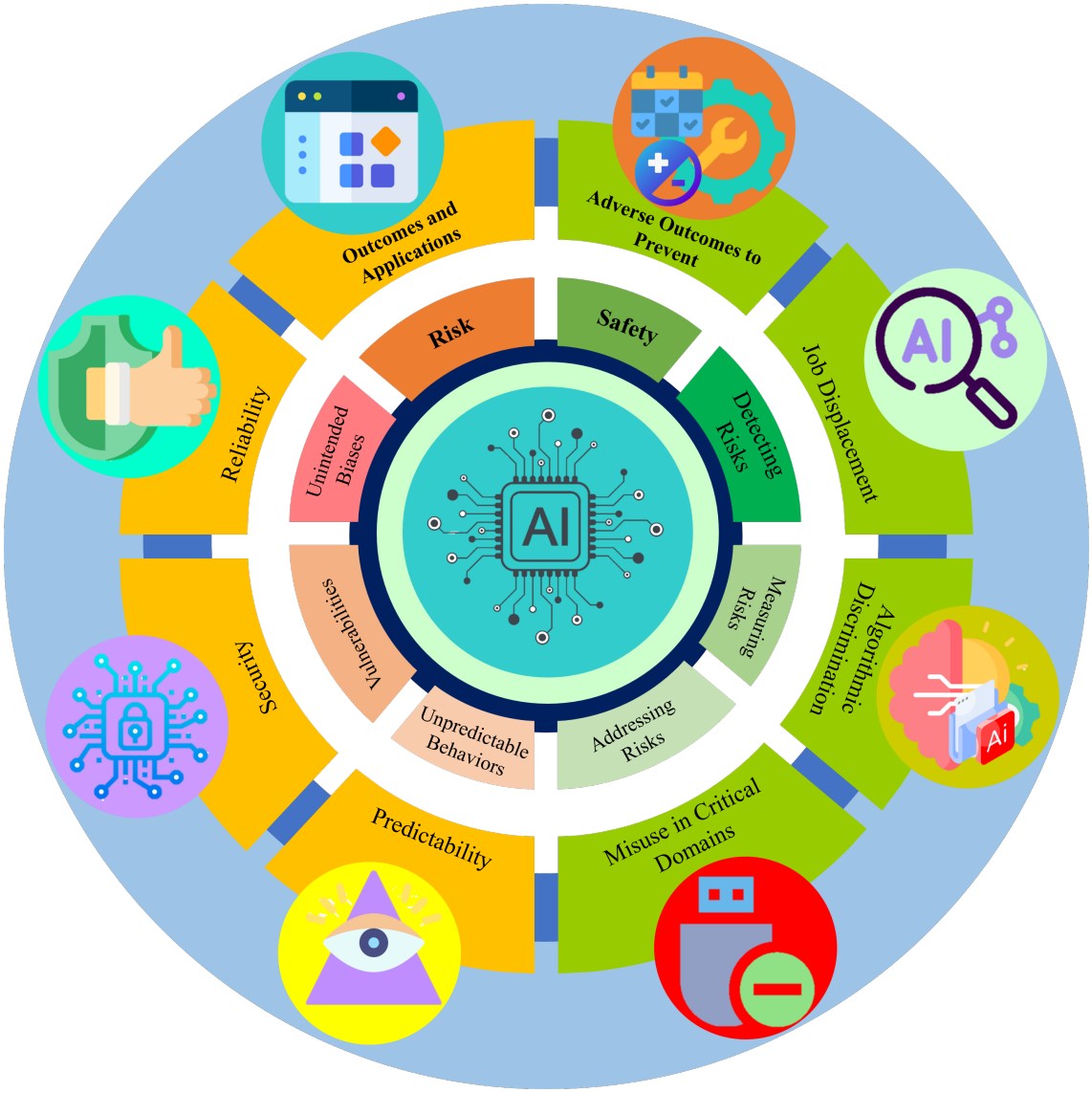

Figure 3. Essential components and connections of AI safety research.

highlights important components of AI safety research investment, and this number underscores the importance of safety research to help mitigate risks from bias, bugs, and other unexpected behaviour in AI systems. They can be unexpectedly biased or flawed, novel in harming ways that put the public at significant risk. For instance, reinforcement learning (RL) agents can exhibit dangerous behaviours if not well-aligned, particularly in safety-critical applications such as autonomous vehicles and healthcare [41]. The theory of safe reinforcement learning (SafeRL) seeks to enable RL agents with unrelated goals and secure behavioural skills [41]. Additionally, the SafeRL algorithm implementation is complex and comes in many challenging ways, requiring one unified, effective lean framework for training. In addition, excitement and uncertainty from the rapid adoption of more advanced AI models have led to significant funding by large AI corporations, such as the UK's £100 million investment in a new "Foundation Model Taskforce" [10].

Nevertheless, while the sociotechnical requirements of real AI existential risk are not met by the standard technical agenda for AI safety [10], it is more comprehensive and politically viable with appropriate iteration. From a software engineering perspective, long-term AI safety concerns the prevention of harm from scaling as capabilities increase above the human level in both functional and programmatic domains toward artificial general intelligence (AGI) or high-level machine intelligence (HLMI) [42]. These discussions are critical yet absent from software engineering venues. This gap must be closed to support favourable future AI/safety and SE developments. Robust methodologies for identifying, quantifying, and mitigating these risks are thus a key component in improving the trustworthiness of AI systems—increasing their reliability, security, and predictability to ensure that adverse outcomes such as job dislocation from automation or algorithmic discrimination do not occur when using AI outside carefully controlled environments. Highlight AI's dual challenges — bias and job loss — and the need for ethical frameworks and regulatory measures to mitigate these concerns. Bias in AI systems is a significant issue because it can reinforce existing inequalities and discrimination, highlighting the need for sound governance frameworks to promote fairness and accountability [43]-[45]. The EU AI Act sets the standard with its strict guidelines to minimize such bias [46], and it is a regulatory framework that other countries may look to replicate. Note that the foreseen job displacement opportunity characterizes a vision of it as a danger to be countered by policies on retraining for the workforce and adaptation so that technological progress does not fuel unemployment and other tensions but human-AI integration [45]. These discussions indicate that incorporating ethics and increasing public awareness is essential for ethical AI technology deployment [40] [45]. Ethical considerations are operationalized through focused AI safety research.

