Standardization and the Governance of Artificial Intelligence

Dave Lewis, Harshvardhan J. Pandit, P.J. Wall, David Filip

The topic of trustworthy Artificial Intelligence (AI) has attracted wide attention from governments, companies and international bodies as they strive to address the ethical and societal risks that emerge as performance of data-driven machine learning scales and improves at pace. Rapid and disruptive advances are evident across a range of applications including: digital content search and selection; business data analysis and decision making; natural language understanding, generation and translation; and speech and video processing. Concerns include the disruption caused by automation of tasks previously requiring human intelligence and communication skills; the ability to produce previously unattainable insights from integrating large data streams monitoring human behavior and the dangers of automated decisions persisting or magnifying socially unwanted biases. These concerns are exacerbated by the opaque nature of modern deep learning systems that renders their internal decision-making unintelligible even to practitioners and poses significant challenges in attempts to provide clear human-understandable explanations of AI decisions. Well publicized episodes have already highlighted existing problematic applications of AI and eroded public trust. These include: the profiled individual targeting of online content and advertising; decision recommendations in the criminal justice system; performance drift in medical diagnosis; and safety critical automobile and airplane management. The ethical and societal implication of these episodes have amplified the call for common international approaches to the development, operation and governance of trustworthy AI systems.

The explosion of interest in trustworthy and ethical AI was heralded by the publication of the Asilomar AI Principles resulting from an assembly of academic and industry actors organised by the Future of Life Institute at the Asilomar conference grounds in California in 2017. At the time of writing in autumn 2020, an online tracker of AI initiatives provided by the Council of Europe counts over 350 different initiatives, dominated by national authorities, international organisations and the private sector.

To date, influential international contributions include:

- The report on Ethically Aligned Design (EAD 2019) from the Global Initiative on Autonomous and Intelligent Systems of the Institute of Electrical and Electronic Engineers (IEEE), which presents a wide ranging international expert review to provide a set of principles for such design, as well as discussing the influence of classical ethics, professional ethics and the role of different different moral worldviews or value systems such as Western Ethics, Buddhism, Ubuntu and Shinto.

---

1 [https://futureoflife.org/ai-principles/](https://futureoflife.org/ai-principles/)
The European Commission’s High Level Expert Group’s ethics guidelines on Trustworthy AI (HLEG 2019), which is informing the EU’s declared drive to become a centre for trustworthy AI development and to legislate on the topic.

The OECD Council on AI’s Recommendations (OECD 2019), which were adopted by the OECD Council of Ministers on 22 May 2019 as the basis for further collaboration on AI policy.

(HLEG 2020) aka ALTAI which develops principles of (HLEG 2019) into a checklist for self-assessment by practitioners.

To date however, most similar proposals offering guidance on trustworthy AI and its ethical dimensions have focussed on defining overriding principles that organisations, governments or societies may apply to the problem or reflective questions to aid deliberation with organisations. Jobin et al. (2019) identify that there is an apparent consensus on the importance of ethical principles of transparency, justice, non-maleficence, responsibility and privacy across 84 such policy document, whereas other common principles of sustainability, dignity and solidarity in relation to labor impact and distribution garner far less attention. Other smaller analyses have also identified gaps in ethical reasoning. One example is the Access Now (2019) programme which analyses and identifies a lack of attention by national authorities to the use of AI by governments and in weapon systems. In addition, Hagendorff (2019) notes a general focus on individual rather than collective harms such as loss of social cohesion and harm to democratic systems. Amongst private sector proposals, Wagner (2018) notes a tendency to diminish the need for government regulation, whilst Calo (2017) warns that ethical principles may present barriers to market entrants. Furthermore, Whittlestone et al. (2019) note the lack of analysis of the tensions between stated ethical principles and commercial imperatives.

Trustworthy AI Standard Development

Given these natural differences, continued work on normalising principles internationally may not serve to advance international consensus in a practical way. Instead, many organisations and governments are shifting focus from the identification of principles to the means for turning those principles into practice. As with the development of trustworthy and ethical AI principles, approaches to related practice are also very diverse, and often reflect the capabilities and positions in power structures of those considering specific approaches. Understanding existing standardised ICT development and organizational management practices offer insight into the extent to which they may provide a basis for standardising practice in governing the development and use of trustworthy AI.

Standards Developing Organizations (SDOs) are themselves very diverse and vary in their attitude to addressing specific ethical issues. The IEEE’s global initiative on ethically aligned design for autonomous and intelligent system has spawned the IEEE 7000 standards working group that places ethical issues at its heart (Adamson et al 2019). The working group is already addressing ethically-aligned standards including ethical design processes, transparency for autonomous systems, algorithmic bias, children, student and employee data governance, AI.

https://ethicsinaction.ieee.org/p7000/
impact on human well-being, and trustworthiness rating for news sources. Data governance, of a series of standards development activities.

