AI Governance: Exploring Key Regulatory Approaches



George Mason and Greg Chance look at the state of global AI regulations. They review the characteristics of key regulatory approaches that have been proposed and discuss their strengths and weaknesses as well as the future of AI regulation.

Artificial Intelligence (AI) has experienced considerable growth over the past decade, primarily due to the significant advancements made possible through deep learning techniques. The current generation of AI markedly outperforms its predecessors, being capable of a diverse range of tasks that previously were exclusive to skilled humans. At some activities, AI can even exceed human performance.

As AI continues to become proficient at human tasks and those beyond human capabilities, there is mounting concern about how it may impact society. Specifically, some worry that AI may be detrimental to humanity. This has led to an increase in demand for *trustworthy* AI that retains the benefits of AI whilst minimising its risks.

The concept of 'trustworthy' AI is loosely defined, with numerous definitions of varying degrees of formality appearing across the literature. Furthermore, alternative words such as 'responsible', 'assured', 'ethical' and 'safe' are sometimes used in place of trustworthy. Regardless of the exact term used, common characteristics that appear in definitions for trustworthy AI include fairness, transparency, robustness, accuracy, accountability and security.

To ensure the use and development of trustworthy AI, jurisdictions worldwide have begun drafting AI regulations that outline rules and principles designed to instil trust in AI systems. In this article, we delve into the regulatory landscape for AI, beginning with the motivation for such regulations, followed by an analysis of the characteristics of key regulatory approaches and then a discussion of their respective strengths and weaknesses and the future of AI regulations.

The Need for Regulations

Until recently, the risks associated with AI were minimal or only applicable in specific circumstances: AI had limited capabilities, meaning there was less criticality, and even the most capable forms of AI typically required specialist equipment and expertise to use.

Presently, however, AI can perform numerous tasks at a human level and is even accessible on consumer-grade equipment by people with little or no expertise. These factors significantly increase the potential risks that AI poses to society, including:

- Mass unemployment – As AI continues to master new abilities, human jobs are becoming increasingly vulnerable. Unlike other technological advances throughout history, the flexible nature of AI to learn an ever-widening range of skills could make it challenging for people to find new roles in the face of redundancy.
- Privacy violations – AI can be used in ways that violate privacy; for example, mass surveillance through facial recognition systems. Moreover, the use of large amounts of data to train AI systems creates the potential for mishandling and leakage of personal information.
- Bias and discrimination – Improperly trained AI systems can exhibit biases and discriminatory behaviours. For instance, when AI is used in the recruitment process, unjust hiring decisions may be made based on factors that should not affect the selection procedure.
- Physical or psychological harm – AI systems installed within physical environments, such as self-driving cars, could behave dangerously in situations where the AI encounters unseen data. Additionally, excessive reliance on tools like chatbots could lead to issues such as social isolation and a weakening of critical-thinking skills.
- Social unrest – Generative AI techniques can synthesise disruptive yet false content. For example, deepfakes that depict notable figures presenting misinformation or chatbots that can produce convincing yet fabricated information. This problem is magnified as the content propagates across social media.

Concerns such as these and more motivate the need to regulate the use and development of AI systems.

Key Regulatory Approaches

The terrain of AI regulations is rapidly evolving. The exact timeline for when the first regulations were proposed is unclear, but most key efforts have occurred within the past three years. As interest in trustworthy AI continues to grow, it is expected that new regulations will be created whilst existing proposals undergo regular revisions.

Notable efforts towards AI regulation have been made by the UK, EU, China, Brazil and Canada. Other jurisdictions such as India are also engaged in discussions on the development of AI regulations; however, their progress is less advanced. Interestingly, the US has

not yet proposed a comprehensive set of federal regulations for AI; although, it has introduced numerous laws and guidelines relating to this field.

Except for China, to date, no jurisdictions have yet finalised and enforced regulations at a national or international level. The following is an overview of those regulations that are most mature.

UK



In March 2023 (last updated in August 2023), the UK published a whitepaper outlining its proposed AI regulations [1]. Being so recent, they may be subject to significant revisions. A key priority of the regulations is to foster innovation amongst AI developers by minimising the burden of stifling legislation.

The regulatory framework is principle-based, encompassing five key principles:

1) safety, security and robustness
2) appropriate transparency and explainability
3) fairness
4) accountability and governance and
5) contestability and redress

To ensure the longevity of the regulations' relevance, the whitepaper provides the broad definition of AI as being a system that is adaptive and autonomous.

The regulations are sector-based, involving existing regulatory bodies to enforce the principles appropriately for each sector. Instead of detailing penalties for noncompliance within the regulations themselves, it falls to the relevant regulatory body to determine appropriate sanctions as needed.

The regulators tasked with enforcing the principles will vary by sector; at present, it remains uncertain exactly who will be responsible for each area. However, it is likely that the Digital Regulation Cooperation Forum (DRCF) may take an overarching role in this regard.

EU



Of all AI regulations still being drafted, the EU's are closest to being finalised. First proposed in April 2021, the drafting process is nearing its conclusion; the European Parliament intends to begin their enforcement in 2024 [2].