5. Implementing Robust Testing and Validation

The testing and validation processes need to be even more robust to ensure that the AI systems work dependably in real-world scenarios. Comprehensive testing can detect technical errors, vulnerabilities, and bias in AI algorithms before deployment, which is beneficial because it lowers the chance of system malfunctions or unintended consequences. For example, anticipatory thinking and a more adaptable model risk audit (MRA) framework can allow organizations to operationalize the identification of risks at the level they exist within models by working to deliver responsible AI deployments that move beyond performance evaluation with an emphasis on issues such as robustness checking, secure deployment readiness explainability and fairness throughout its lifecycle [47]. Moreover, automatically generated test cases for AI-based autonomous systems can support coverage and efficiency while at the same time promoting transparency, which is a critical element for making a valid safety case in the adaptive system context of what should happen [48]. The reliability of AI applications is another important challenge because they will need to be designed with high-level standards and adequately

protected from novel risks, such as discrimination towards people while processing personal data [17]. As illustrated in Figure 4, the proposed AI risk management framework defines governance as a cross-domain function that informs and integrates the other three abilities: mapping, measurement, and management of AI risks [49]. Deep lifecycle assessments and other new governance techniques are seen as legally permissible in the industrialized world as ways to address such problems and provide better control mechanisms [11]. The second concern is embedding ethical requirements at management levels—especially in middle- and top-level management—to promote trust [5] by meaning a part of the development process. Rigorous testing and validation will improve AI technologies' reliability and accountability to regulations and public standards, which should increase user/ stakeholder trust. Safety knowledge is formalized and implemented using a robust validation and testing pipeline.

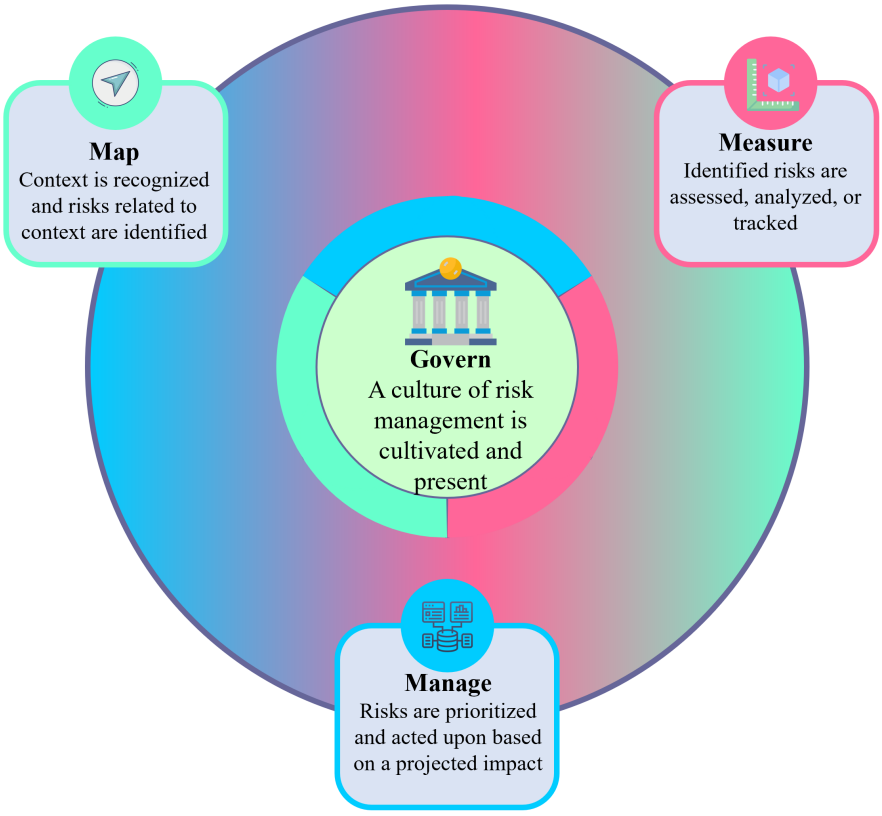

Figure 4. Functions organize AI risk management activities at their highest level to govern, map, measure, and manage AI risks. Governance is designed to be a cross-cutting function to inform and be infused throughout the other three functions, redrawn from [49].

6. Establishing Regulatory Bodies

This is why we need special agencies that check AI implementation for future ethical standards, legislative compliance, and newness issues. To this end, these bodies will be tasked with overseeing cases of the application and outcomes of AI in real-world scenarios, following up on complaints or breaches where they arise to

push forward any regulatory provisions that encourage responsible use. Continual enhancement and deployment of algorithms are required while ensuring safety assurance processes [15], which underscore the necessity for a regulatory framework capable of striking a balance between innovation on one side and ensuring credibility and keeping pace with new technologies on the other. For Europe, the Fourth Industrial Revolution signified a need for pertinent utopian reforms from regulation and adaptation of AI utilization to create opportunities while mitigating risks and ensuring that legal regulations comply with freedom-related human rights [4]. In the absence of systemic regulation, there is a danger that self-regulation may replace this, and we will move further towards unfettered use of AI in business [30], highlighting insufficient controls to ensure widespread implementation at scale can be trusted by businesses. The technical maturity of ethical, trustful, and legal AI is still beginning, while there is a need to shift the regulatory framework to make it evolve from abstract requirements into concrete operational commands for providing tighter oversight throughout the entire lifecycle of AI [11]. While it is correct that global regulatory agencies such as the US Food and Drug Administration are struggling to keep pace with new policies designed to protect patients from poorly performing AI tools, regulations raise important questions about how ethical concerns should be managed and who—a developer of an AI solution or their user—can hold accountability for those breaking the rules [50]. Regulatory bodies can manage the risks associated with AI technologies and support innovation while maintaining social trust by putting in place clear guidelines that are ensured through oversight, which will keep ethical concerns under check and promote accountability.

