Governance for Digital Humanism: The Role of Regulation, Standardization, and Certification

Clara Neppel and Patricia Shaw

Abstract Assuring that digital systems and services operate in accordance with agreed norms and principles is essential to foster trust and facilitate their adoption. Ethical assurance requires a global ecosystem, where organizations not only commit to upholding human values, dignity, and well-being but are also able to demonstrate this when required by the specific context in which they operate. We focus on possible governance frameworks including regulatory and non-regulatory measures, taking as an example AI systems. Thereby, we highlight the importance of considering the specific context, as well as the entire life cycle, from design to deployment, including data governance. Socio-technical, value-based standards, and certification schemes are introduced as enabling instruments for operationalizing responsible and ethical approaches to AI in line with upcoming regulatory requirements.

1 Introduction

AI systems can be used to positively impact humanity for good, *provided it* is designed, developed, deployed, and decommissioned responsibly. This requires creators of AI and users of AI to go beyond the legal requirements (where they exist) and take a whole ecosystem approach to ethically manage the risks and impact AI can have on fundamental rights, human dignity, and human flourishing and sustainability, in short, on people and the planet.

Operationalizing responsible and ethical approaches to AI requires both a top-down and a bottom-up (inclusive of stakeholders) approach to AI and data governance, without which no organization can effectively (1) map (namely, identify AI legal, societal, economic, environmental, and technological risks and plot them to

C. Neppel
IEEE, Vienna, Austria
e-mail: c.neppel@ieee.org

P. Shaw (✉)
IEEE, Beyond Reach Consulting Services Limited, Worksop, UK
e-mail: trish@beyondreach.uk.com

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_21

the relevant product/service and personnel responsible for those risks), (2) manage, (3) measure, (4) mitigate, or (5) monitor their AI or hold themselves accountable for the outputs and outcomes in the short, medium, and long term.

We live in a global AI market, for which there is a clear need for a global coordinated response, but with direct relevance to local contexts when it comes to AI. Regulatory requirements have (as at the date of writing) been jurisdictionally bound, leaving swathes of the world simply having to respond voluntarily rather than dutifully following mandatory legal requirements. For any global response to be effective, it will require the following ecosystem conditions: standards, certification, trustmarks, audit, and, most importantly, stakeholder engagement to not only provide assurance of responsible innovation but to help define the all-important guardrails for safe and trustworthy AI for a global digital world with unique and contextually bound application domains.

2 Background to AI Principles, Regulation, and Standards

2.1 The Principles

There are a number of principles and frameworks seeking to identify and/or provide a taxonomy for AI ethics and values that are to be applied to AI systems and that potentially could be applied universally. These principles were developed by a large number of entities, including international organizations and other governments, industry, and professional organizations, e.g., UNESCO, OECD, and IEEE.

A mapping exercise was undertaken by the Berkman Klein Center at Harvard University, which published "A Map of Ethical and Rights-based Approach to Principles for AI"[1] (see Fig. 1).

In its mapping exercise, the Center found that there was a great degree of commonality in the approaches that many principles, guidelines, and frameworks called for. Key themes included:

- International human rights
- Promotion of human values (such as autonomy, agency, dignity, empathy, and well-being)
- Professional responsibility
- Human control of technology
- Bias, fairness, and non-discrimination
- Transparency and explainability
- Safety and security

[1] Fjeld, Jessica and Achten, Nele and Hilligoss, Hannah and Nagy, Adam and Srikumar, Madhulika, 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,: https://ssrn.com/abstract=3518482 or https://doi.org/10.2139/ssrn.3518482

