AI ETHICS AND GOVERNANCE: TOWARDS THE IMPLEMENTATION OF THE HUMAN-CENTERED AI (HCAI) FRAMEWORK

ALIZA D. RACELIS

ABSTRACT

Artificial intelligence (AI) has profoundly transformed our lives. Concerns about AI's evolution and potential dangers have led to the development of regulations in various regions. While the regulatory approaches differ, they share a goal: ensuring AI benefits society while minimizing negative impacts. This paper argues that these regulations should incorporate the principles from the Human-Centered AI framework. This is a shortened version of a paper delivered at the *Northern Philippines Business Research Conference* in February 2025.

Keywords: Artificial Intelligence • Governance • Human-centered AI •

Artificial intelligence (AI) has profoundly changed and will continue to change our lives. AI is being applied in an increasing number of fields and scenarios, such as autonomous driving, medical care, media, finance, industrial robots, and internet services. The widespread application of AI and its deep integration with the economy and society have improved efficiency and produced benefits. At the same time, it will inevitably impact the existing social order and raise ethical concerns. Ethical issues, such as privacy breaches, discrimination, unemployment, and security risks brought about by AI systems, have generated significant concern among individuals. Consequently, AI ethics has become an important research topic in academia as well as a topic of common concern for individuals, organizations, societies, and countries.

Recent developments in AI have generated significant interest from media and the general public. As AI systems (e.g., robots, chatbots, avatars, and other intelligent agents) evolve from being perceived as tools to being perceived as autonomous agents and teammates, an important focus of research and development is understanding the ethical impact of these systems. Critical questions have emerged: What does it mean for an AI system to make a decision? What are the moral, societal, and legal consequences of their actions and decisions? Can an AI system be held accountable for its actions? How can these systems be controlled once their learning capabilities bring them into

states that are possibly only remotely linked to their initial, designed setup? Should such autonomous innovation in commercial systems even be allowed, and how should AI use and development be regulated? These and many other related questions are currently the focus of much attention.

The Philippines Context

In the Philippines, the challenges posed by AI must be understood within the broader context of business ethics concerns identified by educators and practitioners—particularly the country's widespread poverty that persists despite impressive economic growth. Business ethics professors and practitioners are encouraged to ensure the ethical use of AI in business operations, especially how algorithms can produce outcomes that lead to unintended consequences, such as discrimination, job displacement, privacy violations, and other societal impacts. The Philippines is steadfast in achieving the 17 Sustainable Development Goals (SDGs) by 2030. AI-based research initiatives of its Department of Science and Technology (DOST) are anchored in these goals to achieve better healthcare, economic growth, clean energy, smart cities, smart farming, and climate change mitigation. However, the Philippines' House Bill No. 7913 primarily focuses on establishing the Philippine Artificial Intelligence Council and the AI Research and Development Program. While the bill acknowledges the potential benefits of AI, it falls short in explicitly addressing crucial aspects such as the ethical implications, fairness, potential biases, and the societal impact of AI technologies.

Global AI Ethics Standards and Frameworks

The adaptation of principles and concepts for AI ethics should be based on internationally recognized standards. In November 2021, UNESCO adopted the *Recommendation on the Ethics of Artificial Intelligence*, marking a significant milestone in developing global standards for AI ethics. Supported by all 193 member states, this recommendation serves as a normative framework to address ethical concerns related to AI and to foster trustworthiness throughout the AI system life cycle. It places transparency, fairness, and the protection of human rights and dignity at its core. Along with these, the Center for AI and Digital Policy (CAIDP) emphasizes addressing the connection between AI and human rights. CAIDP, a non-profit organization, is committed to ensuring that advancements in AI contribute to a more equitable and fair society. It advocates for a world where technological advancements are made in hrmony with respect for human rights, rule of law, and democratic institutions.