The different regulatory responses by countries to AI, covering the spectrum of regulation levels, underscores the need for international regulatory consistency in AI governance, which could be facilitated by international organizations such as the Organization for Economic Co-operation and Development (OECD) and the UN. While the European Union's General Data Protection Regulation (GDPR) is considered a high watermark of strict data protection and privacy principles, the decentralized and more market-driven approach in the United States is declared more in keeping with its ideology and economy [45] [51]. On the one hand, China and Japan combine state-led direction with market-driven innovation, exemplifying different regulatory strategies in Asia [45]. The necessity for promoting harmonization of Artificial Intelligence laws and regulations in line with other regulations, such as GDPR, to address challenges and advocate on issues such as bias, transparency, and accountability in AI systems [45] [52]. International organizations such as the OECD and the UN are central to developing harmonized principles and governance models by encouraging flexible regulatory frameworks that reconcile safety, ethics, and innovation [53]. International Regulatory Co-operation (IRC), a practice that describes removing barriers to trade and catering to global economic and technological development, has been led by developed countries that design the IRC systems [54]. The need for standardized safety norms

and international consensus emerges—lessons from the International Atomic Energy Agency (IAEA) nuclear safety regulations offer insights into the challenges posed by the unique risk of AI technologies [52]; therefore, international co-operation on the governance of AI is vital to achieve ethical advancement and amplify social advantages whilst alleviating risks [45] [53]. Effective oversight involves regulatory bodies that accredit, monitor, and enforce.

Comparative Synthesis

However, the legal regimes of AI governance in the EU, the US, and China vary across their frameworks, enforcement mechanisms, and guiding principles. In this way, the EU AI Act creates a risk-based, legally binding framework that is strongly based on the principles of transparency and accountability and on safeguarding individual rights and guarantees in specific high-risk contexts [55] [56]. The US has adopted a sectoral standards-based approach, focusing on existing legislation and voluntary measures like the NIST AI Risk Management Framework to shape industry practice, thereby facilitating innovation; however, it lacks comprehensive regulation [51] [55]. China's approach, however, is that a state-driven governance philosophy, where heavy-handed state control of AI and its use is prioritized, often at the expense of privacy [55], is imposed by way of binding administrative regulations to achieve fast AI deployment in China. Such varied policies not only affect domestic compliance but also play a crucial role in international regulatory dynamics, so that concerted action on the part of governments will be required in order to address and respond to the challenges resulting from AI technologies [57].

7. Encouraging Transparency and Explainability

Promoting transparency and explainability in AI systems is important for building trust and comprehension among both users of the technology, as it can often make decisions that are "black boxes"—i.e., difficult or impossible to interpret from a human perspective. With the help of HCAI, human oversight of AI systems and human decision-making over the processing and reasoning of smart systems will be guaranteed (**Figure 5**) [58]. The XAI has recently gained significant attention as one of the key research areas, which is an effort to grow interpretability through saliency maps, attention mechanisms, rule-based explanations, and model-agnostic approaches [18]. The proposed EU AI Act also embeds the escalation of requirements for transparency and human oversight, even if it is not mandating XAI (which stresses documentation or a clear hand-over), discussing that crystalline build context is essential to establish compliance as well as addressing relevant considerations about black box behaviour inherent in opaque AI systems [21]. Pragmatic approaches such as XAI and auditing standards should be implemented to incorporate ethics in AI and ensure accountability and transparency. They play a significant role in overcoming the issue of the black box in complex AI, which can unlock interpretability in the decision-making process for

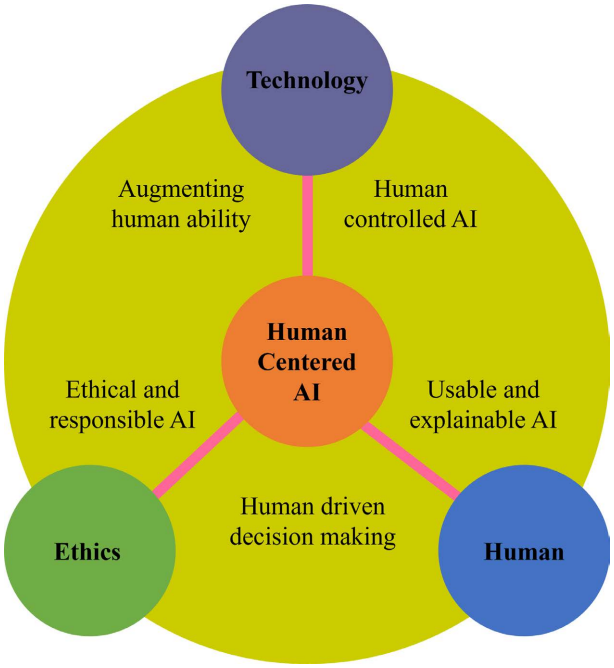

Figure 5. Human-centered AI combines humans, ethics, and technology, redrawn from [58].

AI users [59] [60]. Techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have been significant in demystifying AI operations, facilitating the identification of bias in AI processes to achieve compliance with ethical standards and essential regulations like GDPR [60] [61].

Furthermore, applying XAI in autonomous systems can also significantly improve safety and accountability in high-stakes situations such as healthcare and finance that rely on complex, nontransparent algorithms [61] [62]. Third-party audits and ethics reporting frameworks strengthen accountability by creating responsibility for various stages of AI development, thus connecting a theoretical approach to ethics in AI to actionable solutions [59]. The multidisciplinary processes contribute to XAI, which, in turn, will produce socially practical explanations and will ultimately improve public trust [63]. Overall, with the increasing implementation of AI in sensitive systems, the XAI's role in enhancing transparency and accountability is only expected to grow, fostering responsible AI innovation and application [60] [61]. Although transparency is generally seen as an ideal, there are mixed opinions on its need for implementation due to research that has demonstrated the inclusion of algorithmic details that are more theoretically vague, such that they can enable dismissal or faulty assumption [64]. In addition, the incipient domain of deceptive AI is highlighted as a counter-story to transparency; instead, not all the AI systems would be fully transparent, and there might be better human-AI interactions when deception strategies were enabled on behalf of some algorithms [64]. However, the fragility of trust and ethical concerns demand more nuanced considerations. Various visual explanation

techniques, such as Grad-CAM, Ablation-CAM, Score-CAM, and Eigen-CAM, are being examined to reveal the decision-making processes of convolutional neural networks, thereby improving transparency and accountability in AI systems [65]. Providing AI systems with interpretable explanations for their decisions can alleviate concerns around bias, discrimination, and ethical issues, drive responsible use of AI in different industries, and eventually help establish a more reliable, transparent ecosystem where AI can be trusted. Transparency and explainability yield accountability and trust.