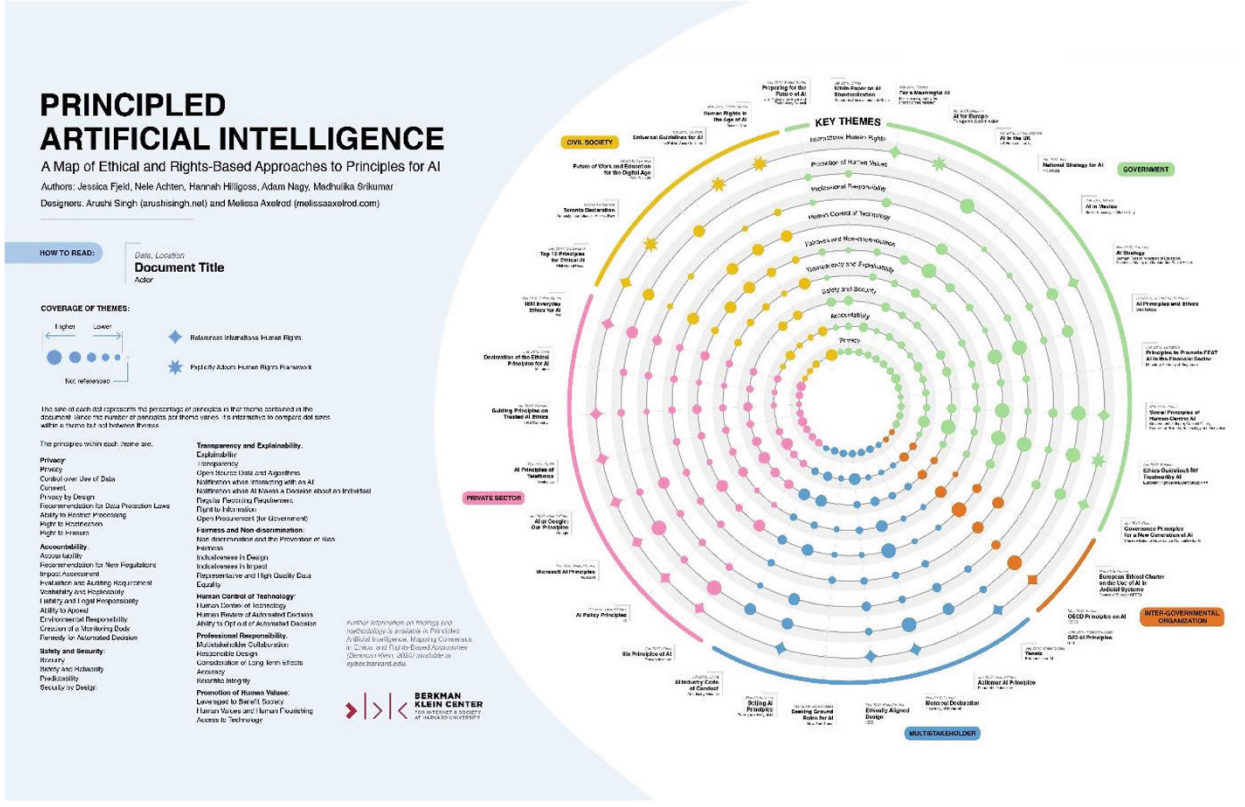

Fig. 1 Source and copyright: Berkman Klein Center (Source: https://cyber.harvard.edu/publication/2020/principled-ai licensed under a Creative Commons attribution 3.0 Unported license: https://creativecommons.org/licenses/by/3.0/legalcode)

- Accountability
- Privacy

The key challenges for many of those principles that were mapped from across the globe are that they are voluntary and therefore are not enforceable and lack the clarity of law and so are often not fully defined leaving their operationalization open to interpretation.

2.2 The Role of Regulation

First proposals for regulating the use of AI systems have already been tabled in different parts of the world to address the specific challenges of AI systems and to provide a trustworthy ecosystem for all affected stakeholders. These regulatory proposals aim to provide AI developers, deployers, and users with requirements and obligations regarding specific uses of AI.

The EU AI Act sets out a risk-based approach, where the obligations for a system are proportionate to the level of risk that it poses. The Act outlines four levels of risk: low-risk systems, limited or minimal risk systems, high-risk systems, and systems with unacceptable risk. We see risk focus and risk proportionality increasingly being used by governments and regulators when designing and delivering regulations with the aim to improve their effectiveness and efficiency.

In parallel, the Council of Europe started a negotiation process for a legal instrument on the development, design, and application of AI based on the Council of Europe framework for human rights, democracy, and the rule of law. If adopted by several countries across the world, this instrument has the potential to act as an international treaty on artificial intelligence.

Besides addressing concerns through legislation and regulations, what can be called "hard law," non-regulatory means known as "soft law" can also set substantive expectations but are not directly enforceable by governments. The OECD AI principles (OECD AI Global Principles Overview, n.d.) are an important example of soft law. They represent one form of such programs where high-level norms are created by a multilateral organization with the intention of setting baseline expectations for the management of AI.

In summary, both hard law and soft law seek to define high-level requirements and obligations for the application of AI systems.

2.3 The Standards

While we do not currently (at the time of writing) have an overarching international legal treaty or convention on AI, and national law and regulation is still in the making, standards are potentially our only way to provide for a consistent technical and/or socio-technical approach to design, develop, and deploy AI systems in a trustworthy and sustainable manner.

As set out above, principles and regulatory requirements are at a high level of abstraction and often need further interpretation for a given context or industry. For instance, transparency can have different meanings to different actors in different sectors. An accident investigator and the average user of an autonomous system would surely have different expectations. The investigator would need to access technical details, such as the source code, whereas the user would need explanations about the system's actions or recommendations, in the name of transparency. This illustrates why having a common understanding of broad and shared principles is key to establishing trust in an ecosystem.

Open and consensus-based processes are the best means for agreeing not only on the definition of principles and requirements but also on how these principles would be implemented and validated. Standards are what can help turn principles into practice and help make AI (and more pertinently AI assurance) interoperable between businesses (or governments) and borders. Standards provide definitions for the principles and a way forward in how to interpret them and apply them in the AI life cycle. Standards can be technical (placing on their users technical requirements) and/or socio-technical (placing on their users processes and/or methodologies in the design, development, and use of technical requirements to achieve human-centered societal outcomes).

2.4 The Role of Standards and Certification

Standards can provide for a technical or non-technical specification, recommend practices, prescribe processes, or describe detailed requirements that must or should be fulfilled to either achieve particular outcomes or for the purposes of compliance and conformity. Examples of standards used every day include IEEE 802.11 WLAN standard and the ISO 27001 information security and management systems standard.

The necessary level of trust in socio-technical systems can only be achieved if affected stakeholders openly address the expected benefits and risks for the given context, as well as necessary tradeoffs associated with them. Stakeholders should include technologists, human scientists, regulators, and civil society. Several initiatives echo this mindset, including OECD, Council of Europe, or IEEE.

Traditionally, standardization deals with technical issues, such as quality, interoperability, safety, or security. In order to help organizations apply abstract AI principles to concrete practices, the IEEE Standards Association has been developing socio-technical standards in parallel to technical standards. Socio-technical standard working groups convene technologists with stakeholder groups and focus on things like defining different levels of transparency for incremental needs or impact assessment of AI systems on human well-being and the environment.

