Global Challenges in the Standardization of Ethics for Trustworthy AI

Dave Lewis*, Linda Hogan, David Filip and P. J. Wall

ADAPT Centre, Trinity College Dublin, Ireland
E-mail: dave.lewis@adaptcentre.ie; linda.hogan@adaptcentre.ie;
david.filip@adaptcentre.ie; pj.wall@adaptcentre.ie
*Corresponding Author

Abstract

In this paper, we examine the challenges of developing international standards for Trustworthy AI that aim both to be global applicable and to address the ethical questions key to building trust at a commercial and societal level. We begin by examining the validity of grounding standards that aim for international reach on human right agreements, and the need to accommodate variations in prioritization and tradeoffs in implementing rights in different societal and cultural settings. We then examine the major recent proposals from the OECD, the EU and the IEEE on ethical governance of Trustworthy AI systems in terms of their scope and use of normative language. From this analysis, we propose a preliminary minimal model for the functional roles relevant to Trustworthy AI as a framing for further standards development in this area. We also identify the different types of interoperability reference points that may exist between these functional roles and remark on the potential role they could play in future standardization. Finally we examine a current AI standardization effort under ISO/IEC JTC1 to consider how future Trustworthy AI standards may be able to build on existing standards in developing ethical guidelines and in particular on the ISO standard on Social

Responsibility. We conclude by proposing some future directions for research and development of Trustworthy AI standards.

1 Introduction

The recent step change in the performance of machine learning and deep neural networks has propelled Artificial Intelligence (AI) systems to the forefront of commercial ICT innovation. In parallel, public and policy concerns have emerged about the trustworthiness of AI systems and more broadly the ethical issues as AI confers new levels of rapid centralization of power in economies and societies. Concerns include the disruption caused as AI automates tasks previously requiring human intelligence and communication skills and AI's ability to produce previously unattainable insights for organizations that can integrate large data streams monitoring human behavior. The concerns are exacerbated by the opaque, sub-symbolic nature of modern deep learning systems that renders their internal decision-making unintelligible and poses significant challenges in attempts to provide clear human-understandable explanations of how results were generated. Well publicized problematic application of AI in criminal justice, education, medical diagnosis and safety critical vehicle management have served to highlight the need for new practices to ensure AI is developed and operated in a trustworthy manner. This raises the question of whether it is both possible and desirable to attempt to standardize these new practices. Reducing ambiguity in how different stakeholder in the AI value chain communicate with each other and with society in general about the trustworthiness of AI system may help improve the understanding and management of the risks involve and thereby reduce a major impediment to realizing the economic and societal benefits of AI. On the other hand standardization related to the trustworthiness of AI must contend with several major headwinds. The rate of AI technology development presents a fast moving target for international consensus building, which could outpace standardization efforts. AI is also increasingly perceived as a key strategic and economic technology for nation states, which may impact their willingness to participate in standardization that is perceived as a brake on rapid AI innovation. AI capabilities are highly concentrated in the hand of a few commercial actors with access to the largest data sets, resulting in an asymmetry of power with respect to other commercial players and even

public sector bodies, which may skew consensus-building towards standards that benefit the widest range of stakeholders.

Nevertheless, many national governments and international bodies have undertaken studies on the societal impact of AI. Many of these identify similar concerns related discrimination and bias reduction; accountability and liability; transparency and explainability; dependability and controllability; and privacy and security. Many also propose broader alignment with established models of human rights and goals of improving human well-being, fairness and sustainability. Proposed mitigations typically break down into: technical and organizational solutions to bias, explanation and transparency; training for and organizational oversight over AI developers; and potential state regulation of aspects of AI deployment.

These approaches are often couched broadly in terms of ethical approaches to the creation and operation of AI systems, with the aim of developing *Trustworthy AI*. In cases where international relevance is sought and where concrete guidelines cannot therefore be grounded in the laws and values of a single jurisdiction, higher level principles are often proposed as a theoretical grounding to the development of guidelines and possible future regulation. Some proposals for Trustworthy AI are grounded in the set of values established in a particular polity. The European Commission's High Level Expert Group (HLEG) guidelines on Trustworthy AI [1] proposes principles grounded in the European Charter of Fundamental Rights. The IEEE work on Ethically Aligned Design (EAD) for Autonomous and Intelligent Systems (A/IS) [2] derives its grounding principles from a wide ranging international expert review, based on a range of culturally diverse ethical traditions. However, these principle-led models approaches can present problems when applied to specific use cases. This manifests itself for example in situations where the harm to the individual must be weighed against benefits to the wider population, or where the autonomy of a minority group needs to be balanced against the economic opportunities of the majority. Stakeholders are likely to struggle to resolve such ethical problems where the conflicts or the relative priority between different grounding principles are raised or where there is doubt on the breadth of acceptance of the grounding principles across the population to which they are applied. For instance, the EU guidelines focus on the EU fundamental rights as grounding principles, but centred on those rights related to individuals without referring to the right of solidarity, which addresses the right to collective action in the workplace. The IEEE EAD also acknowledges that its grounding principles may be skewed towards the Western worldview of ethics that are more dominant in its membership

and accessible to its English-speaking deliberations. Though it acknowledges that AI ethics may be grounded in different moral worldviews or systems of values such as Western Ethics, Buddhism, Ubuntu, Shinto, it does not provide guidance on how to address these differences.

The use of grounding principles presents challenges when aiming to develop common ethical best practices for AI design, development and operation and for the governance of these activities in organizations, across value chains and within societies. Attempting to develop standards that require consensus on values and moral outlook is likely therefore to be curtailed by the variation in value systems encountered across the globe. Accommodating differences in value sets, even within a single jurisdiction, is the domain of legal interpretation where appropriate legislation exists, and of politics where the law is not well established. This raises the question therefore of the degree to which consideration of human values or ethical principles should be included in or even used to drive the formation of international standards. Standards Developing Organizations (SDOs) are very diverse and vary in their attitude to adopting specific value sets. To take an example, ethics and societal concerns are not entirely out of scope of 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, international treaties as well as being widely adopted by companies worldwide. ISO/IEC and other de jure SDOs are formally grounded in the principles of the World Trade Organization (WTO) to reduce barriers to international trade. However, they are expected to take ethical stance where it can be affected by standardization. Some examples within JTC 1 include the work of subcommittee (SC) 39 on ecological sustainability of IT and through IT, and of SC 40 on IT Governance. Standards are therefore a mixed common good that is expected to have positive ethical impact on business and industry practice as well as reducing interoperability barrier to commerce and shared best practice.