A different approach is taken by the ISO/IEC Joint Technical Committee 1 (JTC1) which was established by the International Standardization Organization (ISO) and the International Electro-technical Commission (IEC) in 1987 to develop, maintain and promote standards in the fields of Information Technology (IT) and Information and Communications Technology (ICT). Expert contributions are made via national standards bodies and documents (over 3000 to date) are often used as technical interoperability and process guideline standards in national policies and international treaties, as well as being widely adopted by companies worldwide. Statements of relevance to UN Sustainable Development Goals (SDGs) and Social Responsibility Guidelines are an inherent part of all new standardisation projects proposed in JTC 1. AI standards are addressed together with big data technology standards by the JTC 1 subcommittee (SC) 42 which was formed in 2017 and first met in spring 2018. In common with other JTC 1 standardization activities, this places a strong emphasis on ensuring consistency with existing process and interoperability standards as well as reuse of existing terms and concepts to provide industry with a consistent body of applicable standards. JTC 1/SC 42 is therefore addressing AI related gaps within existing standards for management systems, risk management, governance of IT in organisations, IT systems and software quality, etc.

Characteristics of Trustworthy AI

JTC 1/SC 42 has developed a technical report that sets out some of the core concepts and issues for standardisation related to Trustworthy AI (ISO/IEC 24028:2020). Trustworthiness is defined as the ability to meet stakeholders’ expectations in a verifiable way. When applied to AI, trustworthiness can be attributed to services, products, technology, data and information as well as organizations when considering their governance and management. Characteristics we might expect a trustworthy entity to exhibit could be related to: technical characteristics such as reliability, robustness, verifiability, availability, resilience, quality; bias and robustness; stakeholder-related ones such as ethics, fairness and privacy; as well as management- and governance-related characteristics such as transparency, explainability, accountability and certification.

The remainder of this article considers the prospect for international standards development related to these different characteristics of trustworthy AI.

Technical-related Characteristics

A wide range of technical characteristics relate to trustworthiness, however these can to some extent build on existing standards related to software development, testing, quality and safety.

---

3 https://www.iso.org/committee/6794475.html
4 https://www.iso.org/standard/81230.html
5 https://www.iso.org/iso-31000-risk-management.html
6 https://www.iso.org/standard/62816.html
7 https://www.iso.org/standard/64764.html
8 https://www.iso.org/standard/77608.html
The specific technical characteristics of modern AI such as machine learning, deep learning and neural networks raise novel technical concerns that are being addressed in new standards. These forms of AI are developed through assembling a dataset related to a problem domain that is used to train a model that can exhibit effective and reliable behaviour when presented with similar data in an application. This therefore raises the need to test both datasets and the behaviour of models to identify and rectify unwanted biases where certain groups of people are treated differently to others; e.g. people of colour in judicial decision making or women in processing job applications. New testing standards are also needed to ensure the performance of AI models is robust when the data being processed starts to diverge from that used to train it. Further, the opaque nature of decision-making of many deep learning AI models requires new standards to explain automated decisions in a human understandable manner.

**Stakeholder-related characteristics**

Stakeholder engagement is already well established in the design of human-computer interaction through frequent engagement with the intended users to develop and test though prototypes their interaction with the design. Such user engagement on AI application design will be important, especially in relation to the design of mechanisms for interacting with users via any natural language interface and when providing explanations of AI decisions. For AI however, the focus on ethical issues requires engagement with stakeholders that are more broadly considered than immediate users or customers of an AI system. Appropriately therefore, ISO standards define a stakeholder as any individual, group, or organization that can affect, be affected by, or perceive itself to be affected by a decision or activity, or more broadly as any entity with an interest in an organization’s decisions or activities. The ISO standard for social responsibility places recognition of and engagement with stakeholders at the centre of its guidelines (ISO 26000:2010) and is therefore referenced in SC 42 standards activities related to governance, as well as other ethical and societal issues. Social responsibility guidelines are structured in relation to how stakeholders are classified in relation to established legislative and regulatory structures. These social responsibility stakeholder classes can be characterised as: people in general in relation to human rights issues; workers in relation to labour practices; future generations in relation to environmental issues; value chain partners, i.e. suppliers and customers, in relation to fair operating practices; individual users in relation to consumer issues; and people in locales where organisations operate in relation to community involvement and development. Social responsibility issues therefore provide a good basis for exploring common approaches to issues that have not yet attained widespread consensus as ethical principles, including fairness in distributing the benefits and harms resulting from the use of AI, global considerations of operating in locales with differing levels of protective legislation for stakeholders, labour displacement through automation, as well as the carbon and pollutant emission growth associated with some AI use.