8. Fostering Public Awareness and Education

As a result, raising public awareness and education on AI is very important in helping foster more informed decision-making and correcting common misconceptions. In this regard, as AI technologies are increasingly deployed in society, the public needs to be aware of what will benefit us and where its opportune deployment falls under grounds that bear risks alongside ethical concerns. Targeted educational techniques, as demonstrated through an eight-week course called "AI in Everyday Life," are necessary to enable more of the public to understand better the capabilities and limitations of AI-powered tools [66]. Given the general lack of public awareness about AI compared to other technology areas, improving AI literacy is something for everybody—from childhood schooling to adulthood [67]. Given the critical role AI plays in shaping information environments to which that public sphere is exposed—e.g., social media platforms but also more broadly [68]. It seems imperative to create awareness among the wider population about how those tools affect societal visibility and agenda-setting of truly democratic undertakings.

Furthermore, AI art enables the public to develop more efficient collective literacies of what AI is and does by connecting technical systems and structural powers while teaching, experiencing, and translating comprehension into interpretation rather than just information [69]. The new disruptive fallen earth caused by the powerful AI technologies like ChatGPT in this era of post-web education should be a serious re-thinking of these predatory educational systems to connect between its current state and well-accelerating reality to maturity to ensure not just quality teaching and similar activities but societal needs [70]. A public education advocacy campaign will empower individuals to engage with AI technologies competently and constructively, guiding the development and deployment of AI by promoting agreed-upon societal values and ethical considerations. Publicity enables an informed public and gives legitimacy.

9. Encouraging Human-AI Collaboration

This is essential as we look at a society where humans work hand in glove with AI yet manage to mitigate its dark side. Rather than delegating human capabilities to a robot, it would be great if this partnership could augment and complement what humans do instead of replacing societal economic functions across sectors like

productivity, creativity, and decision-making. The idea of Human-AI Teaming (HAT) is an example of this approach; however, with AI as a team member (AI as a subordinate agent), not just another tool can compensate for the strengths and weaknesses of each other to reach their joint performance possible level [26]. Enabling human-robot interaction is required without a doubt to facilitate useful collaboration, but human-centered AI must ensure that in the age of AI itself, it remains faithful only to our values and objectives, investing ethically in at least some mutual advantage Human-Centered AI (HCAI) [28]. A focus on user empowerment, ethical considerations, and shared decision-making is needed to build trust and promote users' agency, such as staff. Finally, the emergence and sustenance of collective intelligence in human-AI systems may be supported by developing sociocognitive architectures, which take a holistic approach to socio-technical system design [71]. Behavioural synchronization, such as Intentional Behaviour Synchrony (IBS) is a newfangled technique that may be used to establish trust and cooperation. Among AI decisions with human expectations, certain actions are taken to engender the feeling of similarity between a human partner and an AI counterpart [72]. Organizations embedding these underlying and intertwined insights and frameworks can design AI technologies that not only enable human capacities individually or collectively but can also conform to ethical standards, leading to a possible world having more beneficial impacts of AI on humanity and well-being [71]. In addition, cyberattacks may trigger such conflicts (**Figure 6**), such as false data injection (FDI) on the sensor, which is equivalent to sensor faults in terms of consequences [73]. Collaboration between humans and AI centres on augmentation rather than replacement.

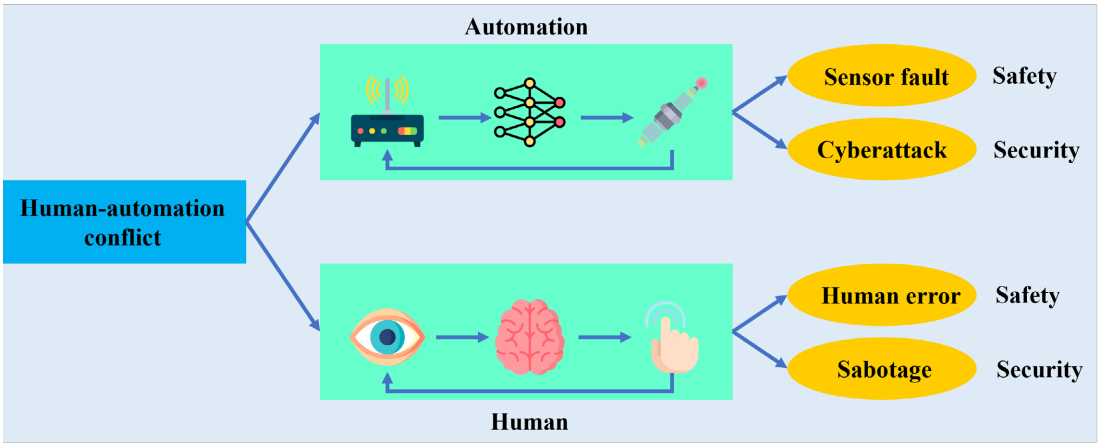

Figure 6. Human-automation conflict, redrawn from [73].

10. Developing Value-Aligned AI

Value-aligned AI, an approach to ensuring that AI systems prioritize human well-being, fairness, and safety while reducing potential harm to humanity, is significant in the broader interest of humanity. The setup of AI technologies includes involving ethical issues in AI development to enable their operations to follow

ethical guidelines and requirements while reflecting societal values. Human-centered AI is characterized by user empowerment through personalized experiences, explainable AI, and consideration of ethical concerns such as fairness, transparency, accountability, or the lack thereof, privacy protection, and ensuring user rights are maintained, or biases are averted [28]. On the other hand, value alignment, capturing the essence of the value alignment problem, hinders the realigning intelligence focus on provably values-aligned intelligence, while social science presents a formal [7] conceptual framework where the formal reasoning focuses on human values [7]. Human-AI collaborative interaction also refers to mutual decision-making, where users have control over the AI and promote their optimal well-being and autonomy, utilizing AI to make AI technologies benefit people and create a better future for humanity [28]. Such an approach, which focuses AI design on the users' needs and through interdisciplinary interaction involving all stakeholders, can enable more ethical use of AI and aversion to challenges like those associated with some AI applications by making its extensive use positively impact society. Value alignment embeds social norms into system goals.