As acknowledge more generally with respect to Value Sensitive Design [3], a clear understanding of differences in values may be required in communicating trustworthy characteristics of AI systems internationally. Further, the need for workable international frameworks for communicating ethical issues related to Trustworthy AI are likely to become increasingly important as the

development of AI is rapidly advancing on a more globally multi-polar basis than previous phases of ICT adoption such as personal computing and the Internet. The lowering of barriers to access the technical skills and computing power needed to implement effective AI; the virtual nature of AI-based goods and service that allows them to be accessed easily across jurisdictions; the asymmetry in power between large multinational organizations using AI and the people impacted by their use; and the immense technical and legal expertise that large organizations can bring to bear on any regulatory conflicts, all point to the need for globally usable ethical guidelines that populations and governments can employ effectively regardless of their size.

In this paper, we examine the challenges of developing international standards for Trustworthy AI that aim both to be globally applicable and to address the ethical questions key to building trust at a commercial and societal level. We begin by examining the validity of grounding international standards on human right agreements in order to contribute to the ethical grounding for such standards. We then examine in more detail the major recent proposals from international professional and transnational governmental organizations on the ethical governance of Trustworthy AI systems in terms of scope and use of normative language. From this analysis, we establish a preliminary minimal model of the functional roles that may be relevant to future standards development, both in implementing process standard or as parties communicating across value chains for Trustworthy AI. We identify the different types of interoperability reference points that may exist between these functional roles and remark on the potential part they may play in future standardization. Finally we examine a current AI standardization effort under ISO/IEC JTC1 to consider how future Trustworthy AI standards may be able to build on existing standards in developing ethical guidelines and in particular the ISO standard on Social Responsibility. We conclude by proposing some future directions for research and development of Trustworthy AI standards.

2 Universal Values and Human Rights

Recent years have seen a significant focus on the development of ethical principles for Trustworthy AI amongst intergovernmental and national bodies, as well as amongst professional associations. 2019 has seen both the EU's Ethics Guidelines for Trustworthy AI [1], and the Organization for Economic Cooperation and Development (OECD) Recommendation on Artificial Intelligence (which is the first inter-governmental standard on AI) [5],

as well as comparable initiatives from national governments including Australia, Canadian, and France. The adoption of the Principles for Responsible Stewardship of Trustworthy AI by the G20 (including China and Russia) in the summer of 2019, which includes the OECD recommendations, is also noteworthy in this regard. Professional initiatives such as the 2017 Asilomar Principles on Beneficial AI [4], and the IEEE's Ethically Aligned Design (EAD) are complemented by the adoption of professional research standards in Big Data, robotics and other related fields, and by international efforts such as the UN's AI For Good Series that connect the drive for Trustworthy AI with other relevant matters of global concern, such as the Sustainable Development Goals. These various principles, guidelines and recommendations aim to establish an ethical framework that sets clear parameters for the development and deployment of AI globally, although the model is of soft governance rather than binding international law. As discussed in the next section, there is an apparent convergence on normative principles among the various bodies, regardless of whether the language is that of principle, guideline or recommendation. The language of human rights and claims to universality are fundamental to these international and intergovernmental frameworks, including in the G20 Principles for Responsible Stewardship of Trustworthy AI. In addition to the convergence on the core principles, there is also a common commitment to the development of safeguards that will give technical expression to the ethical principles. They each also adopt an approach that firmly embeds these global initiatives on AI ethics and governance within a democratic and participatory model of politics. This is the case also with the G20 statement, which is drawn from the OECD principles and recommendations. However, Section 2 of the G20 statement, which is concerned with issues of international co-operation, is noted but not adopted by the G20 countries, thus suggesting ambivalence on the part of some members to aspects of the statement on international co-operation for Trustworthy AI.

While this consensus, both on the overall objective of Trustworthy AI and on many of the principles required to advance it is to be welcomed, the likelihood of its achievement is beset by intersecting philosophical, cultural, political, organizational and technical challenges. Central among these challenges is the continuing ambivalence regarding the foundational categories of universal values and human rights, categories that are central to the Trustworthy AI agenda. This ambivalence is in large part a result of the decolonization of ethics which has not only challenged the concept of universality itself but has also engendered a radical critique of many of

the fundamental ethical categories (e.g. autonomy, reason, human rights) on which the current declarations and statements regarding Trustworthy AI rely. The case of human rights exemplifies the point. The 1948 UN Declaration of Human Rights is asserted to be 'a common standard of achievement for all nations and peoples' [6], grounded in universal moral values, and applicable everywhere. Together with the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights it is the international reference point for human rights norms. It forms the basis of the EU Charter of Fundamental Rights, the ASEAN Human Rights Declaration and other regional instruments, and is the backdrop to many of the Trustworthy AI declarations. Yet even as it is the reference point for universal values, its legitimacy continues to be contested. Of relevance for this discussion is not the relativist critique of those like Adamantia Polis who regard human rights as 'a Western construct of limited applicability', [7] but rather critics like the Singaporean delegation at the 1993 World Conference on Human Rights in Vienna who condemn the ethnocentrism of the rights enumerated in the UN Declaration, and who argue for a re-examination of the specific human rights in the Declaration with a view to developing a more globally representative account of human rights and universal values [8]. Although this 'Asian values debate' resurrected the skepticism about the universality of the current codification of human rights, nonetheless Asian values advocates asserted that they accepted the claim that human rights are universal, but that they also believed that such rights ought to be contextualized 'against a dynamic and evolving backdrop of norms, histories, cultures, religions, and national and regional particularities.', Article 8 of the Final Declaration of the Regional Meeting for Asia of the World Conference on Human Rights 1993 [9]. The most cited features of the Asian cultural backdrop that affect the region's instantiation of universal human rights include: a communitarian rather than individualistic world-view; deference to tradition and to elders; strong role-related duties and the prioritization of economic and social over civil and political rights. Moreover, the renewed visibility of Asian moral world-views, particularly Confucianism, in global ethical debate amplifies the challenge to the universal validity of specific moral categories, such as autonomy, and puts in question the exclusive framing of AI governance standards through the lens of the individual and his/her rights.