One example of such a standard is the IEEE 7000™-2021 Model Process for Addressing Ethical Concerns During System Design (IEEE 7000™-2021 Standard - Addressing Ethical Concerns During Systems Design, 2023). The standard guides developers in making their products and services compatible with the ethical values

of the communities in which technical products and services are placed and used. The standard gives step-by-step guidance to organizations on how to care for stakeholder values from the early conception of a system all through its development and later deployment. To elicit values of ethical relevance, the standard applies utilitarianism, virtue ethics, and duty ethics and recommends to also reach out to the culturally and spiritually founded ethical traditions of local cultures.

IEEE 7000 has four primary processes to build ethical systems: concept of operations and context exploration, value elicitation and prioritization, ethical value requirements identification, and risk-based design. These are complemented by a transparency management process.

The role of standards and certification (and in particular AI ethics standards like that seen in the IEEE 7000 suite of standards) is about creating the right behaviors across the AI life cycle and creating the right environment and ethics culture for businesses to interoperate across the AI value chain.

While AI ethics standards set the bar of what processes need to be in place to help achieve certain ethical outcomes, certification is about providing assurance that the necessary processes, policies, practices, and procedures are put in place between parties so that they can fulfil their own legal compliance requirements; manage risk; understand their dependencies, interdependencies, and limitations; and appropriately mitigate and monitor risks.

In conclusion, standards are about how you do it and the good (and often best) practice an organization puts in place, but certification is about testifying publicly to what has been done by the organization to get it AI ethics ready.

2.5 What Is AI (and Data) Governance and Why Is It Necessary?

Artificial intelligence (AI), or more pertinently an AI System, according to the OECD is as follows: "*AI system*: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy."

This definition is set out in OECD/LEGAL/0449 AI Recommendation, which was adopted on May 22, 2019. At the time of writing this chapter, while it was not documented in the Official Draft of the EU AI Regulation, it was recognized that this definition had also been accepted by the European Parliament as the official definition of AI for the purposes of the EU AI Regulation.

As an AI system is neither created nor operated in a vacuum, certain other definitions also accompanied the definition of an AI system under the OECD Recommendation. These include recognition of the AI life cycle and the AI value chain where a variety of actors and stakeholders play a part.

"*AI system lifecycle*: AI system lifecycle phases involve: *i)* 'design, data, and models'; which is a context-dependent sequence encompassing planning and design,

data collection and processing, as well as model building; *ii)* 'verification and validation'; *iii)* 'deployment'; and *iv)* 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.

"*AI knowledge*: AI knowledge refers to the skills and resources, such as data, code, algorithms, models, research, know-how, training programs, governance, processes and best practices, required to understand and participate in the AI system lifecycle."

"*AI actors*: AI actors are those who play an active role in the AI system life cycle, including organizations and individuals that deploy or operate AI."

"*Stakeholders*: Stakeholders encompass all organizations and individuals involved in, or affected by, AI systems, directly or indirectly. AI actors are a subset of stakeholders."

AI governance must therefore recognize the complex ecosystem within which AI is designed, developed, deployed, monitored, and overseen, as well as decommissioned.

When we talk of governance of AI, firstly we cannot leave data out of the equation. For a technology that is data-driven, where, how, and when you got your data and for what purpose matter.

To that end, AI governance must include data governance as two but intertwined ecosystems. Indeed, the European Commission proposed together with its AI strategy also a data strategy to establish the right regulatory framework regarding data governance, access, and reuse. The provenance and quality of data matters. Data (especially if it is personal identifiable data) is potentially also subject to separate regulatory regimes in different jurisdictions. If not completely separate regulations, the interpretation of them can be unique to localized contexts and regulators. Data governance requires assessment and evaluation of the data used in data-driven technologies at every stage of the data life cycle, which is a separate ecosystem in and of itself to that of the AI life cycle but forms an intricate part of the AI life cycle.

The data life cycle (like the AI life cycle) has various stages where the type of data and treatment of the data must be observed, analyzed, and in some cases modified (whether for accuracy or for format, for structuring or for profiling within wider database, or for being matched or merged with other data sets), actions logged, and decisions recorded. The data life cycle typically consists of (1) collection, (2) collation, (3) storage, (4) decisions and inferences made, (5) reporting the story, (6) distributing and sharing, and (7) disposal[2].

How data is treated or what decisions are made will affect the AI system (Fig. 2).

Data can be used at different touchpoints across the whole of the AI life cycle. Depending on how the data is used and when in the AI life cycle will determine its impact. Data is used for training the AI system; testing and evaluating the AI system

[2]Holt, Alison, Data Governance – governing data for sustainable business (BCS, The Chartered Institute for IT 2021, Swindon, UK)