One area where stakeholder considerations are leading to standards relevant to trustworthy AI is in data protection. Existing data protection and privacy laws, such as the European Union’s General Data Protection Regulation (GDPR) and the California Consumer Privacy Act already support common development practice in privacy by design and data protection impact assessments and common operational practices through obligations to data subjects related to
notification of and consent to processing of an individual’s data. The transparency and accountability required regarding information retained by organisations, either directly or indirectly in trained AI models, has the potential to support new approaches to detecting and mitigating AI bias. Work toward machine-readable privacy policies and interoperable data processing consent receipts demonstrate paths to standards in this area. Obligations related to automated profiling and algorithmic decision-making may also motivate standardisation on the capabilities of AI training datasets and AI models, as well as on explainable AI.

Consistent with the emphasis on personal characteristics and behaviour of virtue ethics, the IEEE Ethically Aligned Design and others have suggested that the ethical considerations of AI could be based on a professional oath undertaken by AI practitioners, similar to the Hippocratic oath universally undertaken in the medical profession. Adherence to ethical standards of practice are already expected from members of professional bodies such as the Association of Computing Machinery and the IEEE. However, such professional membership is rarely required of practitioners or even necessarily as an aid to a successful professional ICT career. Therefore, unlike medical and legal professionals, the breach of professional ethics in ICT lacks the sanction of a bar to practice. Further, many governments and companies actively pursue the broadening of access to AI development skills as an economic imperative through low-cost training, which is bolstered by rapid advances in AI development tools to make them accessible to non-ICT or data science professionals. It is unlikely therefore that any form of standardised practitioner oath will have a major impact on trustworthy AI practice.

Governance-related Characteristics
Standards are already available to address the governance of IT by organisations that are extended to address the governance of AI. AI governance would address how the governing body of an organization directs the management of AI in an organization through specification of strategies and policies, how it evaluated proposals and plans from managers to implement these and how it monitors the performance and conformance of that implementation. For organisations using AI, such governance needs to address issues and risk management related to areas such as: automated decision-making; the handling and processing of data; the alignment of AI development and operation to values expressed in policies; compliance to regulatory and other requirements of external authorities; and being transparent and accountable against stakeholder expectations. Apart from data protection, the regulatory and legislative environment for AI is not currently well developed, AI governance standards may need to develop to support different models meeting obligations to external authorities. Possible structures for governing AI could include new national (Tutt 2016) and international (Erdelyi & Goldsmith 2018) regulatory bodies that engage with organisations to co-regulate the technology. Alternatively, self-regulation may be possible by internal ethics boards that may help organizations implement best practice (Calo, 2013; Polonetsky et al. 2015). However, AI governance presents a number of challenges for both organisations and regulators. Scherer (2016) identifies challenges including: reaching stable consensus on what defines AI for governance purposes; the widening access to AI skills and computing infrastructure that serve to obscure AI developments from regulators; the diffusion of AI development over locations and jurisdictions globally; the emergence of impacts of an AI system only after being deployed in an
application; the resistance of modern machine learning to yielding clear explanations of its behaviour; the potential for AI-driven autonomous systems can behave in ways that may elude monitoring or outpace the control of those responsible for them. More broadly, development of regulation is challenged by a variety of factors including: the pacing problems where AI technology develops faster that societies’ ability to regulate it; the international economic and military competition that may impede cooperation needed between major powers in developing common standards; the perceived impediment of regulation to AI innovation and its economic and social benefits; and the power asymmetry caused by AI implementation capability being concentrated in the hands a a small number of large corporations (Calo 2017). To date these factors have not visibly impeded standards development, with JTC 1/SC 42 actively developing AI governance, AI risk management and AI management system (AIMS) standards. Uncertainty about the direction of AI regulation may also direct governance standards to AI specialisations of proposed self regulation standards such as technology ethics review panels (CWA 17145-1 2017) and ethical impact assessments (CWA 17145-2 2017).

References

Wagner, B., Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping” in proc Artificial Intelligence, Ethics and Society, New Orleans, USA, Feb 2018
Whittlestone, J., Nyrop, R., Alexandrova, A., Cave, S., The Role and Limits of Principles in AI Ethics: Towards a focus on tensions, in proc Artificial Intelligence, Ethics and Society, Honolulu, Hawaii, USA, Jan 2019