Asian moral world-views are not the only ones that have acquired a renewed visibility in global moral debate about universal values and human rights. African, indigenous and First Nation ethical traditions, previously

obscured through the ubiquity and persistence of the colonialism and the dominance of Enlightenment morality, have also begun to impact debates about the nature and substance of the values that are claimed to be universal and that are invoked in Trustworthy AI discussions. The IEEE's recommendation that research be undertaken to address the possible relationships between Ubuntu and A/IS value design is but one example of the recognition that African ethical and political categories may have the capacity to re-orient the shape of the trustworthy AI conversation. Ubuntu points towards an ethic that gives precedence to relationality and that contests the current prioritization of the individual subject [2, p. 56]. The recent discussion paper entitled Two-Eyed AI, prepared for the Canadian Commission for UNESCO, and which draws on indigenous values to shape the future direction of AI, is another case in point [10].

Communitarian world-views like Confucianism and previously untheorized categories like Ubuntu challenge many of the assumptions about how universal values and human rights are expressed in specific contexts, and what they imply for the ethical implementation of AI. So too do Hindu, Buddhist, Daoist, Shinto and other worldviews that express the moralities of populations around the world. The use of AI in the Chinese social credit system is a case in point. While critics point to the threats to individual freedom and autonomy, the government emphasizes the important societal and planning benefits that the system brings, and the media report expressed support for the system amongst the Chinese population. Whether this is an example of how a communitarian world-view re-interprets the relative significance of individual and community interests, or of a reluctance to criticize a government policy is unclear. However, it highlights the complexities of developing Trustworthy AI in a global setting, even when there appears to be consensus on the principles.

3 Comparison of Stakeholder Models, Scope and Normative Language

We now consider further how proposed sets of principles for Trustworthy AI could map into new international standards that have potential to contribute to normative behavior of organizations involved in developing and using AI system globally. It is beyond the scope of this paper to provide a comprehensive survey of policy papers as can be found in [13, 27, 28]. Instead we focus on prominent, consensus-led, internationally authoritative works that are relevant to the development of standards in support of Trustworthy AI.

Our analysis moves beyond existing term-based surveys [13, 27] however, and we focus instead on the use of normative language in describing principles. These works analyzed are: the IEEE Ethically Aligned Design (EAD) Edition 1 [2], which is the result of a major international, multidisciplinary, multistage consultation; the EU HLEG Guidelines on Trustworthy AI [1], which was developed by a 50-strong, multi-disciplinary expert group for the EU; and the OECD Council on AI Recommendations, which were adopted by the OECD Council of Ministers on 22 May 2019 [5]. These documents are not directly comparable in their purpose and intent. The IEEE EAD addresses the ethically aligned creation and use of AI, which is a subject of professional practice and interest for many IEEE members. This work specifically analyses Autonomous/Intelligent Systems (A/IS), but we use the term AI system synonymously here. The HLEG document addresses the need for ethical guidance in EU AI policy-making and AI development, deployment and use by organizations in Europe. The OECD document aims to provide governments with common recommendations that aim to foster both innovation and trust in AI, while being sufficiently flexible to weather rapid changes in the field.

These documents differ in their use of normative language. To classify styles of normative language we can refer to the principles and rules for drafting documents used by ISO and JTC1 [11]. The classifications used therein for normative statements are:

- A *requirement*, which is an objectively verifiable criteria which must be met without deviation to claim conformance to the containing standards;
- A *recommendation* which suggest a possible choice or course of action without exclusion of others; and
- A *permission*, which convey consent or liberty to do something.

While the documents analyzed do not claim adherence to this specific style of normative language, it does provide a widely used basis for a comparison of normative intent and scope of individual statement. The quantity of statements from each work that were classified as a requirement ("req"), a recommendation ("rec") or a permission ("perm") were tabulated as summarized in Table 1.

Different authoring styles were adopted between the three documents. Differences in normative language style are also evident within the EAD and HLEG documents, which may be due to their diverse authorship. For instance in the HLEG document, statements that an action "must" be taken, that it was "required" or that it was "critical" that it was taken, were classified

Table 1   Summary of normative statements classification against principle area in OECD [5], HLEG [1] and EAD [2] documents

Categories of Normative Statements	Total Number of Statement	Number of Req. Statements	Number of Rec. Statements	Number of Perm. Statements	Normative Status Unspecified
OECD	19	0	19	0	0
1. Inclusive growth, sustainable development and well-being	1	–	1		–
2. Human–centred values and fairness	4	–	4		–
3. Transparency and explainability	6	–	6		–
4. Robustness, security and safety	7	–	7		–
5. Accountability	1	–	1		–
HLEG Trustworthy AI Requirements	68	21	39	7	1
1. Human agency and oversight	16	2	7	6	1
2. Technical Robustness and Safety	12	6	6	–	–
3. Privacy and Data Governance	6	4	2	–	–
4. Transparency	9	1	8	–	–
5. Diversity, non-discrimination and fairness	6	1	4	1	–
6. Societal and environmental well–being	7	1	6	–	–
7. Accountability	12	6	6	–	–

(*Continued*)

Table 1   Continued

Categories of Normative Statements	Total Number of Statement	Number of Req. Statements	Number of Rec. Statements	Number of Perm. Statements	Normative Status Unspecified
IEEE EAD Principles	43	13	29	0	1
1. Human Rights	7	4	3	–	–
2. Well-being	1	1	–	–	–
3. Data Agency	4	2	2	–	–
4. Effectiveness	9	1	8	–	–
5. Transparency	3	1	1	–	1
6. Accountability	6	1	5	–	–
7. Awareness of Misuse	6	1	5	–	
8. Competence	7	2	5	–	–

as requirements. Statements that an action "should" occur were classified as recommendations. In the EAD document, statements using "must", "shall" and "is needed" were classified as requirements. Recommendations were taken from statements that an action "should" be taken, though in the EAD, statements that an action was "important" or "advisable" were also classified as recommendations. Actions that were stated as "can" or "may" occur were classified as permissions. This classification does not make the distinction made in [11] between a permission (denoted by use of "may") and possibilities or capabilities (denoted by the use of "can"), since it was not apparent that this was a distinction intended by the document authors. Further statements that seem to have some normative intent, but were expressed without use of any of these normative keywords were classified as "unspecified".