Like the European Union's General Data Protection Regulation (GDPR) in 2018, the EU AI Act could become a global standard, determining to what extent AI has a positive rather than negative effect on life. The EU's AI regulation is already making waves internationally. In late September

2024, Brazil's Congress passed a bill creating a legal framework for artificial intelligence. There are, however, several loopholes and exceptions in the EU law. These shortcomings limit the Act's ability to ensure that AI remains a force for good.

UNESCO's Recommendations on the Ethics of AI appear to present the most robust AI guidelines among the global guidelines. Their recommendations have set the standard and served as a benchmark for developing other AI guidelines. They recommend adapting principles for an ethical framework that promotes responsible development and use of AI technologies. UNESCO's guidelines emphasize the importance of human rights, transparency, explainability, and accountability in AI systems.

Navigating the Regulatory Landscape

We can define regulation of AI broadly as including not only legislation and government policies but also professional norms and technical standards. Central to this task is the question, What parameters are required? Although national and international government bodies play a defining role here, other players are also influential. Defining rules for something as extensive, complex, and versatile as a system technology brings numerous challenges, problems, and dilemmas. One of the best known is the "Collingridge dilemma." On the one hand, a new technology is difficult to regulate in the early phase because much remains unclear regarding its workings and effect. Moreover, the need for regulation is initially less apparent. Later, once the technology's effects on society are more conspicuous, it becomes clear what regulation is needed and why. By then, however, many of the decisions taken earlier are difficult to reverse. A further complication is that power structures develop around a technology, and these cannot be modified easily or quickly. The Collingridge dilemma is exemplified by the architecture of the internet, which was developed in a spirit of openness and market freedom. Today it is clear that many safety and security issues were not adequately addressed by the original design. Rectification of these design flaws at this stage would require large sections of the internet to be completely restructured. .

The Current State of AI Governance

Embedding or integrating AI into society depends on the existence of frameworks, and therefore regulation. Now that the technology is making the transition from the lab to society, its effects on the economy and society are subject to widespread scrutiny. This has led to debate about the nature of the regulatory measures needed to ensure that AI is properly integrated in society and government processes. Attention has focused not only on the opportunities but also on AI's potential negative consequences. Hundreds of guidelines, codes of conduct, private standards, public-private partnership models and certification schemes have been developed with a view to both promoting opportunities and addressing adverse repercussions. One of the

more important initiatives is the *European Union's AI Act*. Many existing legal provisions and frameworks are potentially applicable to AI, ranging from fundamental rights to liability law, intellectual property rights and the rules on archiving and evidence. In other words, the effects of AI are now controlled through a wide range of frameworks and specific rules, many more of which are likely to be laid down in the years ahead.

Distinguishing Human Intelligence from Artificial Intelligence

Given these challenges, a new ethic of technological development, based on the unconditional priority of public interest and security of the individual, ought to be developed. A critical distinction must be made: the distinction between human intelligence and "artificial intelligence." According to Turing, artificial intelligence mimics humans in the process of preparing and making decisions. This kind of intelligence is very useful in organizational activities, as it offers opportunities to improve human performance by extracting relevant information from large datasets and by predicting unexpected events, by doing so in a fraction of the time it takes humans to do it. Through its imitative abilities, AI is able to identify information patterns that optimize work-related trends. However, humans possess cognitive abilities that represent true intelligence – human intelligence. Being in an open system, humans must respond accordingly to exogenous influences. This mode requires a creative approach to the formation of future strategy, manifested in the ability to correctly respond to sudden changes in the situation and to anticipate the possible developments, as well as to correctly perceive distorted information. All this requires a rational and radical concept of "responsibility."

The Race for Trustworthy AI

It has been argued that a race to AI regulation ought to be pursued, with ever-louder calls being made for regulators to look beyond the benefits and ensure that AI is trustworthy – that is, legal, ethical, and robust. Besides minimizing risks, such regulation could facilitate AI's uptake, boost legal certainty, and also contribute to advancing countries' positions in the race. Indeed, a new playground for global regulatory competition seems to be emerging, which in the best-case scenario pushes governments—amid uncertainty as to the technology's impact, the impact of regulatory intervention, and the cost of non-intervention—to find the most appropriate balance between protection and innovation. By striving for such balance in their own distinct manners, countries can compete through regulation to attract those ingredients that render them a competitive force on the global AI market, while exploring the best approaches to protect their citizens.