11. Limitations

Additionally, research is needed to map out a clear log of the critical areas of AI regulation. First, there is an urgent need for more granular studies on AI across various applications and respective sectorial regulatory challenges. This is important because different applications of AI are likely to present particular risks and so require bespoke regulatory interventions. Future research should also track how things change over the years within AI development and implementation so regulatory regimes can be timely adjusted as technology progresses. It is also important for the system as AI develops quickly, and unexpected capabilities might be acquired. Existing data and research gaps should be addressed, including greater consideration of the range of regulatory practices in various global settings that impact how AI might be governed. The way the European Union is regulating AI, with an accent on freedom and human rights, will be very different from what it sees in its current tech ethos rivals US or China. So, we need comparative studies to understand best practices. Third, future research should focus more on understanding the practical implementation questions arising from AI regulation policies, such as challenges in making policy decisions and coordination with various stakeholders. This involves creating effective operational rules and accountability mechanisms to ensure the quality of AI systems and legal compliance throughout their lifecycle. Overcoming these practical barriers will allow for more impactful and integrated regulation by all sectors. Sustainability should be taken into account when regulating AI, and the impact on the carbon footprint of AI technologies should be reduced as far as possible for this reason—with human rights instruments correctly balancing between individual claims to predictive processing and collective ecological interests. Overcoming these limitations in future work will further empower AI regulators to create more effective and fair policies that

harness the ethical innovation potential of this technology for society. A major limitation is that views from the Global South are underrepresented, with regulatory agendas there likely to stress infrastructure needs, capacity development, and contextual rights. Future efforts should involve regional experts and cover multilingual material, including case studies, in order to provide balanced recommendations across the world.

12. Future Research Direction

Future research on AI regulation should be geared towards formulating flexible regulatory mechanisms that can keep pace with the fast-paced tech progress. This means developing flexible regulations that can grow at the same rate as AI. That is also why embedding those AI principles into the design and governance of any new technology is essential, as well as focusing on ethical, responsible legal norms that must frame every aspect of societal need. Global and cross-cultural perspectives on regulatory practices to advance understanding of differences in approaches amongst regions, including Europe as evidenced by its "twin strand" approach; the US with an emphasis on freedom flowed through human rights case law associated with MIT v IBM3; China emphasizing innovation (and security) seamlessly grounded-conceptually in benevolence. Academic collaborations require a convergence between disciplines as diverse as law and computer science, coming together with environmental space in response to two dual transformations: digitization and sustainability studies. Inclusive decision-making processes that engage stakeholders are important since AI international law is co-produced and enforced through interactions with multiple actors—private firms, industry associations, civil society, etc. This legal and social framework must be supported by an effective monitoring and evaluation mechanism to test the regulatory effectiveness of societal impact. It may require extensive lifecycle assessments and new governance solutions to fill operating gaps and offer better control mechanisms.

Finally, the plan for further research on AI auditing techniques, impact assessment frameworks, and standardized criteria for ethical assessment ensures ethical oversight and responsible AI deployment. AI auditing—majorly discussed—systematically evaluates AI systems against predefined expectations [74] and is crucial for ensuring these systems comply with legal and industry standards. So, the responsible AI question bank provides a systematic prism for risk assessment, complementing fairness, transparency, and accountability principles with emerging regulations and improved AI governance [75]. Underscores the need for ethical frameworks guiding AI's societal and technical challenges, emphasizing fairness, accountability, and transparency to minimize risks like biases and privacy violations [76]. The Ethical Analysis Framework (EAF) is an approach that systematically assesses fairness, transparency, and accountability in AI systems and highlights the importance of using ethically sound data in shaping AI's moral implications [77]. These takeaways point towards the need for future research to establish robust auditing tools, comprehensive cognitive assessment frameworks,

and standardized metrics to promote ethical and responsible deployment of AI systems and explore methodologies for assessing transparency, accountability, and fairness in AI models.

Moreover, the question of environmental AI sustainability—including transparency mechanisms and design for sustainability—is now being worked on to mitigate climate-related externalities related to carbon-intensive deep learning computation with large models. Through these directions, future research can also play a key role in informing the development of ethical, resilient, and flexible AI regulations to foster innovation in ways that protect broader societal interests and values across various contexts. Adopting this whole-of-government approach to AI will ensure that these technologies are developed and implemented reliably, backing all of society.

13. Conclusion

AI regulation is a complex, multi-domain challenge requiring cross-domain thinking and strategy. A step towards building ethical AI frameworks, explored in this paper, is a step toward aligning these systems with human values and, therefore, with societal norms. This article underscores the complexity of AI regulation and the importance of a balanced strategy that promotes innovation while establishing ethical and oversight measures. The key insights are that investing in AI safety research could help to proactively mitigate some of these risks and the importance of rigorous testing and validation in ensuring the reliability and safety of AI systems. Independent governing bodies can ensure consistent oversight and accountability, and transparency and explainability are crucial for maintaining public trust in AI systems. Further, spreading awareness and enlightening people about the potential and pitfalls of AI will help them responsibly thrive in an AI-fueled world. By making co-habilitation between humans and AI more oriented toward augmentation than competition, we can ensure AI complements human genius. In the end, the future of AI will depend on building value-aligned AI systems, ongoing research, and ethical oversight. The action plan described here is a step in the right direction, but implementing it will necessitate continuous cooperation between policymakers, scientists, industry leaders, and the broader public.

Credit Authors Statement

Hong Yu: Conceptualization, Investigation, Methodology, Formal Analysis, Writing–original draft. Conceptualization, Investigation, Methodology, Visualization, Data curation, Formal Analysis, Resources, Writing–original draft. Writing—review and editing, Supervision, Funding Acquisition, Resources. All authors have read and agreed to the published version of the manuscript.

Ethics Statement

The author has no ethics issues to report.