In some cases this coding identified normative type statements that made more than one potentially verifiable statement. In such cases the statement was separated into separate statements to enable a clear assessment of the scope of normative statements represented in these documents.

As can be see from Table 1, the OECD document offers relatively few normative statements, all of which are recommendations. This is consistent with the document's aim to provide standard guidance to national governments in support of development of their own policies. Specifically, it provides a basis for the OECD Committee on Digital Economy Policy to monitor

and report on implementation, but this does not involve any certification or enforcement duties. The HLEG document, with 68 normative statements, offers more detailed guidance on achieving Trustworthy AI, with 21 of those stated as requirements. Though the process by which statements were formed as requirements, recommendations or permissions is not explicitly explained, the presence of normative requirement statements does indicate an intent that certification or enforcement may be envisaged within EU policy at some point. The document explicitly positions itself as a complement to a separate lawful AI component. The HLEG's guidance on public policy considerations relevant to legal and regulatory aspects of Trustworthy AI were included in a subsequent document [12], indicating these aspects are being actively explored within the EU policy sphere. The EAD document yielded 43 normative statements, 13 of which were stated as requirements. This document was prepared as part of the broader IEEE initiative that includes on-going work on the IEEE P7000 Standards Series focused on ethically aligned A/IS issues [26]. The selection of the type of normative statement used may therefore reflect the authors' sense of priority for future standards on Trustworthy AI systems under this series.

In terms of scope, all three documents adopt high-level principles addressing accountability, transparency and human wellbeing. Support for human rights is covered explicitly in the EAD document and through reference to human-centered values in the OECD recommendation and to human agency in HLEG guidelines. The HLEG guidelines, and to a lesser extent the OECD recommendation, also address sustainability concerns. All three cover issues of fairness and non-discrimination in AI-supported decision making. As AI is primarily data-driven, issues of privacy and data protection are also touched upon by all three documents, though the EAD seemingly advocates a specific technology-supported approach rather than general principles in this area. The correct technical operation of AI systems is addressed as robustness and safety in the OECD and HLEG documents, while the EAD addresses this as AI system effectiveness. Finally the EAD document places more emphasis than the other two on professional competence (consistent with its role as a professional organization) and the misuse of AI. These observations are similar to a wider recent analysis of 15 papers and reports that propose guidelines for AI Ethics, including earlier versions the HLEG and EAD work [31]. This highlights that the broad focus on privacy, accountability, fairness and transparency (as measured by mentions in the analyzed texts) corresponds to those issues that are amenable to technological rather than institutional or cultural solutions.

3.1 Stakeholder Model

Ethically-driven approaches to the Trustworthy AI must focus on both gaining and maintaining trust of specific stakeholder populations and communicating on trust-building measures between stakeholders. This can be achieved by using definitions that provide a clear context for specific characteristics of Trustworthy AI, such that a change to the context may trigger a critical re-evaluation of the stated characteristic [14]. In this sense, it is insufficient to simply refer to the 'Trustworthy AI', but to specify who trusts whom in what aspects of AI creation, operation and use. Such stakeholder contextualization therefore applies equally to considerations of specific Trustworthy AI principles such as transparency, explainability and accountability. Contextualization requires careful identification of stakeholders and clear understanding of their involvement at different points in the AI system lifecycle and related value chains. Stakeholders can hold different views on appropriate characteristics for Trustworthy AI based on the different values they adhere to or seek to observe. However, if the characterization and decision-making process can be captured in an open and interoperable form, this may reduce barrier to international trade in Trustworthy AI products and services. Given the uncertain and rapidly evolving nature of Trustworthy AI, standardized terms and structures may allow for open exchange of local organizational, national or regional deliberations on Trustworthy AI that enable convergence to best practices while accommodating different approaches to address ethical issues in different cultural or regulatory contexts.

As we consider the development, implementation and enforcement of technical interoperability and process standards for Trustworthy AI, it is clear that different stakeholders may hold differing views of the relative importance of different proposed normative statements. The standardization of terms and a conceptual framework for Trustworthy AI should therefore enable clear, unambiguous communication between different stakeholders, so that these different viewpoints can be understood, made explicit, shared and resolutions sought. Such stakeholder communication may address: how different stakeholders may be affected by AI systems; how any assets that are valued by different stakeholders are used or affected by AI use; and how the use of AI in a product or service relates to values held by different stakeholders. In business, Stakeholder Theory [15] highlights the benefit of an approach to decision making that looks beyond the fiduciary obligation of management to generate profits for shareholders while complying with local legislation, and considers benefits to other types of stakeholders in an organization, including:

employees, customers, management, suppliers, creditors, government and regulators, society in general, and the natural environment (which can be considered as a proxy representing the stake held by future generations).

There is not yet a clear consensus on a typology of stakeholders related to the use of AI in products or services. The EAD document distinguished between AI creators and operators. The OECD document identifies AI actors as anyone playing an active role in the AI system lifecycle, including users and any other actor involved in or affected by an AI system. The HLEG document adopts a definition of stakeholders that encompasses both active AI value chain actors and other affected actors. As all documents address data privacy and data management concerns, the actors providing data for the creation of AI should also be considered explicitly. To varying degrees, the documents also make reference to different forms of oversight authority, though with no explicit commitment of whether this should be government appointed, at either national or transnational levels, an industry-level regulator or an oversight body within an organization. There have been some proposals for possible structures for co-regulation conducted in cooperation between industry and government, such as new national regulatory bodies responsible for the governance of big data and it use for algorithms [16] as well as internal ethics boards that may help organizations implement best practice [17, 18]. There has also been a proposal for the structure of an ethical oversight body proposed as the result of EU research into responsible research and innovation [19]. However, there seem to be few concrete policy proposals in place to establish the public regulatory bodies and standards needed to implement such co-regulation. This is in part due to uncertainty about the direction and impact of AI technology, which promotes a 'wait-and-see' approach [31], but may also in part be due to the lack of relevant technical expertise available to public policy bodies in comparison to the large digital platforms and other major private adopters of AI [17].