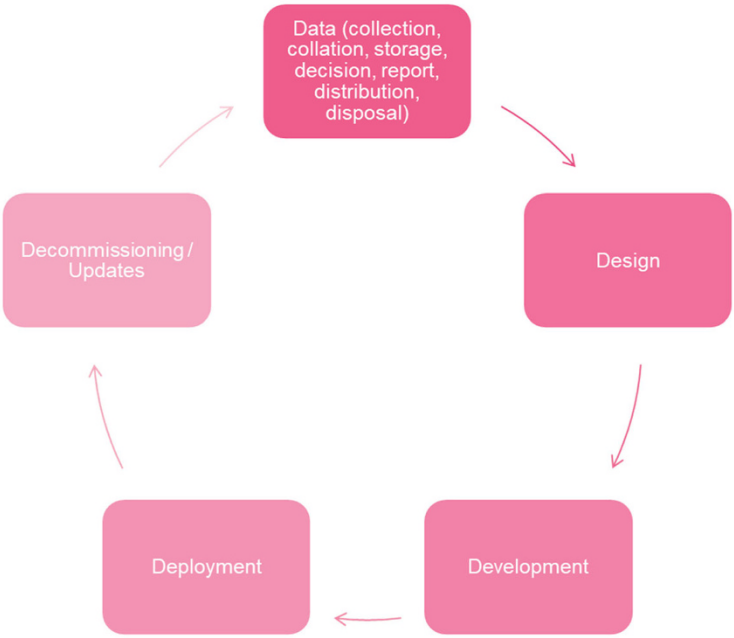

Fig. 2 Data life cycle (Reproduced with the permission of the copyright holder Beyond Reach Consulting Limited)

prior to going to market or being put into service, for verification, or when it is fully operational may set parameters and determine inferences and links made between data variables, features, and attributes.

Governance is the requirement to hold the providers of an AI system to account and to have designated roles aligned with responsibilities to hold the AI system (and the organization designing, developing, deploying, operating, maintaining, and decommissioning the AI system) to account. Fundamentally, it is to have oversight of an AI system to manage it, map the risks, mitigate the risks, and monitor them and (should it be necessary) to have the mandate to turn it off (with that all important "kill switch"), reset it, update it, and provide alternative operations for business continuity and disaster recovery.

An AI system, unlike static software applications, is dynamic. Machine learning, and in particular, deep learning, has the potential to make constant small but iterative changes to the AI system, such that it is perceived as "self-learning." The outcomes of such an AI system (hereafter AI Outcomes) can vary depending on their application domain, context, and audience. AI outcomes can result in societal, ethical, environmental, economic, technological, and legal risks and impacts that may change over time or only become apparent after a significant period of use. Some AI outcomes may transpire in the short term, but others may only occur over the medium or longer term. It is because of this agile and dynamic nature of AI that any AI governance framework applied to it itself cannot be a "one stop shop," never to be

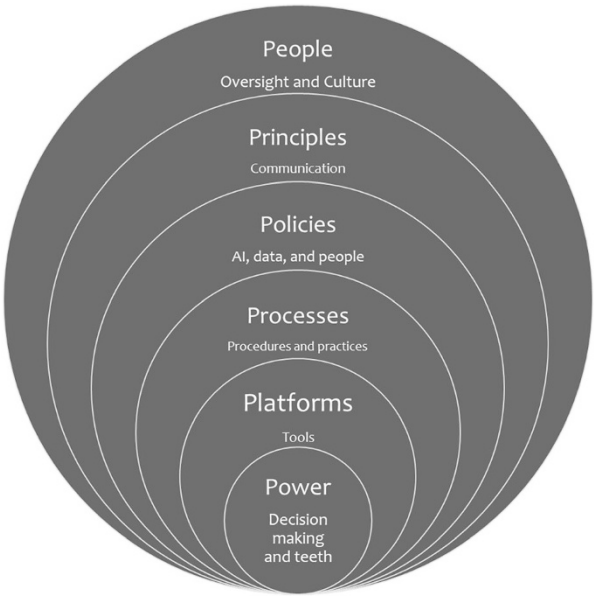

Fig. 3 The six *P*s of AI governance (Reproduced with the permission of the copyright holder Beyond Reach Consulting Limited)

revisited again. Nor can it take a "one-size-fits-all approach." AI governance must be iterative (like the AI life cycle) and continuous (beyond an AI system being put into action in a live environment): map . . . . manage . . . measure . . . mitigate . . . monitor the risks and . . . repeat.

To devise an agile and iterative AI governance framework, it needs to be a holistic approach, which requires an organization to have a four *M*s approach, (1) multilayered, (2) multidisciplinary, (3) multifaceted, (4) multijurisdictional and/or multicultural, and to have the six *P*s in place: (1) people; (2) principles; (3) policies; (4) processes, practices, and procedures; (5) platforms; and (6) power (Fig. 3).

Ultimately, an AI Governance Operating Model should encompass both the 4*M*s and the 6*P*s. Ideally these would all be mapped in a centralized organization-wide Global Risk and Compliance (GRC) Register referencing a centralized repository of all AI use in an organization aligned to domain, product and platform, as well as the data repository containing details of data provenance and the data's limitations (whether they be contractual or purpose limitations), and reporting would be to an empowered, with four *I*s (independence, influence, insightful, and informed), ethics advisory board engaged iteratively just as the AI governance is managed, mitigated, and monitored iteratively. Herein lies the key to successful AI governance, and that is where the ethics advisory board provides the all-important oversight over and above the day-to-day operational management and governance. In an ideal world, independent oversight of AI systems, which are high risk and have the potential to have a negative impact or unintended consequences on people and planet, such as large foundational models, ought to be mandatory.

Having governance structures in place to deal with the day-to-day operations and management of an AI system is one thing, but having an independent board other than that of the executive or non-executive organizational board (depending on the organization's structure) to help oversee and provide an element of that all-important stakeholder insight (as experts and experienced individuals for a variety of disciplines and backgrounds, the ethics advisory board itself can add to the stakeholder voices) will help hold the organization internally to account for itself.