Acknowledgements

The author wishes to thank the College of Communication and Information Engineering, Chongqing College of Mobile Communication.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Huang, C., Zhang, Z., Mao, B. and Yao, X. (2023) An Overview of Artificial Intelligence Ethics. *IEEE Transactions on Artificial Intelligence*, **4**, 799-819. https://doi.org/10.1109/tai.2022.3194503

[2] Reddy, S. (2023) Navigating the AI Revolution: The Case for Precise Regulation in Health Care. *Journal of Medical Internet Research*, **25**, e49989. https://doi.org/10.2196/49989

[3] Al-Hwsali, A., Alsaadi, B., Abdi, N., Khatab, S., Alzubaidi, M., Solaiman, B., *et al.* (2023) Scoping Review: Legal and Ethical Principles of Artificial Intelligence in Public Health. In: Mantas, J., *et al.*, Eds., *Healthcare Transformation with Informatics and Artificial Intelligence*, IOS Press, 640-643. https://doi.org/10.3233/shti230579

[4] Owczarczuk, M. (2023) Ethical and Regulatory Challenges Amid Artificial Intelligence Development: An Outline of the Issue. *Ekonomia i Prawo*, **22**, 295-310. https://doi.org/10.12775/eip.2023.017

[5] Agbese, M., Mohanani, R., Khan, A. and Abrahamsson, P. (2023). Implementing AI Ethics: Making Sense of the Ethical Requirements. *Proceedings of the* 27*th International Conference on Evaluation and Assessment in Software Engineering*, Oulu, 14-16 June 2023, 62-71. https://doi.org/10.1145/3593434.3593453

[6] Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., *et al.* (2023) AI Ethics Principles in Practice: Perspectives of Designers and Developers. *IEEE Transactions on Technology and Society*, **4**, 171-187. https://doi.org/10.1109/tts.2023.3257303

[7] Osman, N. and d'Inverno, M. (2023) A Computational Framework of Human Values for Ethical AI.

[8] Barletta, V.S., Caivano, D., Gigante, D. and Ragone, A. (2023) A Rapid Review of Responsible AI Frameworks: How to Guide the Development of Ethical AI. *Proceedings of the* 27*th International Conference on Evaluation and Assessment in Software Engineering*, Oulu, 14-16 June 2023, 358-367. https://doi.org/10.1145/3593434.3593478

[9] Anderljung, M., *et al.* (2023) Frontier AI Regulation: Managing Emerging Risks to Public Safety.

[10] Lazar, S. and Nelson, A. (2023) AI Safety on Whose Terms? *Science*, **381**, 138. https://doi.org/10.1126/science.adi8982

[11] Lucaj, L., van der Smagt, P. and Benbouzid, D. (2023) AI Regulation Is (Not) All You Need. 2023 *ACM Conference on Fairness Accountability and Transparency*, Chicago, 12-15 June 2023, 1267-1279. https://doi.org/10.1145/3593013.3594079

[12] Lupo, G. (2023) Risky Artificial Intelligence: The Role of Incidents in the Path to AI Regulation. *Law, Technology and Humans*, **5**, 133-152. https://doi.org/10.5204/lthj.2682

[13] Adnan, M., Xiao, B., Ali, M.U., Bibi, S., Yu, H., Xiao, P., *et al.* (2024) Human

Inventions and Its Environmental Challenges, Especially Artificial Intelligence: New Challenges Require New Thinking. *Environmental Challenges*, **16**, Article ID: 100976. https://doi.org/10.1016/j.envc.2024.100976

[14] Engstrom, D.F. and Haim, A. (2023) Regulating Government AI and the Challenge of Sociotechnical Design. *Annual Review of Law and Social Science*, **19**, 277-298. https://doi.org/10.1146/annurev-lawsocsci-120522-091626

[15] Li, P., Williams, R., Gilbert, S. and Anderson, S. (2023) Regulating AI/ML-Enabled Medical Devices in the UK. *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, Edinburgh, 11-12 July 2023, 1-10. https://doi.org/10.1145/3597512.3599704

[16] Sateli, B., Castillo, F.D. and Moshtagi, R. (2023) Towards Determining the Criticality of AI Applications: A Model Risk Management Perspective. In: *Proceedings of the Canadian Conference on Artificial Intelligence*, Canadian Artificial Intelligence Association, 1-12. https://doi.org/10.21428/594757db.74665221

[17] Poretschkin, M., *et al.* (2023) Guideline for Trustworthy Artificial Intelligence—AI Assessment Catalog.

[18] Thalpage, N. (2023) Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems. *Journal of Digital Art & Humanities*, **4**, 31-36. https://doi.org/10.33847/2712-8148.4.1_4

[19] Theis, S., Jentzsch, S., Deligiannaki, F., Berro, C., Raulf, A.P. and Bruder, C. (2023) Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work. In: Degen, H. and Ntoa, S., Eds., *Artificial Intelligence in HCI*, Springer, 355-380. https://doi.org/10.1007/978-3-031-35891-3_22

[20] Mehrotra, S., Centeio Jorge, C., Jonker, C.M. and Tielman, M.L. (2023) Building Appropriate Trust in AI: The Significance of Integrity-Centered Explanations. In: Lukowicz, P., *et al.*, Eds., *Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, IOS Press, 436-439. https://doi.org/10.3233/faia230121

[21] Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., *et al.* (2023) The Role of Explainable AI in the Context of the AI Act. 2023 *ACM Conference on Fairness Accountability and Transparency*, Chicago, 12-15 June 2023, 1139-1150. https://doi.org/10.1145/3593013.3594069

[22] Floridi, L. and Cowls, J. (2022) A Unified Framework of Five Principles for AI in Society. In: Carta, S., Ed., *Machine Learning and the City: Applications in Architecture and Urban Design*, John Wiley & Sons Ltd., 535-545.

[23] Ashraf, Z.A. and Mustafa, N. (2024) AI Standards and Regulations. In: Qidwai, M.A., Ed., *Intersection of Human Rights and AI in Healthcare*, IGI Global, 325-352. https://doi.org/10.4018/979-8-3693-7051-3.ch014

[24] Jobin, A., Ienca, M. and Vayena, E. (2019) The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, **1**, 389-399. https://doi.org/10.1038/s42256-019-0088-2

[25] Subash, B. and Whig, P. (2024) Principles and Frameworks. In: Bhattacharya, P., *et al.*, *Ethical Dimensions of AI Development*, IGI Global, 1-22. https://doi.org/10.4018/979-8-3693-4147-6.ch001

[26] Xu, W. and Gao, Z. (2024) Applying HCAI in Developing Effective Human-AI Teaming: A Perspective from Human-AI Joint Cognitive Systems. *Interactions*, **31**, 32-37. https://doi.org/10.1145/3635116

[27] Wen, H. (2023) Alert of the Second Decision-Maker: An Introduction to Human-AI Conflict.