To advance discussion of trustworthy and ethical AI standards, we suggest the following stakeholder classes as a potentially sufficient, but broadly acceptable starting point. The following distinct organizational functions can be identified in the AI value chain (noting that a single organization could undertake several such roles):

- *Data Provider*: An organizational function or individual activity that provides the data that is used to create an AI system. For example under the EU General Data Protection Regulation (GDPR) [30] these would correspond to a data controller and data subject roles respectively.

- *AI System Creator*: An organizational function that designs, develops and trains an AI system.
- *AI System Operator*: An organizational function that develops and deploys a product or service that uses one or more AI system.
- *AI User*: An organizational function or individual activity that consumes a product or service that uses one or more AI system.

In addition, we identify two more organizational functions that operate somewhat orthogonally to the above value chain roles in addressing ethical and societal issues. Though defined in somewhat abstract terms, they provide a basis for considering these roles in the application of standards while acknowledging that their constitution, responsibilities and roles are poorly understood, may evolve over time and vary between different markets, jurisdictions, application domains and global cultures:

- *Oversight Authority*: An organizational function recognized by all other actors as exercising authoritative oversight of the operation of the AI value chain and on its affect on any stakeholders.
- *Associate Stakeholder*: Any stakeholder that is affected, or considers itself affected, by the creation, operation or use of a product or service using one or more AI systems, but that is not itself an actor in the AI value chain or exercising oversight over it.

The identification of functions for Oversight Authority (OA), Associate Stakeholder (AS), Data Provider (DP), AI System Creator (AC), AI System Operator (AO) and AI User (AU) therefore enables identification of likely interoperability reference points between these functions that may benefit from standardization. In particular, the reification in this model of Associated Stakeholders, AI Users and Oversight Authorities provides a framing that accommodates greater engagement with societal stakeholders including collective or cooperative action amongst stakeholders. Further, the identification of reflexive interoperability points for the functions enables consideration of interoperating co-existence of variations in operating practices resulting from global variation in value systems as described in Section 2. These potential interoperability reference points between functional units are depicted in Figure 1. They represent a potential Interoperability Reference Model for exploring the focus and scope of standards for Trustworthy AI. Table 2 outlines some potential benefits of standardization at each of these reference points and suggest potential topics for future Trustworthy AI standardization efforts in each case. These potential efforts are discussed further in the following sections.
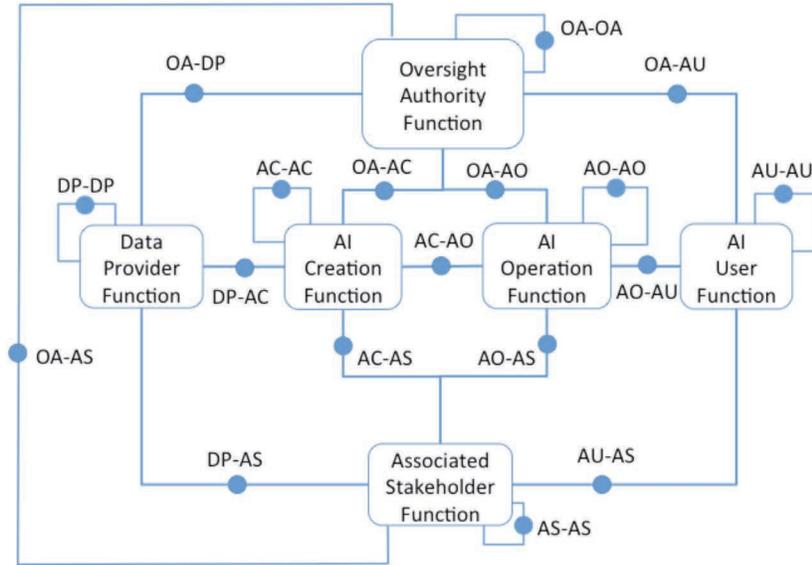
Figure 1   Interoperability reference model for standardizing Trustworthy AI.

Table 2   Potential scope of Trustworthy AI reference point standardization

Ref. Point	Benefits of Standardization	Possible Topics for Standardization
DP-DP	Data sharing value chains, including with data subject, in support of trustworthy AI	Exchange of data and meta-data on data provenance and treatment, terms for data sharing, restrictions on purpose to which data can be put and auditing of usage.
DP-AC.	Risk management and governance of data feeding AI creation.	
AC-AC	Integrative models of AI creation using separately sourced machine learning models, e.g. transfer learning	AI System provision. Ex-ante: Vendor declaration of conformance for AI system. Ex-post: integration, test and incident reports
AC-AO	Risk management and liability governance in adopting AI, e.g. via well specified bias and effectiveness envelopes	
AO-AO	Risk management and liability governance in the integration of products and services using AI	Vendor declaration of conformance for AI-based products and services

Table 2 Continued

Ref. Point	Benefits of Standardization	Possible Topics for Standardization
AO-AU	Transparent, comparable consumer assessment of risks and effectiveness of using AI-based products and services	
AU-AU	Exchange of consumer assessment of AI-based products and services in support of collective action, e.g. class actions, consumer rights	User-friendly ethical impact assessment reporting, sharing and indexing
AS-AS	Exchange of stakeholder impact assessment of AI-based products and services in support of collective action, e.g. digital rights activism, regulatory policy influence	
AU-AS	Broader engagement between users benefiting from AI-based products and services and other impacted stakeholders	
OA-AS	Collection of incident and external impact reports. Stakeholder engagement in the development of policy and regulations	Stakeholder engagement mechanisms. Grievance and redress reporting mechanisms
DP-AS	Assessment of second order stakeholder impacts on data sharing, e.g. on relatives of people providing DNA data	
AC-AS	Engagement of potentially impacted stakeholders with AI creation process.	
AO-AS	Engagement of potentially impacted stakeholders with the process of developing and deploying AI-based products and services.	
OA-DP	Ease data protection reporting for DP operating across jurisdictions, and eases OA interactions with different DPs.	Records of data sharing, data regulation compliance actions and data protection impact assessments

(*Continued*)

Table 2   Continued

Ref. Point	Benefits of Standardization	Possible Topics for Standardization
OA-AC	Eases collection of monitoring and compliance data of AI design and development.	Ex-ante: Data, Ethical and Human Rights impact assessments. Ex-post: AI system testing procedures and results.
OA-AO	Eases reporting of monitoring, compliance and incidence reporting.	
OA-AU	Eases reporting of AI ethical impacts and user tracking of accountability and redress actions.	
OA-OA	Exchange of monitoring and compliance data between authorities at different levels, e.g. in-company, national, transnational, and for cases spanning jurisdictions, e.g. involving multinational DP/AC/AO.	

