Poster: Towards Trustworthy AI: Legislation, Techniques and Challenges

George Mason and Greg Chance

Frazer-Nash Consultancy

Extended Abstract *With continuing increases in compute power, data availability and theoretical advances, Artificial Intelligence (AI) has progressed from being primarily an academic discipline to being a household term. Indeed, the use of AI is becoming ubiquitous in society, ranging from benign tools, such as email spam filters, to assisting with advanced tasks, such as financial analysis.*

Recent developments of AI – most notably, deep learning – have transformed it into a highly versatile technology, able to carry out a variety of complex activities; however, this has sparked concerns that its utilisation may have a disruptive effect on society (Gillespie et al., 2023). For example, foundation models, trained to perform a wide range of tasks, have the potential to make entire careers redundant; generative models, capable of creating convincing yet false information, could lead to confusion and conflict; and black box models, whose decisions may have life-changing consequences yet cannot be justified. Furthermore, there is growing apprehension that future forms of AI – more advanced and capable of affecting society in deeper ways than current forms – may unintentionally be trained to learn behaviours that do not align with human preferences (Russell, 2022).

These concerns, and others, bring rise to the concept of 'trustworthy AI'. Although there isn't yet a widely accepted formal or legal definition for what qualifies as trustworthy AI, across the literature (including academic and governmental), recurring qualities that constitute trust in AI include safety, fairness and transparency; and governance to ensure that it is used ethically, accountably and contestably (Chance et al., 2023).

To realise trustworthy AI, jurisdictions globally have begun drafting legislation on the responsible use and development of AI: regulatory frameworks, AI-specific laws, general IT laws that can apply to AI and guidelines have all been proposed (IAPP, 2023). Furthermore, the research community has responded by developing techniques to assess the risks of AI systems and to ensure that such systems incorporate the appropriate qualities that inspire trust (Li et al., 2023).

Of the legislation under development, it is regulatory frameworks that offer the greatest safeguards to ensure trust in AI. These frameworks, currently being developed in the UK (Department for Science, Innovation & Technology, 2023), Canada (Innovation, Science and Economic Development Canada, 2023), Brazil

(OECD, 2023) and the EU (European Commission, 2023), are designed from the ground up to account for a wide range of AI uses and developments – including efforts to future-proof against new forms of AI. China has also developed and enforced several AI regulations; however, whilst these regulations do include various rules for trustworthy AI, they are generally formulated to suit Chinese national interests (China Law Translate, 2022).

Across the regulatory frameworks still under development, the key principles for trustworthy AI outlined in each are broadly similar yet not identical. Common principles include safety and accuracy (AI will behave as it is expected to), transparency (how has AI made its decisions), fairness and equality (AI is free from racial, gender or other forms of discrimination), accountability (who is responsible for the use and development of AI systems) and human oversight (decisions made by AI will align with human values). Themes of other principles that are featured in some regulations include environmental conscientiousness, respect for labour rights and democratic values.

AI-specific laws, such as those proposed in the US, focus on particular instances of using and developing AI; for example, US law mandates that federal government employees who work with AI must undergo training to learn its capabilities and risks (AI Training Act, 2022). General IT laws that can apply to AI, such as New Zealand's Algorithm Charter (New Zealand Government, 2020), could prove helpful in mitigating its risks; however, they may fail to accommodate the inherent characteristics of AI that distinguish it from other classes of technology (importantly, its adaptable nature to potentially learn behaviours beyond those that were assessed). Guidelines, for example, those created in Japan (Ministry of Economy, Trade and Industry, 2022), offer only weak safeguards: adherence to them is not mandatory.

Whilst legislation prescribes what trustworthy AI must or must not do, or how it can or cannot be used, it does not provide the means to achieve it. Therefore, in parallel with legislative efforts, researchers are devising techniques for both assessing risk and assuring trust.

Risk assessment is the process of evaluating the trustworthiness of an AI system; various metrics have been developed to accomplish this. Assessing risk includes the preliminary step of determining which metrics are relevant to stakeholder values; for instance, an autonomous vehicle may require rigorous evaluation of its safety and accuracy, but less so its fairness and equality. Metrics used for risk assessment range from subjective to objective: subjective metrics, such as those used to assess how ethical a system is, include questionnaires and polls – debate is increasingly required (Chance et al., 2023). Oppositely, for functional properties, objective metrics can be used, such as AI-specific standards (The Alan Turing Institute, 2022).

The principles of trustworthy AI require different classes of techniques to be assured. For safety and accuracy, formal methods using mathematical analyses and proofs can be employed to monitor and shape behaviour (Mason et al.,

2017), simulations can be done to evaluate a system's performance (Singh et al., 2021) and testing of the system's code can identify implementation errors (Braiek and Khomh, 2020). For transparency, there is explainability to justify why AI has made a decision (Hassija et al., 2023) and traceability to examine all the factors that created the AI (Mora-Cantallops et al., 2021). For fairness and equality, data should be representative and free from errors (Clemmensen and Kjærsgaard, 2023) and appropriate models should be selected to mitigate bias (Ferrara, 2023). For accountability, there is redress to contest unjust decisions by AI (Fanni et al., 2023) and auditing for assessing the overarching process of developing AI systems (Falco et al., 2021). For human oversight, human-in-the-loop AI architectures require a human to always be included in the decision-making process; similarly, human-on-the-loop architectures ensure that a human can always intervene with an otherwise fully autonomous system (Nahavandi, 2017).