2.6 *Key Areas for Any Responsible AI Governance Operating Model*

Operationalizing responsible and ethical approaches to AI requires both a top-down and a bottom-up (inclusive of stakeholders) approach to AI and data governance, without which no organization can effectively map, manage, measure, mitigate, and monitor their AI or hold themselves accountable for the outputs and outcomes in the short, medium, and long term.

Furthermore, operationalizing responsible and ethical approaches to AI requires a holistic and values-based approach to governance, requiring an understanding of what it means to put ethical principles and their foundational requirements in practice to an organization. This requires mapping the risks (legal, reputational, ethical, and societal) and the benefits both to the business and all its ecosystem stakeholders. This is the approach of the IEEE CertifAIEd framework. The main idea is that the riskier from an ethical perspective an AI system of interest is, the deeper into the levels of the framework the duty holder needs to interrogate.

The IEEE has published its core CertifAIEd ontological specifications[3] detailing the first-tier level of enquiry and provides businesses and governments, any duty holder from within the AI system of interest, with a great starting point to look holistically at the organization as well as the technology and its outcomes. It's intended to be a holistic and outcomes-based approach to AI ethics. Furthermore, it is also intentionally able to be adaptable and flexible to meet the needs of the local application domain and its context.

The CertifAIEd framework promotes awareness, intelligence, and ethics and provides a firm foundation for any AI governance operating model based on four key areas:

[3] https://engagestandards.ieee.org/ieeecertifaied.html

- Accountability[4]
- Algorithmic Bias[5]
- Transparency[6]
- Ethical Privacy[7]

More criteria suites under CertifAIEd are to follow.

As highlighted above, many principles and frameworks exist that do not provide a clear definition or an interpretation to allow them to be operationalized with any level of consistency. The CertifAIEd framework and criteria suites provide both the definitions and credible ways to evidence that "ethical foundational requirements" (operations that provide for and promote ethical practices and behaviors) have been met.

2.7 Accountability

According to the IEEE's CertifAIEd ontological specification, to put in place accountability over an AI system means:

> **ethical accountability:** A contextual set of values pertaining to accountability and the satisfaction of a framework of expectations concerned with taking responsibility for actions, omissions, and outcomes and their ethical consequences (such as justice, redress, preservation of autonomy, self-determination, self-selected communities/locum and intimacies, and where issues of dignity and well-being in the use of technology are pertinent).

The framework further specifies how such ethical accountability is to be interpreted:

- Ethical accountability needs to be human-centric: when humans who are part of the accountability construct whether that be governance and oversight roles and responsibilities or it be part of an ethics advisory board, committee, panel, etc., the duty holder draws from a wide variety of dimensions being diverse and

[4] https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEE_CertifAIEd_Ontological_Spec-Accountability-2022.pdf#:~:text=Abstract%3A%20The%20IEEE%20CertifAIEd%E2%84%A2%20criteria%20for%20certification%20in,ethical%20performance%20is%20the%20goal%20of%20this%20work.

[5] https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEE%20CertifAIEd%20Ontological%20Spec-Algorithmic%20Bias-2022%20%5BI1.3%5D.pdf#:~:text=Abstract%3A%20The%20IEEE%20CertifAIEd%E2%84%A2%20criteria%20for%20certification%20in,ethical%20performance%20is%20the%20goal%20of%20this%20work.

[6] https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEE%20CertifAIEd%20Ontological%20Spec-Transparency-2022.pdf#:~:text=Abstract%3A%20The%20IEEE%20CertifAIEd%E2%84%A2%20criteria%20for%20certification%20in,ethical%20performance%20is%20the%20goal%20of%20this%20work.

[7] https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEESTD-2022%20CertifAIEd%20Privacy.pdf

inclusive to ensure that accountability is kept "human centric," i.e., humans at the heart of it and humans in the loop of AI governance and cognizant of real human impact based on the variety of human experiences and expertise.

- Ethical accountability is of a multidimensional nature. What and who is accountable and responsible for an action or omission in an organization depends on the structure of the organization, the roles held within the organization, the clarity of reporting lines, and how well supervised or not staff (or contractors) are within an organization. Furthermore, each role may interpret what is going on in an AI system differently depending on their own expertise and experience, and the interaction between colleagues in any governance construct may also be susceptible for group and power dynamics—both positively and negatively.
- Attitudes, behaviors, culture, and institutionalized norms and practices have a role to play in accountability. Poor behaviors, culture, and perceived normalized practices in an organization can lead to a vicious circle. In contrast, good behaviors, a culture that takes responsibility and seeks to do better and be ethical, and an environment of seeking excellence and best practice can lead to a virtuous circle. The presumption here is that poor and unethical practices ultimately lead to bad outcomes.
- Upholding law is seen as complementary to accountability as failure to comply with law tends to result in enforcement of better practices and/or liability. Depending on whether law exists to hold organizations to account, or whether it goes far enough, will determine how much it would truly overlap with ethical accountability. That said, law tends to be promulgated in response to unethical behaviors and practices that are deemed unacceptable by a civilized society. While law in the area of AI is awaited, frameworks like CertifAIEd concerning accountability will be crucial in demonstrating the trustworthiness of organizations in their design, development, and deployment (as well as decommissioning) of AI systems.