A key aim of the EU regulations is to uphold human rights. To achieve this objective, it adopts a tiered, risk-based approach that categorises AI systems based on their potential disruption to society into four levels of risk: unacceptable, high, low and minimal. Furthermore, the regulations empower those negatively affected by AI to seek redress.

The regulations' definition of AI is based on the one given by the Organisation for Economic Cooperation and Development (OECD). In contrast to the UK's definition, this definition is more detailed, specifying behaviours and techniques that characterise an AI system.

Enforcement of the regulations will occur at both national and international levels. Nationally, each member state of the EU is responsible for enforcing the regulations within their country; the specific regulatory bodies who will be responsible for enforcement may vary by country.

Internationally, enforcement will be overseen by the European Artificial Intelligence Board, comprising a representative from each member state. The regulations outline specific financial penalties for noncompliance, with a maximum penalty of up to 35 million euros or 7% of annual worldwide turnover.

China



As of the time of writing, China is the only jurisdiction to have implemented AI regulations at a national level: in March 2022, it enforced its regulations on the management of algorithmic recommendations [3]. Since then, China has continued to develop and enforce further regulatory measures.

The motivations behind Chinese regulations differ somewhat to other jurisdictions. China has an ambitious goal to become the world leader for AI development and these regulations represent a substantial step in this direction. Furthermore, they are important for managing the creation and dissemination of content with respect to ensuring national interests.

A further difference in the Chinese regulations is that they do not appear to provide a general-purpose definition of AI. Instead, they list specific technological behaviours relevant to the purpose of each regulation; this may encompass technologies not traditionally considered to be AI whilst potentially excluding others broadly recognised as such.

Despite these differences, the regulations do have some similarities to regulations from other jurisdictions, such as stipulating that AI systems are accurate, fair and safe.

The regulations are enforced by the Cyberspace Administration of China, with penalties for noncompliance seemingly financial in nature.

Brazil



Brazil has proposed various AI regulations dating back to 2019; however, these were met with limited approval from Congress. The most recent proposal, which replaces previous ones, was introduced in May 2023 [4]. It shares many similarities to the EU's, from which it likely drew inspiration.

Similarities include using the OECD's definition of AI as a basis and adopting a risk-based approach. Additionally, both proposals prioritise human rights; however, the Brazilian regulations seem to emphasise the importance of this even more so by making it central to their regulations.

The main enforcement body for the regulations is yet to be determined; however, it is likely to be the Brazilian Data Protection Authority (ANPD). Penalties for noncompliance vary, including warnings; fines of up to 50 million Brazilian reals or 2% of annual revenue; publicising the infringement; and prohibition or limitations placed on the offender's ability to do further AI development.

Canada



Like Brazil, Canada's approach to AI regulation (proposed in June 2022) shares certain similarities to the EU's; in particular, prioritising human rights and defining AI based on the OECD's definition [5].

In common with the UK's proposed regulations, Canada follows a principle-based approach, emphasising the importance of transparency, accountability, data protection and ethics whilst minimising disruption to innovation. It is possible that the UK was inspired by Canada when drafting their regulations.

Enforcement of the regulations is likely to be centralised, with the task falling to The Minister of Innovation, Science and Industry in conjunction with an AI and Data Commissioner. Violations of the regulations will incur financial penalties of up to 25 million Canadian dollars or 5% of total revenue.

US



The US has not yet proposed a comprehensive set of federal regulations for AI; although, some regulations have been implemented at a state level. So far, the US has primarily approached AI governance through the development of individual laws tailored to specific contexts.

However, in October 2023, the White House published an executive order on the safe, secure and trustworthy development and use of AI, outlining eight guiding principles for the creation of trustworthy AI systems [6]. Although the executive order is not a comprehensive set of regulations, it suggests that one may eventually follow.

Discussion

AI regulations are still a new concept, with only China having finalised and enforced any such regulations at a national level. All other regulatory efforts are still in the drafting stage.

Among those under development, the EU's regulations are closest to completion; they are expected to be enforced during 2024. These regulations will likely prove highly influential on the development of other AI regulations worldwide. Indeed, it seems that they already have been instrumental in the development of Brazil's regulations and may have served as inspiration for Canada's.

The UK aims to strike a balance between ensuring trust and promoting innovation amongst developers.

The UK's regulatory framework presents an interesting variation compared to others. Whereas most other regulations primarily focus on ensuring that AI is trustworthy, the UK aims to strike a balance between ensuring trust and promoting innovation amongst developers. China's regulations also stand out due to their unique blend of trustworthiness-focussed aspects and provisions aimed at ensuring national security by managing the creation and distribution of content.

Although the US has not yet drafted any regulations, given their global prominence in the development of AI and the issuance of an executive order focussed on the need for trustworthy AI, it can be expected that federal AI regulations are likely to be created in the future.