4 Trustworthy AI and Social Responsibility

To explore the direction of standardization for Trustworthy AI we consider now current activity in ISO/IEC JTC1. JTC1/SC 42 on AI has currently 5 Working Groups (WG). WG3 deals specifically with AI Trustworthiness. WG 3 evolved from a Study Group on AI Trustworthiness that was convened in the Beijing inaugural plenary in Spring 2018. WG 3 started working on its first three technical reports going out of the second SC 42 plenary in Sunnyvale, California, in October 2018. This group works closely with WG1 on Foundational Standards (including concepts and terminology), WG 2 on Big Data, WG 4 which is gathering use cases and applications, and WG5 which is addressing AI computational methods. SC42 also participates in a joint WG with SC 40 (Organizational and IT Governance) exploring new AI Governance requirements.

An AI Trustworthiness overview technical report (ISO/IEC DTR 24028) is currently under ballot since mid 2019. Reports on robustness of neural networks (ISO/IEC WD TR 24029-1) and bias in AI decision-making systems (ISO/IEC WD TR 24027) are underway. Of most relevance is a newly approved work item on ethical and societal concerns (ISO/IEC WD TR 24368). In addition, WG 3 is developing a Risk Management standard for AI (ISO/IEC WD 23894). This is intended as a specialization of the established ISO 31000 risk management standard and SC 42/WG 3 needs to work closely with ISO TC 262 Risk Management. WG 3 is also currently performing a gap

analysis on a number of JTC 1 and other ISO or IEC standards. For instance, development of an AI software testing standard in the ISO/IEC/IEEE 29119 series on software testing is under consideration.

In general, SC 42 is not trying to reinvent wheels, they are looking at the existing standards within the ISO/IEC ecosystem and elsewhere, mapping areas out from the point of view of AI standardization in technical reports and then addressing identified gaps in existing standardization with AI-targeting specializations. Imminent areas under considerations include risk management (as explained above), general governance (ISO TC 309), IT governance (SC 42 JWG 1 with SC 40), systems engineering (SC 7, SQuARE and Software testing) and functional safety of vehicles (ISO 26262 series).

Of relevance to ethical consideration, in 2019 WG3 established a group to document ethical and societal issues of AI. One restriction in this group was that it must avoid adopting any specific value set and should consider principles that could be globally adopted. This made it difficult to adopt existing Trustworthy AI principle such as those laid out in the OECD, HLEG and EAD documents. It has therefore considered basing its analysis on the existing ISO 26000 standard on Social Responsibility [25]. While this is not specific to AI, or even to ICT, it provides a grounding for considering social responsibility issues that has already been accepted by ISO. ISO Guide 82 provides guidelines for considering social responsibility issues in other standardization activities [20]. This provides a set of 37 issues, grouped under 6 core areas of social responsibility, which together with issues of organizational governance can be used as a form of checklist to consider in formulating new standards. These areas are: Human Rights, Labor Practices, the Environment, Fair Operating Procedures, Consumer Issues and Community Involvement and Development. In Table 3 we provide a summary of a mapping of the requirement statements extracted from the OECD, HLEG and EAD documents in section 3 to the Social Responsibility issues under each area. This provides a possible starting point for considering social responsibility for AI Trustworthiness standardization with a global perspective. Statements related to the design and development of AI that did not align obviously with individual social responsibility issues were tabulated under Organizational Governance. Note that some statements were mapped to more than one issue.

This mapping reveals that these existing documents provide relatively comprehensive coverage of issues related to Human Rights and Consumer Issues. This is due largely to normative statements on transparency, account-ability and privacy actions mapping to resolution of grievances under human

Table 3   Mapping of ISO 26000 Social Responsibility issues [25] to areas addressed by OECD [5], HLEG [1] and EAD [2] principles

ISO 26000 Social Responsibility Issues	OECD	HLEG	EAD
Organizational Governance	2	15	6
Human Rights	13	41	32
1. Due diligence	3	4	8
2. Human rights risk situations	2	4	3
3. Avoidance of complicity	–	2	2
4. Resolving grievances	5	19	9
5. Discrimination and vulnerable groups	1	6	2
6. Civil and political rights	2	4	5
7. Economic, social and cultural rights	–	2	3
8. Fundamental principles and rights to work	–	–	–
Labor practices	2	4	4
1. Employment and employment relationships	–	1	–
2. Conditions of work and social protection	–	–	–
3. Social dialogue	–	1	3
4. Health and safety at work	2	1	1
5. Human development and training in the workplace	–	1	–
The Environment	1	6	0
1. Prevention of pollutions	–	–	–
2. Sustainable resource use	1	2	–
3. Climate change mitigation and adaptation	–	2	–
4. Protection of the environment, biodiversity and restoration of natural habitats	–	2	–
Fair Operating Procedures	0	4	0
1. Anti-corruption	–	–	–
2. Responsible political involvement	–	2	–
3. Fair competition	–	–	–
5. Promoting social responsibility in the value chain	–	1	–
6. Respect for property rights	–	1	–

(*Continued*)