[28] Usmani, U.A., Happonen, A. and Watada, J. (2023). Human-Centered Artificial Intelligence: Designing for User Empowerment and Ethical Considerations. 2023 5*th International Congress on Human-Computer Interaction, Optimization and Robotic Applications* (*HORA*), Istanbul, 8-10 June 2023, 1-7. https://doi.org/10.1109/hora58378.2023.10156761

[29] Xu, W. and Gao, Z. (2024) Applying HCAI in Developing Effective Human-AI Teaming: A Perspective from Human-AI Joint Cognitive Systems. *Interactions*, **31**, 32-37. https://doi.org/10.1145/3635116

[30] Analytica, O. (2023) As It Considers Regulation Canberra Must Also Boost AI. Emerald Expert Briefings.

[31] Aptasari, A.D. (2024) Legal Transformation for the Achievement of SDGs: Integration of Multidisciplinary Approaches towards 2030. *RSF Conference Series*: *Business, Management and Social Sciences*, **4**, 154-159. https://doi.org/10.31098/bmss.v4i1.870

[32] Rantala, S., Jabbour, J. and Närhi, J. (2023) Global Environmental Knowledge Synthesis: What's in It for National Action? *Sustainability*: *Science, Practice and Policy*, **20**, Article ID: 2291883. https://doi.org/10.1080/15487733.2023.2291883

[33] de Graeff, N., Jongsma, K.R. and Bredenoord, A.L. (2021) Experts' Moral Views on Gene Drive Technologies: A Qualitative Interview Study. *BMC Medical Ethics*, **22**, Article No. 25. https://doi.org/10.1186/s12910-021-00588-5

[34] Tangyi, B. (2024) Concurrent Evidence: A Framework for Using Evidence from Multiple Disciplines. Leiden Madtrics.

[35] Carcary, M. (2018) Personal Data Protection: Insights in the Digital Context.

[36] Fisher, E.J.P. and Fisher, E. (2023) A Fresh Look at Ethical Perspectives on Artificial Intelligence Applications and Their Potential Impacts at Work and on People. *Business and Economic Research*, **13**, 1-22. https://doi.org/10.5296/ber.v13i3.21003

[37] Novelli, C., Hacker, P., Morley, J., Trondal, J. and Floridi, L. (2024) A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities. AI Board, Scientific Panel, and National Authorities. https://doi.org/10.2139/ssrn.4817755

[38] Nikolinakos, N.T. (2023) EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies—The AI Act. Springer.

[39] Almeida, V., Mendes, L.S. and Doneda, D. (2023) On the Development of AI Governance Frameworks. *IEEE Internet Computing*, **27**, 70-74. https://doi.org/10.1109/mic.2022.3186030

[40] Batool, A., Zowghi, D. and Bano, M. (2025) AI Governance: A Systematic Literature Review. *AI and Ethics*, **5**, 3265-3279.

[41] Ji, J., *et al.* (2023) Omnisafe: An Infrastructure for Accelerating Safe Reinforcement Learning Research.

[42] Gros, D., Devanbu, P. and Yu, Z. (2023) AI Safety Subproblems for Software Engineering Researchers.

[43] Patil, S.M. (2024) Regulatory Frameworks for Ethical AI Development in Coding. *International Journal for Research in Applied Science and Engineering Technology*, **12**, 2018-2024. https://doi.org/10.22214/ijraset.2024.63445

[44] Hullurappa, M. (2024) Exploring Regulatory Dimensions in Computing and Artificial Intelligence through Comprehensive Analysis. *FMDB Transactions on Sustainable Computing Systems*, **2**, 74-83. https://doi.org/10.69888/ftscs.2024.000199

[45] Kashefi, P., Kashefi, Y. and Ghafouri Mirsaraei, A. (2024) Shaping the Future of AI: Balancing Innovation and Ethics in Global Regulation. *Uniform Law Review*, **29**, 524-

548. https://doi.org/10.1093/ulr/unae040

[46] Matai, P. (2024) Comprehensive Guide to AI Regulations: Analyzing the EU AI Act and Global Initiatives. *International Journal of Computing and Engineering*, **6**, 45-54. https://doi.org/10.47941/ijce.2110

[47] Munz, P., Hennick, M. and Stewart, J. (2023) Maximizing AI Reliability through Anticipatory Thinking and Model Risk Audits. *AI Magazine*, **44**, 173-184. https://doi.org/10.1002/aaai.12099

[48] Ebert, C., Weyrich, M. and Vietz, H. (2023) AI-Based Testing for Autonomous Vehicles. SAE Technical Paper 1228.

[49] Tabassi, E. (2024) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST Trustworthy and Responsible AI. National Institute of Standards and Technology.

[50] Gottlieb, S. and Silvis, L. (2023) Regulators Face Novel Challenges as Artificial Intelligence Tools Enter Medical Practice. *JAMA Health Forum*, **4**, e232300. https://doi.org/10.1001/jamahealthforum.2023.2300

[51] Chun, J., Witt, C. and Elkins, K. (2024) Comparative Global AI Regulation: Policy Perspectives from the EU, China, and the US.

[52] Cha, S. (2024) Towards an International Regulatory Framework for AI Safety: Lessons from the IAEA's Nuclear Safety Regulations. *Humanities and Social Sciences Communications*, **11**, Article No. 506. https://doi.org/10.1057/s41599-024-03017-1

[53] Huang, K., Joshi, A., Dun, S. and Hamilton, N. (2024) AI Regulations. In: Huang, K., *et al.*, Eds., *Generative AI Security*: *Theories and Practices*, Springer, 61-98. https://doi.org/10.1007/978-3-031-54252-7_3

[54] Zhao, L. (2023) International Regulatory Cooperation. In: Zhao, L.Y., Ed., *Modern China and International Rules*: *Reconstruction and Innovation*, Springer, 231-261. https://doi.org/10.1007/978-981-19-7576-9_9

[55] Kuzior, A. (2024) Navigating AI Regulation: A Comparative Analysis of EU and US Legal Frameworks. *Materials Research Proceedings*, **45**, 258-266. https://doi.org/10.21741/9781644903315-30

[56] Arora, A.S., Saboia, L., Arora, A. and McIntyre, J.R. (2025) Human-Centric versus State-Driven: A Comparative Analysis of the European Union's and China's Artificial Intelligence Governance Using Risk Management. *International Journal of Intelligent Information Technologies*, **21**, 1-13. https://doi.org/10.4018/ijiit.367471

[57] Coromina, M.P.-U. (2024) Análisis comparado de los distintos enfoques regulatorios de la inteligencia artificial en la Unión Europea, EE. UU., China e Iberoamérica. *Anuario Iberoamericano de Justicia Constitucional*, **28**, 129-156. https://doi.org/10.18042/cepc/aijc.28.05

[58] Chameera De Silva, T.H. (2023) Human-Centered Artificial Intelligence: The Solution to Fear of AI. IX Interactions.