Despite advances in legislation and research, and even though current AI can be trusted in certain contexts, trustworthy AI in a general sense remains a goal; it is not yet the status. Although progress is being made with legislation, most of it is still in the drafting stage; moreover, its effectiveness is yet unproven. Furthermore, whilst many techniques to provide assurances of trustworthy AI have been developed (and continue to be developed), some offer only vague assurances, are too restrictive or have limited applicability (Hassija et al., 2023); and others are vulnerable to attack, such that they could be manipulated into giving assurances for systems that are not actually trustworthy (Noppel et al., 2022). More development is needed for trustworthy AI techniques so they are mature enough to be applied widely and effectively.

An additional challenge, as outlined in the UK's proposed AI regulations (Department for Science, Innovation & Technology, 2023), is that efforts to ensure AI trustworthiness should not stifle innovation. Unreasonably high expectations of trust, excessive legislation and overly restrictive techniques may dissuade people from using or developing AI in systems where it could otherwise flourish.

Acknowledgments We thank all of our colleagues in the Digital Systems Assurance group at Frazer-Nash Consultancy for their valuable expertise and feedback throughout this project. Particular thanks go to Rose Gambon, the group leader, and Lee Ramsay for their support that went beyond technical advice and leadership.

References

AI Training Act (2022). Public Law 117–207, § 136 Stat. 2238. United States.

Braiek, H. B. and Khomh, F. (2020). On testing machine learning programs. Journal of Systems and Software, 164, pp. 1–24. doi: 10.1016/j.jss.2020.110542.

Chance, G., Abeywickrama, D. B., LeClair, B. et al. (2023). Assessing Trustworthiness of Autonomous Systems. arXiv:2305.03411v2 [cs.AI].

China Law Translate (2022). Provisions on the Management of Algorithmic Recommendations in Internet Information Services. https://www.chinalawtranslate.com/en/algorithms/. Accessed 18th October 2023.

Clemmensen, L. H. and Kjærsgaard, R. D. (2023). Data Representativity for Machine Learning and AI Systems. arXiv:2203.04706v2 [stat.ML].

Department for Science, Innovation & Technology (2023). A pro-innovation approach to AI regulation. Great Britain. (CP 815).

European Commission (2023). The Artificial Intelligence Act. [online]. https://artificialintelligenceact.eu/. Accessed 5th October 2023.

Falco, G., Shneiderman, B., Badger, J. et al. (2021). Governing AI safety through independent audits. Nature Machine Intelligence, 3, pp. 566–571. doi: 10.1038/s42256-021-00370-7.

Fanni, R., Steinkogler, V. E., Zampedri, G. et al. (2023). Enhancing human agency through redress in Artificial Intelligence Systems. AI & Society, 38, pp. 537–547. doi: 10.1007/s00146-022-01454-7.

Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. arXiv:2304.07683v1 [cs.CY].

Gillespie, N., Lockey, S., Curtis, C. et al. (2023). Trust in Artificial Intelligence: A Global Study. The University of Queensland and KPMG Australia. doi: 10.1426/00d3c94.

Hassija, V., Chamola, V., Mahapatra, A. et al. (2023). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation, pp. 1–30. doi: 10.1007/s12559-023-10179-8.

IAPP Research and Insights (2023). Global AI Legislation Tracker. [online]. https://iapp.org/resources/article/global-ai-legislation-tracker/. Accessed 4th October 2023.

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 5th October 2023.

Li, B., Qi, P., Liu, B. et al. (2023). Trustworthy AI: From Principles to Practices. ACM Computing Surveys, 55(9), pp. 1–46.

Mason, G., Calinescu, R., Kudenko, D. et al. (2017). Assured Reinforcement Learning with Formally Verified Abstract Policies. In: Proceedings of the 9th International Conference on Agents and Artificial Intelligence. Science and Technology Publications, pp. 105–117. doi: 10.5220/0006156001050117.

Ministry of Economy, Trade and Industry (2022). Governance Guidelines for Implementation of AI Principles. [online]. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf. Accessed 5th October 2023.

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E. et al. (2021). Traceability for Trustworthy AI: A Review of Models and Tools. Big Data and Cognitive Computing, 5(20), pp. 1–14. doi: 10.3390/bdcc5020020.

Nahavandi, S. (2017). Trusted Autonomy Between Humans and Robots: Toward Human-on-the-Loop in Robotics and Autonomous Systems. IEEE Systems, Man, and Cybernetics Magazine, 3(1), pp. 10–17. doi: 10.1109/MSMC.2016.2623867.

New Zealand Government (2020). Algorithm charter for Aotearoa New Zealand. [online]. https://www.data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter/. Accessed 5th October 2023.

Noppel, M., Peter, L. and Wressnegger, C. (2022). Backdooring Explainable Machine Learning. arXiv:2204.09498v1 [cs.CR].

OECD (2023). Brazil's path to responsible AI. https://oecd.ai/en/wonk/brazils-path-to-responsible-ai. Accessed 18th October 2023.

Russell, S. (2022). Artificial Intelligence and the Problem of Control. In: Werthner, H., Prem, E., Lee, E. A. et al. (eds) Perspectives on Digital Humanism. Springer, Cham. pp. 19–24. doi: 10.1007/978-3-030-86144-5_3

Singh, V., Hari, S. K. S., Tsai, T. et al. (2021). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 122–128.

The Alan Turing Institute (2022). AI Standards Hub. [online]. https://aistandardshub.org/. Accessed 5th October 2023.