Table 3 Continued

ISO 26000 Social Responsibility Issues	OECD	HLEG	EAD
Consumer Issues	3	30	21
1. Fair marketing, factual and unbiased information and fair contractual practices	2	4	–
2. Protecting consumers' health and safety	–	3	1
3. Sustainable consumption	–	–	–
4. Consumer service, support, and complaint and dispute resolution	–	18	12
5. Consumer data protection and privacy	1	4	8
6. Access to essential services	–	1	–
7. Education and awareness	–	–	–
Community Involvement and Development	4	5	12
1. Community involvement	4	3	10
2. Education and culture	–	–	–
3. Employment creation and skills development	–	–	–
4. Technology development and access	–	1	2
5. Wealth and income creation	–	–	–
6. Health	–	1	–
7. Social investment	–	–	–

rights and to consumer service, support, complaint and dispute resolution and data protection and privacy under consumer issues. Sustainability is covered in the existing documents, but not extensively, which is perhaps consistent with the relatively simple relationship perceived between computing and the environment. Of more concern is relative lack of coverage for: labor issues, especially given the considerable concerns about labor market disruption arising from AI-driven automation; and fair operating procedures, given the oft-commented dominance of a small number of multinationals in developing AI-based products and services [17]. The relatively low coverage of issues around community involvement and development also reinforces the need to address the engagement with stakeholders as identified in the interoperability model presented in section 3.1, including consideration of culturally grounded differences in ethical outlook as discussed in section 2. Similar conclusions were raised in a recent study of European proposals for AI regulation at the EU and national levels [21]. This study identified gaps

in addressing impacts of AI: on freedoms of expression and of the press; on vulnerable communities subject to decision by, but not directly using AI; on due process in law; on economics right and access to work; on the use of AI by public authorities and on collective harm in favor of individual rights.

5 Conclusions and Further Work

The standardization of Trustworthy AI is in its infancy. Intense interest in the subject in industry and at national and international policy level has resulted in proposed statements of principles or requirements for ethical and trustworthy AI by the OECD, the EU and the IEEE that demonstrate some level of consensus in the topics that require attention, including accountability, transparency, fairness and non-discrimination, privacy and data protection, human well-being and human rights and sustainable development. All three works vary in the depth of focus and the strength of normative expectations they apply to these broad areas. However, the analysis of their normative content informed a proposed stakeholder model that identifies organizational and societal functions over which we establish an Interoperability Reference Framework as the basis for exploring future standardization for Trustworthy AI.

ISO/IEC JTC1/SC42 is already well placed to develop standards that extend existing standards in Risk Management, IT Governance and System Testing that can be readily deployed in the identified Data Producer, AI Creator and AI Operator functions. This provide a common set of procedures, concepts and terms that can be use to develop standardized vendor certificates of conformance at the DP-AC, AC-AO and AO-AU reference points that make up the AI value chain. Such standardization effort can build on existing proposals for supporting trustworthiness in the exchange of AI training data sets [22, 23].

Developing standards that address concerns outside of the AI value chain, i.e. concerning the Oversight Authority and Associated Stakeholder functions as modeled in section 3.1, represents a more significant challenge. This is in part due to the slow and varying response of national governments to AI regulation, especially as international competition for global dominance in AI grows. It is also in part due to the variation in the ethical viewpoints and the resulting differences in societal priorities and trade-offs globally as discussed in section 2. International standards development requires international consensus, and cannot directly resolve such global variances. However, the approach taken in ISO 26000, where consideration

of community involvement and development is made an explicit part of an organization's approach to social responsibility, may offer a useful, standards-based approach to managing variation in ethical outlook, especially for multinationals as they operate globally. Common practice is being developed for ethical impact assessment [24] and citizen engagement[1] in responsible research and innovation, that could be extended to modes of stakeholder engagement for use by AI value chain actors. These developments offer rich sources of emerging shared best practice for local community engagement with value chain players that could be used as a basis for standardization that address community involvement and development aspects of socially responsibility for AI. Further, standardized reporting on the conduct and outcome of such community engagement may ultimately provide a medium for oversight by relevant local authorities, i.e. via the OA-AS, OA-DP, OA-AC, OA-AO and OA-AU reference points. The process oriented auditing and accountability employed via these reference point could build heavily on standards terms, measures and processes developed in SC42 under its AI risk management, governance, bias and system engineering projects. In 2019, SC42 also commenced a justification study for the development of a Management Systems Standard for AI that may lead to requirements for process and product certification that could be exposed and assessed by a certifying body in the oversight authority role, via these reference points.

At the same time, new forms of civil society, such as data cooperatives[2] and data trusts,[3] are emerging to deal with the growing datafication of societal issues. These may themselves benefit from and engage in development of intra-community information sharing, i.e. the AS-AS and AU-AU reference points in our model, to help strengthen the effectiveness of these local civil society bodies and counter balancing the centralization of power that AI technology seems to be accelerating us towards.

Acknowlegdements

---

[1]https://www.rri-tools.eu/

[2]https://datacommons.coop/

[3]https://theodi.org/article/what-is-a-data-trust/

References

[1] European Commission's High Level Expert Group, "Ethics Guidelines for Trustworthy AI", April 2019. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

[2] "Ethically Aligned Design – First Edition", The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, IEEE, 2019.

[3] Borning, A. and Muller, B., "Next steps for Value Sensitive Design", Proceedings of CHI 2012, 1125–1134. New York, NY: ACM Press.

[4] "Asilomar AI Principles", Future of Life Institute, 2017. Retrieved from https://futureoflife.org/ai-principles/

[5] OECD/LEGAL/0449. (2019). "Recommendation of the Council on Artificial Intelligence". OECD.

[6] UN General Assembly, Resolution 217A, 1948 https://www.un.org/en/universal-declaration-human-rights/index.html

[7] Pollis, A., "Human Rights: A Western Construct with Limited Applicability", In Human Rights: Cultural and Ideological Perspectives, edited by Pollis, A. and Schwab, P, 1–18, New York: Praeger, 1979.

[8] Dunne, T. and Wheeler, N. (eds.) "Human Rights in Global Politics", New York: Cambridge University Press, 1999.

[9] "Final Declaration of the Regional Meeting for Asia of the World Conference on Human Rights", Bangkok, March 29 to April 2, 1993, https://www.ru.nl/publish/pages/688605/bangkok-eng.pdf

[10] Bourgeois-Doyle, D. "Two-Eyed AI: A Reflection on Artificial Intelligence – A Reflection Paper prepared for the Canadian Commission for UNESCO", Ottowa, Canada, March 2019.

[11] ISO/IEC Directives Part 2: "Principles and rules for drafting and structuring of ISO and IEC documents", 8th edition, 2018. https://www.iso.org/sites/directives/current/part2

[12] "Policy and Investment Recommendation for Trustworthy AI", High Level Expert Group on Artificial Intelligence, 26 June 2019.

[13] Hagendorff, T., "The Ethics of AI Ethics – An evaluation Guideline", https://arxiv.org/pdf/1903.03425.pdf

[14] O'Keefe, K. and Brien, D.O. "Ethical Data and Information Management: Concepts, Tools and Methods", Kogan Page, 2018.