The Rome Call and Interfaith Perspectives

The Rome Call for AI Ethics (www.romecall.org), finalized in February 2020, committed signatories to follow principals of transparency, inclusion, accountability, impartiality, reliability, security, and privacy. Religious faiths have played and will continue to play a role in shaping a world in which human beings are at the center of the concept of development. It was argued at the February 2020 event that the ethical development of AI must be approached from an interfaith perspective. In the face of radical transformations that digital and intelligent technologies are producing in society, the three Abrahamic religions together provide guidance for humanity's search for meaning in this new era.

The Emergence of Human-Centered AI

While the technology-centered approach has dominated the development of AI technology, researchers have individually explored a range of human-centered approaches to address the unique issues introduced by AI technology. These include humanistic design research, participatory design, inclusive design, interaction design, human-centered computing, and social responsibility. To respond to AI ethical challenges, Stanford University established a Human-Centered AI (HCAI) research institution, focusing on ethically aligned design. HCAI suggests strategies that support human self-efficacy, creativity, responsibility, and social connections. Researchers, developers, business leaders, policymakers, and others are expanding the technology-centered scope of artificial intelligence (AI) to include HCAI ways of thinking. This expansion from an algorithm-focused view to embrace a human-centered perspective can shape the future of technology to better serve humanity.

Implementing the HCAI Framework

I recommend the Human-Centered Artificial Intelligence (HCAI) framework for designing and assessing AI systems and tools. HCAI clarifies how to (1) design for high levels of human control and high levels of computer automation so as to increase human performance, (2) understand the situations in which full human control or full computer control is necessary, and (3) avoid the dangers of excessive human control or excessive computer control. Achieving these goals will support human self-efficacy, creativity, responsibility, and social connections. In summary, AI ought to amplify, augment, enhance, and empower people. Educators, designers, software engineers, product managers, evaluators, and government agency staffers can build on AI-driven technologies to design products and services

> ***"A greater emphasis on human-centered AI will reduce fears of AI's existential threats and increase benefits for users and society in business, education, healthcare, environmental preservation, and community safety."***

that make life better for users, enabling people to care for each other. A greater emphasis on Human-Centered AI will reduce fears of AI's existential threats and increase benefits for users and society in business, education, healthcare, environmental preservation, and community safety.

Further Reading

1. Alfiani, F. R. N. & Santiago, F. (2024). A comparative analysis of artificial intelligence regulatory law in Asia, Europe, and America. *SHS Web of Conferences*. https://doi.org/10.1051/shsconf/202420407006
2. European Union (2024). *Artificial intelligence act (Regulation (EU) 2024/1689)*. https://artificialintelligenceact.eu/the-act/
3. Rosales, M. A.; Magsumbol, J. V.; Palconit, M. G. B.; Culaba, A. B.; & Dadios, E. P. (2020). Artificial intelligence: The technology adoption and impact in the Philippines. *27th International Computer Science and Engineering Conference (ICSEC)*. https://doi.org/10.1109/ICSEC59635.2023.10329756
4. Shneiderman, B. (2022). *Human-centered AI. United Kingdom: Oxford University Press.*

Aliza Racelis gifting Executive Director Dr. Julio Amador (Fulbright Philippines) with a copy of her textbook Business Ethics and Social Responsibility, December 2022.

Biography

Aliza Racelis received her PhD in Business Administration from the University of the Philippines in April 2010. She is a Management and Business Ethics professor at the University of the Philippines Business School. Her current research interests are in the areas of Business Ethics, Corporate Governance, Virtue Theory, Social Responsibility, Transcendental Leadership, and Sustainability. In 2018, she was granted a Fulbright scholarship to do research and teaching in the U.S. during the Fall Term.