2.8 Algorithmic Bias

According to the IEEE's CertifAIEd ontological specification, the distinction between algorithmic bias in the context of an AI system and ethical algorithmic bias is:

> **Algorithmic bias**: Automated recommendations and predictions that disproportionately favor one stakeholder entity over another. This may be a negative unethical bias that prevents fair access to education, employment, health care, and economic enfranchisement. It may be a positive ethical bias that weights the AIS and its data use to recommend and predict fair outcomes for identified stakeholders within the context of use for the AIS.
>
> **Ethical algorithmic bias**: A contextual set of values pertaining to a framework of expectations that ensures algorithmic biases that negatively impact individuals, communities, and society have established boundaries of acceptance to protect autonomy and freedoms, where autonomy is defined by one's capacity to direct one's life.

This framing of algorithmic bias and ethical algorithmic bias recognizes that some bias is wanted and desirable and some bias is unwanted and chiefly negative in its results. Important to note that algorithmic bias does contribute to unfair outcomes but is not the sole measure of unfairness. To elaborate on how ethical algorithmic bias can be interpreted in the context of a CertifAIED certification:

- Bias can be introduced and reintroduced at any point during the AI life cycle. To that end, it is important to implement interventions to counterbalance and counteract negative bias, to preserve personhood and individual autonomy.
- It concerns bias that affects humans, so recognizing that bias is a chiefly human endeavor, whether it is in the institutional, systemic, and historic data or again institutional, systemic, historic, cognitive, cultural (and the list goes on) rearing its unwanted head in relation to the designing, the development, the deployment, or even decommissioning of a system, bias is there. It is borne of people, about people, and impacting people.
- Ethical algorithmic bias ought to be complementary to areas of law, which are enforced concerning protection from discrimination and from having barriers to all important freedom. Like we have seen above, what the algorithmic bias may be preferencing or skewed in relation to may not always neatly fall within a protected characteristic, e.g., socioeconomic deprivation.
- Bias cannot realistically be eradicated, and sometimes having intentional and wanted bias is desirable.
- Removing protected characteristics and/or bias considerations may in some instances result in a "blind policy" approach being adopted in respect of an AI system, which itself may cause further bias problems and other undesirable outcomes from the AI system, including inadvertently or uncharacteristically identifying false positives or false negatives. More on the biased impacts of false positives and false negatives can be seen in Joy Buolamwini's papers concerning "Gender Shades" (Gender Shades, kein Datum) and the Netflix film "Coded Bias" (Coded Bias, 2020).

2.9 Transparency

According to the IEEE's CertifAIEd ontological specification, to put in place ethical transparency such that it is clear what an AI System does and how it does it, means:

> Ethical transparency: A contextual set of values pertaining to transparency and the satisfaction of a framework of expectations (preservation of autonomy, self-determination, and self-selected communities/locum and intimacies).

It recognizes that transparency is contextual and local context to the person endeavoring to provide as well as receive transparency and that context is pertinent to the understanding of the AI systems. Ethical transparency can be further interpreted in the context of a CertifAIEd certification as:

- Human centric: it must be transparent to humans and contextually relevant for humans.
- Norms and practices that can either work toward transparency or cause obfuscation and detract from transparency.
- Informational autonomy and empowerment to make informed decisions.
- Without transparency, law cannot easily be enforced, and law cannot be applied. The same applies in respect of ethical foundational requirements for CertifAIEd assessment. Law cannot be truly determined and applied without transparency. Furthermore, without transparency, accountability, privacy, and algorithmic bias protections cannot be easily applied. In short, transparency is the cornerstone ethical requirement to most other ethical and legal requirements. It's foundational.

2.10 Ethical Privacy

According to the IEEE's CertifAIEd ontological specification, to safeguard privacy in an AI system means to go above and beyond mere what is legally required but to consider all facets of the private sphere of a person including their data and to understand them contextually to that person. Ethical privacy is therefore:

> A contextual set of values pertaining to privacy and the satisfaction of a framework of expectations (preservation of autonomy, self-determination, and self-selected communities/locum and intimacies).

Context matters and privacy are no less contextual in respect of (1) what is being disclosed or hidden and (2) the context of what the item(s) are that are being disclosed or hidden and indeed (3) where the privacy is being exerted such as in the home or in one's home life. For example, people consider information about their sexual health or orientation, religion or belief or political associations, and biometrics as sensitive personal data. In contrast, while not always sensitive, financial information is often deemed highly confidential and socially may be taboo to talk about. Furthermore, information about who a person is friends or associates with or which sports clubs they belong to may be seen as private but less sensitive depending on their context and what is intended to be done with the information.

Ethical privacy is not just about personal data being protected under legislation like EU GDPR; it is about going beyond the law, exploring the rights and freedoms of individuals and the collective. It is keeping privacy human centered rather than merely data centered.