[15] Freeman, R.E. and McVea, J. A Stakeholder Approach to Strategic Management (2001). Darden Business School Working Paper No. 01–02. Available at SSRN: https://ssrn.com/abstract=263511 or http://dx.doi.org/10.2139/ssrn.263511

[16] Tutt, A. (2016). "An FDA for Algorithms". SSRN Electronic Journal, 83–123. https://doi.org/10.2139/ssrn.2747994

[17] Calo, R. (2017). "Artificial Intelligence Policy: A Roadmap". SSRN Electronic Journal, 1–28. https://doi.org/10.2139/ssrn.3015350

[18] Polonetsky, J., Tene, O., and Jerome, J. (2015). "Beyond the common rule: Ethical structures for data research in non-academic settings". Colo Tech L J, 13, 333–368.

[19] "Ethics assessment for research and innovation – Part 1: Ethics committee", CEN Workshop Agreement CEN/CWA 17145-1:2017 http://satoriproject.eu/media/CWA_part_1.pdf

[20] ISO Guide 82:2014, "Guidelines addressing sustainability in standards", ISO, April 2014, https://iso26000.info/wp-content/uploads/2016/04/ISO_Guide_82_2014E_new_format.pdf

[21] "Mapping regulatory proposals for artificial intelligence in Europe", Access Now, Nov 2018. Retrieved from https://www.accessnow.org/mapping-regulatory-proposals-AI-in-EU

[22] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé, H., and Crawford, K. "Data sheets for Datasets, in Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning", Stockholm, Sweden, 2018.

[23] Bender, E., and Freidman, B., "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science", in Transactions of the Association for Computational Linguistics, vol. 6, pp. 587–604, 2018

[24] "Ethics assessment for research and innovation – Part 2: Ethical impact assessment framework", CEN Workshop Agreement CWA 17145-2, http://www.tekno.dk/wp-content/uploads/2017/08/CWA17145-22017.pdf

[25] "Guidance on social responsibility", International Standards, ISO 26000:2010(E), 2010.

[26] Adamson, G., Havens, J.C., and Chatila, R. "Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems", Proceedings of the IEEE, 107(3), 518–525. 2019, https://doi.org/10.1109/JPROC.2018.2884923

[27] Zeng, Y., Lu, E., and Huangfu, C. "Linking artificial intelligence principles" arXiv preprint arXiv:1812.04814, 2018

[28] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI", January 15, 2020, Berkman

Klein Center Research Publication No. 2020-1. Available at SSRN: https://ssrn.com/abstract=3518482

[29] Calo, R. "Consumer Subject Review Boards: A Thought Experiment", Stanford Law Review Online, 66 (2013), Retrieved from http://www.stanfordlawreview.org/online/privacy-and-big-data/consumer-subject-review-boards

[30] European Union. (2016). Regulation 2016/679 of the European parliament and the Council of the European Union (General Data Protection Regulation). Official Journal of the European Communities, 2014(October 1995), 1–88. https://doi.org/http://eur-lex.europa.eu/pri/en/oj/dat/2003/l_285/l_28520031101en00330037.pdf

[31] Daly, A., Hagendorff, T., Hui, L., Mann, M., Marda, V., Wagner, B., Wang, W., and Witteborn, S., "Artificial Intelligence Governance and Ethics: Global Perspectives", 8 June 2019, https://arxiv.org/abs/1907.03848

Biographies

Dave Lewis is the head of the AI Discipline at the School of Computer Science and Statistics at Trinity College Dublin. He is also Deputy Director of the ADAPT SFI Research Centre for Digital Content Technology. He leads ADAPT's programme of industry collaborative research and its multidisciplinary research theme on Data Governance. His research focuses on the use of open semantic models to manage the Data Protection and Data Ethics issues associated with digital content processing. He has led the development of international standards in AI-based linguistic processing of digital content at the W3C and OASIS and is currently active in international standardization of Trustworthy AI at ISO/IEC JTC1/SC42 serving as an expert contributor to documents on an Overview of Trustworthy AI and Ethical and Societal Issues for AI.

Linda Hogan is an ethicist with extensive experience in research and teaching in pluralist and multi-religious contexts. Her primary research interests lie in the fields of inter-cultural and inter-religious ethics, social and political ethics, human rights and gender. Professor Hogan has lectured on a range of topics in ethics and religion, including Ethics in International Affairs; Ethics of Globalisation; Biomedical Ethics; Human Rights in Theory and Practice; and Comparative Social Ethics. She has been a member of the Irish Council for Bioethics and has been a Board member of the Coombe Hospital, Science Gallery and Marino Institute of Education. She has worked on a consultancy basis for a number of national and international organisations, focusing on developing ethical infrastructures.

David Filip is a research fellow at Trinity College Dublin and is part of the ADAPT Centre. His research address interoperability in digital content technologies by developing and implementing open and transparent technical standards. He focuses on optimizing processes, making science and technology serve business needs, including ethical and societal concerns. He is: convenor of JTC 1/SG 1 Open Source Software; convenor of JTC 1/SC 42/WG 3 Trustworthiness of AI, national mirror chair for National

Standards Authority of Ireland (NSAI) TC 02/SC 18 on AI; Head of the Irish national delegation, ISO/IEC JTC 1/SC 42 AI; Chair & Editor of OASIS XLIFF OMOS TC; Secretary & Lead Editor, OASIS XLIFF TC; and NSAI expert to ISO/IEC JTC 1/SC 38 Cloud Computing, ISO TC 37/SC 3 Terminology management, SC 4 Language resources, SC 5 Language technology.

P. J. Wall is a postdoctoral researcher at the ADAPT Centre (www.adapt centre.ie) in the School of Computer Science, Trinity College Dublin. His research focuses on technological innovation and an exploration of the wider implications associated with the social, cultural, and political aspects of the implementation and use of ICT in the Global South (ICT4D). His current research is based in Sierra Leone and examines the role of mobile technologies in reconfiguring health systems and practices (mHealth), and an exploration of how such mobile devices are implemented, adopted, scaled, and sustained. His interest is in understanding ICT, and specifically mHealth, adoption, and use from a variety of critical realist and interpretivist ontological perspectives.