A consideration that does not seem to be addressed by the regulations is how they are to be enforced at an individual level, such as someone using their personal computer within their home. With the proliferation of open-source models and the ready availability of tools and equipment to work with AI, combined with the abundance of free educational resources

available to learn how to develop and use AI models, it is relatively straightforward for anyone with a modest budget and an internet connection to engage in AI use and development. Whilst the personal use of AI poses fewer risks compared to an organisation or government using it (for instance, mass unemployment is an unlikely risk), numerous other risks persist, such as the capacity to generate misinformation or harm mental wellbeing. It remains unclear how the proposed regulations aim to address this issue; it appears that the current regulatory proposals focus primarily on the use and development of AI by organisations and governments.

Another factor to consider is the effectiveness of the specified penalties for noncompliance. Where they are defined, sanctions are typically in the form of fines; however, if an organisation can enjoy substantial profits from noncompliant AI systems and these profits outweigh any financial penalties, even significant fines may not serve as a deterrent due to the net profit after fines still being greater than that generated by compliant AI systems. In this regard, the UK's regulations may be at an advantage, since penalties will be determined based on what is deemed appropriate within a given sector. This approach could potentially inspire other jurisdictions to follow suit if blanket financial sanctions are found to be ineffective.

The definition of AI provided within each set of regulations is another area that may undergo revision. Definitions range from broad in scope to specific and detailed. Each approach has its advantages and disadvantages: broader definitions can potentially apply to future forms of AI which may function significantly differently to current ones; however, it could become unclear to stakeholders and regulators what types of technology qualify as AI without clear delineation. Stricter definitions have the opposite advantages and disadvantages. It will be interesting to observe how effective each definition proves in practice.

An intriguing question is which regulations are likely to become dominant on the global stage; which, if any, will become the global standard? Factors that could influence this include trade, leadership in AI development, political influence and the proven track record of effectiveness of the regulations. In the current geopolitical climate, however, it is unlikely that a single approach will emerge as dominant; instead, it seems likely that multiple regulations will become established and coexist for the foreseeable future. Stakeholders will need to be cognizant of these different regulations if they wish to conduct business in those jurisdictions.

An intriguing question is which regulations are likely to become dominant on the global stage; which, if any, will become the global standard?

A final point to consider is that regulations represent only one aspect of the strategy to ensure trustworthy AI. Other aspects include the development of tools and techniques to assess the risks of AI systems and to ensure they satisfy the criteria for trustworthiness. Although considerable progress has been made in this area through continuing research efforts, much work remains to be done. Similarly, further work is needed to crystallise the process of enforcing the regulations; specifically, identifying the regulatory bodies responsible for this task and outlining the procedures that stakeholders must adhere to in order to fully comply with the regulations.

Conclusion

Regulation of AI is still in a nascent stage and its future remains uncertain. However, it seems likely that the EU's regulations will take a pivotal role moving forward. Despite the uncertainties surrounding their effectiveness, these regulatory efforts demonstrate a serious commitment from both experts and legislators worldwide to ensure that our future – where AI is expected to hold significant influence – is one of prosperity and not disaster.

References

[1] Department for Science, Innovation & Technology (2023). A pro-innovation approach to AI regulation. Great Britain. (CP 815). ISBN: 978-1-5286-4009-1.

[2] European Commission (2023). The Artificial Intelligence Act. [online]. https://artificialintelligenceact.eu/ Accessed 11th January 2024.

[3] China Law Translate (2022). Provisions on the Management of Algorithmic Recommendations in Internet Information Services. [online]. https://www.chinalawtranslate.com/en/algorithms/ Accessed 11th January 2024.

[4] OECD (2023). Brazil's path to responsible AI. [online]. https://oecd.ai/en/wonk/brazils-path-to-responsible-ai Accessed 13th January 2024.

[5] Innovation, Science and Economic Development Canada (2023). Artificial Intelligence and Data Act. [online]. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act Accessed 13th January 2024.

[6] The White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. [online]. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ Accessed 14th January 2024.

Image Attribution
Clarote & AI4Media / Better Images of AI / User/Chimera / CC-BY 4.0
flags: 20631874 © Jameschipper & 162596679 © Fourleaflover | Dreamstime.com

George Mason • Frazer-Nash Consultancy Ltd

George Mason is an engineer at Frazer-Nash Consultancy, working in the Trustworthy AI team within the Digital Systems Assurance group, where his work focusses on the risk assessment and safety assurance of AI in safety-critical systems. Prior to this, he was a member of the Trustworthy Adaptive and Autonomous Systems and Processes research team at the University of York, where he used formal methods to develop an approach for assuring safety in reinforcement learning. He was awarded a PhD for his research, which won the Best PhD Project award at DCAART'17. His continued research interests lie primarily in the development of trustworthy AI techniques and systems.

Greg Chance • Frazer-Nash Consultancy Ltd

Greg Chance is a consultant in the Digital Assurance Group at Frazer-Nash Consultancy and an Honorary Fellow of the University of Bristol. His interests are AI assurance and methods to develop responsible AI and verification and assurance techniques. He has previously held academic positions researching robotics and AI and has worked on numerous InnovateUK and EPSRC projects about assuring safety of AI systems, including autonomous vehicles, UAVs, soft robots and swarms. He has also developed testing methods for simulation-based verification of AI, including agent-based reinforcement learning and crowdsourcing.

The authors retain copyright of this article. All opinions expressed in this article are entirely the authors' own and do not reflect the views of their current or previous employers.