To elaborate more on what ethical privacy means and how it can be interpreted in the context of a CertifAIEd certification:

- Ethical privacy is highly contextual and is affected by a variety of dimensions, including but not limited to geographical, cultural, and matters pertinent to ethnicity. An example of the latter might be concerning the Maori people and their ethical data principles. They understand personal data being an extension of themselves and their personhood, requiring special privacy and treatment. Personal data for people of Maori ethnicity operates in an especially sacred space.
- What is considered worthy of privacy (or a right or wrong behavior in relation to a person's privacy) can in some jurisdictions be dictated by local laws but can also be determined by localized social, cultural, and moral norms, ethics, and principles.
- Ethics is human focused, so ethical privacy is human centric.
- Ethical privacy does overlap and complements data protection, privacy, and human rights laws, but ethical privacy takes considerations beyond what the law requires, often the law being very data centric or confidentiality centric (recognizing in some common law jurisdictions that privacy entails torts of peeping tom, publication of private facts, defamation, and misappropriation) as opposed to considering wider aspects of interference with personhood or unverified or intrusive inference about personhood.
- It pertains to all aspects of privacy, including physical, emotional, spiritual, psychological, thought-life, economic, and cultural, and within the inner sphere whether in the life analogue or the life online, beyond simple informational privacy and data and data protection concerns.
- Privacy is not always a matter of upholding individual identity or dignity but can pertain to a group or community beyond that of the individual person.
- It recognizes the power in privacy and its correlation with self-determination and autonomy.
- Ethical privacy is something that aligns with a person's personal expectations but also pertains to the integrity of self, the group, or the community.
- Failure to uphold ethical privacy can lead to human dignity being undermined and a greater dependency or reliance on the use of technology, which may determine inclusive or exclusive behaviors.

3 Application of the Principles and the Governance Operational Model

Standards providing details of process, practices, and procedures, coupled with the IEEE CertifAIEd frameworks, can provide a great deal of practical guidance and reference tool on how accountability, algorithmic bias, transparency, and privacy (amongst other tenets of governance) can be mapped, managed, mitigated, and monitored both from the top-down and the bottom-up.

Putting ethical principles into practice realistically needs a "champion" at the very top of an organization (usually C-suite level) who would drive the organizations to

put principles into practice and to be ultimately accountable for governance and the outcomes AI produces. For any governance framework to be effective, it will require financial resourcing and capacity, capability, and competence and a number of other roles and responsibilities across the organization (preferably dedicated personnel and teams) to also be responsible for the AI being managed and monitored on a day-to-day basis. It will also require participation and understanding of the impacts on stakeholders, especially those who are to be impacted by or influenced by the AI system(s) subject to the governance.

For citizens, it means demanding that AI-based public services be fair and transparent. To keep up, public bodies will have to adopt. Companies providing AI-based solutions and establishing internal criteria and measures that cannot be independently verified will not be able to provide a genuine guarantee that the expected criteria are satisfied.

In the fast-changing AI environment, it is important to be innovative, and standards development organizations are no exception. Currently, it can take years to finalize a standard so that it is ready to certify the conformity of products or services. Sometimes AI development and deployment require only a few months; to wait years is unacceptable. Therefore, the development of standards and conformity assessment criteria needs to become more agile so that it can adapt to changes faster.

For this to happen, AI systems developers need new ways to collaborate and achieve consensus faster. Currently, IEEE's CertifAIEd program uses a model-based graphical capture and representation approach for the principal concepts and factors that foster or inhibit the attainment of the desired aim, such as transparency. This allows rapid tailoring to the needs of a sector, such as finance, or a specific use case, such as fraud detection.

4 Use Case: Wiener Stadtwerke (The IEEE CertifAIEd Framework for AI Ethics Applied to the City of Vienna, 2021)

IEEE CertifAIEd's first real-world test was completed in a pilot project between IEEE Standards Association (IEEE SA) and Wiener Stadtwerke. Wiener Stadtwerke is a public service provider owned by the City of Vienna, providing services in the areas of public transport, electricity, natural gas, heating, telecommunications, parking, burial, and cemeteries, to more than two million customers in the Vienna metropolitan region.

In recent years, the Wiener Stadtwerke group has explored several ideas for using AI technology in pilot projects, always adhering to the overall goal of efficiently delivering high-quality services to the citizens of Vienna. One of these was selected for thorough ethical evaluation in the IEEE CertifAIEd pilot with IEEE SA. This is an email classification system (ECS), which is used to automatically assign categories to incoming customer service requests.

The customer service department of Wien Energie (an energy provider belonging to the Wiener Stadtwerke group) receives more than 1000 email requests per day, which need to be briefly skimmed over by a person and assigned to one of about 15 categories. This categorization results in tickets assigned to different teams for processing, where every email is read by a human operator, who will then determine and initiate the appropriate actions and send a reply to the customer. The manual pre-categorization procedure amounts to one person's entire work time per day, even when less than 30 seconds are spent per email. And it is a very repetitive, monotonous, and tiring task. The ECS was developed to automate this pre-categorization step, effectively relieving one customer service operator to focus on actual customer interaction again and thus making better use of their qualifications and training.

Because the described manual procedure has been applied for years, an excellent data collection of several hundred thousand emails with manual category assignments by experts was readily available, providing a very promising starting position for a machine learning approach to the problem. Therefore, a group-internal project was initiated in 2019 to explore the possibility to develop an automatic categorization system from scratch, which gradually led via increasingly mature prototypes to a production-ready email classification system.

5 Assessment of Wiener Stadtwerke's Email Classification System

The first step in the evaluation process was to thoroughly explain the system and its context to a panel of five IEEE experts, including the background and goals of the project, the system's architecture and interfaces, the machine learning component, and the data used for model training, as well as the effects of the new system on people and processes in the organization.

Based on this information, a risk assessment according to the IEEE CertifAIEd framework was conducted. For each of 26 ethical values such as transparency, dignity, trust, and (avoidance of) discrimination, the expert panel rated the likelihood of the ECS to undermine that ethical value, considering concrete potential scenarios in the system's deployment in the Wiener Stadtwerke context. The results of this risk analysis were used to determine the most relevant of the four IEEE CertifAIEd criteria sets for the application—accountability in the case of the ECS. Furthermore, the overall low-risk class of the system resulting from the risk assessment meant that only a subset of the accountability criteria set needed to be addressed in the following step.