[59] Pappachan, P., Moslehpour, M., Bansal, R. and Rahaman, M. (2024) Transparency and Accountability. In: Gupta, B., Ed., *Challenges in Large Language Model Development and AI Ethics*, IGI Global, 178-211. https://doi.org/10.4018/979-8-3693-3860-5.ch006

[60] Rane, J., *et al.* (2024) Enhancing Black-Box Models: Advances in Explainable Artificial Intelligence for Ethical Decision-Making. In: *Future Research Opportunities for Artificial Intelligence in Industry* 4.0 *and* 5.0, Deep Science Publishing, 136-180.

[61] Rane, N.L. and Paramesha, M. (2024) Explainable Artificial Intelligence (XAI) as a Foundation for Trustworthy Artificial Intelligence. In: *Trustworthy Artificial Intelligence in*

Industry and Society, Deep Science Publishing, 1-27. https://doi.org/10.70593/978-81-981367-4-9_1

[62] Chukwunweike, J., *et al.* (2024) Navigating Ethical Challenges of Explainable AI in Autonomous Systems. *International Journal of Science and Research Archive*, **13**, 1807-1819. https://doi.org/10.30574/ijsra.2024.13.1.1872

[63] Fenoglio, E. and Kazim, E. (2024) AI Explainability, Interpretability, and Transparency. In: Lütge, C., *et al.*, Eds., *The Elgar Companion to Applied AI Ethics*, Edward Elgar Publishing, 66-94. https://doi.org/10.4337/9781803928241.00010

[64] Rogers, K. and Howard, A. (2023) Tempering Transparency in Human-Robot Interaction. 2023 *IEEE International Symposium on Ethics in Engineering, Science, and Technology* (*ETHICS*), West Lafayette, 18-20 May 2023, 1-2. https://doi.org/10.1109/ethics57328.2023.10154942

[65] Bhagya, J., Thomas, J. and Raj, E.D. (2023) Exploring Explainability and Transparency in Deep Neural Networks: A Comparative Approach. 2023 7*th International Conference on Intelligent Computing and Control Systems* (*ICICCS*), Madurai, 17-19 May 2023, 664-669. https://doi.org/10.1109/iciccs56967.2023.10142255

[66] Kasinidou, M., Kleanthous, S. and Otterbacher, J. (2023) Artificial Intelligence in Everyday Life: Educating the Public through an Open, Distance-Learning Course. *Proceedings of the* 2023 *Conference on Innovation and Technology in Computer Science Education*, Vol. 1, 306-312. https://doi.org/10.1145/3587102.3588784

[67] Kasinidou, M. (2023) AI Literacy for All: A Participatory Approach. *Proceedings of the* 2023 *Conference on Innovation and Technology in Computer Science Education*, Vol. 2, 607-608. https://doi.org/10.1145/3587103.3594135

[68] Jungherr, A. and Schroeder, R. (2023) Artificial Intelligence and the Public Arena. *Communication Theory*, **33**, 164-173. https://doi.org/10.1093/ct/qtad006

[69] Hemment, D., Currie, M., Bennett, S., Elwes, J., Ridler, A., Sinders, C., *et al.* (2023) AI in the Public Eye: Investigating Public AI Literacy through AI Art. 2023 *ACM Conference on Fairness Accountability and Transparency*, Chicago, 12-15 June 2023, 931-942. https://doi.org/10.1145/3593013.3594052

[70] Wang, F., *et al.* (2023) Climate Change: Strategies for Mitigation and Adaptation. *The Innovation*, **1**, Article ID: 100015.

[71] Gupta, P., Nguyen, T.N., Gonzalez, C. and Woolley, A.W. (2023) Fostering Collective Intelligence in Human-AI Collaboration: Laying the Groundwork for Cohumain. *Topics in Cognitive Science*, **17**, 189-216. https://doi.org/10.1111/tops.12679

[72] Naser, M.Y.M. and Bhattacharya, S. (2023) Empowering Human-AI Teams via Intentional Behavioral Synchrony. *Frontiers in Neuroergonomics*, **4**, Article ID: 1181827. https://doi.org/10.3389/fnrgo.2023.1181827

[73] Wen, H., Khan, F., Ahmed, S., Imtiaz, S. and Pistikopoulos, S. (2023) Risk Assessment of Human-Automation Conflict under Cyberattacks in Process Systems. *Computers & Chemical Engineering*, **172**, Article ID: 108175. https://doi.org/10.1016/j.compchemeng.2023.108175

[74] Cen, S.H. and Alur, R. (2024) From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. *Proceedings of the* 4*th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, San Luis Potosi, 29-31 October 2024, 1-14. https://doi.org/10.1145/3689904.3694711

[75] Lee, S., *et al.* (2024) Responsible AI Question Bank: A Comprehensive Tool for AI Risk Assessment.

[76] Firmansyah, G., Bansal, S., Walawalkar, A.M., Kumar, S. and Chattopadhyay, S. (2024) The Future of Ethical AI. In: Gupta, B., Ed., *Challenges in Large Language Model*

Development and AI Ethics, IGI Global, 145-177.
https://doi.org/10.4018/979-8-3693-3860-5.ch005

[77] R, S., Gambhir, V. and Seth, J. (2024) Investigating the Ethical Implications of Artificial Intelligence and Establishing Guidelines for Responsible AI Development. 2024 *International Conference on Advances in Computing Research on Science Engineering and Technology* (*ACROSET*), Indore, 27-28 September 2024, 1-6.
https://doi.org/10.1109/acroset62108.2024.10743915