Next IEEE SA provided a list of 43 ethical criteria with brief definitions to Wiener Stadtwerke, who were then to provide evidence for each criterion, showing that the respective ethical question or issue is adequately addressed in the system and its context. These criteria range from rather technical aspects such as error analysis, hyperparameter tuning, and mitigation of false positives to more governance-related

aspects concerning the organization, such as adopting a layered approach; avoidance of inaction, delay, and indifference; and human authority and autonomy.

For each of the 43 criteria, Wiener Stadtwerke provided evidence in the form of technical documentation, system architecture and software implementation details, screenshots, meeting slides and meeting minutes, internal and public reports, strategy papers, process and role definitions, organigrams, etc., giving full detail for the respective criterion. A so-called Case for Ethics document was compiled, using a structure and template provided by IEEE SA, where Wiener Stadtwerke provided general information about the system, its background, scope, etc. (similar to step one, but in written and structured form), as well as all the evidence for the 43 accountability criteria. This Case for Ethics, a 150-page document, was then submitted to IEEE for assessment.

Finally, an assessment report was delivered back to Wiener Stadtwerke by IEEE SA. This included specific feedback for each of the 43 criteria from the expert panel members, indicating to what degree the respective criterion was considered fulfilled and what could be done to further improve in the respective area. It also included an overall confirmation that the submitted Case for Ethics justifies recognition and certification through the IEEE CertifAIEd program for Wiener Stadtwerke's email classification system. The expert panel feedback contained also pointers to things that could be further improved.

6 Conclusions

Digital humanism should result in the development and use of trustworthy and sustainable digital solutions. This brings a range of responsibilities that technical communities of developers and engineers alone do not have the need to adopt in isolation. Enablers with a combination of organizational, cultural, and technical skills have the ability to come up with technically based value propositions that align with the ethics and values of their application domain stakeholders. Thus, the governance and risk management structures within organizations will be ultimately responsible for implementing standards, best practices, and audits, as well as training programs and certification for the people who develop and use high-risk systems.

As such, technical and socio-technical standards, and certifications, developed in an open and transparent paradigm, can establish evidence of the extent to which systems and ecosystem stakeholders conform with upcoming regulation or agreed principles. Such standards and certifications would serve as reliable and important governance instruments for regulators, industry, and the ordinary citizen.

In the current dynamic context, effective and efficient standardization, certification, and appropriate governance structures are indispensable elements of a trustworthy ecosystem. We have shown that these elements complement and facilitate the development of responsible regulatory frameworks that guarantee both the uptake of AI systems and address the risks associated with certain uses of this new technology, such as currently assessed by the Council of Europe or the European Commission.

An ethical future—this is a journey, not a one-stop shop. Not only for the businesses designing, developing, deploying, and monitoring AI but also those who procure it and use it as well as those that become the future ethical and responsible AI practitioners.

As AI is borderless, an ethical future also requires interoperability—clearly recognized global standards to provide for consistency and certainty while adapting and being flexible enough to local ethics and values and being contextually relevant.

Finally, there is also a need to train for the jobs of the future, which will likely be multidisciplinary and require interdisciplinarity. Skills will need to cover not only the creation of technologies but also the governance, oversight, as well as the development of policies, laws, principles, standards, certification, conformity assessment, and audit. Future jobs may include value leads, AI ethics certifiers, and auditors. This needs AI ethics literacy, ongoing education, and identification of the skill sets necessary for future competent assessors and trainers in these areas. IEEE (among other bodies) can provide sector and technology-related professional education to skill the future generations. At a given point, this should become a part of mainstream education. In the meantime, raising awareness of AI outcomes and potential risks for people and planet, increasing technical understanding accompanied with the ability to critique the outcomes (both legal and ethical, short term, medium term, and long term) is vital.

Discussion Questions for Students and Their Teachers

1. Mapping AI ethics risks—assessment of risk, impact, scope, and likelihood or severity of an AI system (Table 1)
2. Consequence scanning—an agile practice for responsible innovators (https://doteveryone.org.uk/project/consequence-scanning/).

   Using this tool considers the scope of the ethical risks in short, medium, and long term to a wide variety of potential actors and stakeholders

Table 1 Mapping AI ethics risk matrix (Reproduced with the permission of the copyright holder Beyond Reach Consulting Limited)

Risk	**What is the impact / outcome of the risk?** *(The risk could have multiple impacts (or could be an outcome from an impact) and impact stakeholders differently or have different effects in different application domains and contexts)*	**Scope of impact** *(How many people/how much could it impact)*	**Likelihood** *(How likely is the risk to occur)*	**Severity** *(If the risk were to occur, how severe would that that impact be)*

Table 2 Plotting responsibility (hose responsible, accountable, consulted, or informed (RACI)) to ethical foundation requirements matrix (Reproduced with the permission of the copyright holder Beyond Reach Consulting Limited)

Description of Effect	Stakeholder affected	Describe the requirements necessary to manage /mitigate /monitor that effect	Describe where in the AI lifecycle could those requirements be best managed/mitigated/monitored	Consider who is best placed to manage /monitor and mitigate them (RACI)

Table 3 Interventions matrix (Reproduced with the permission of the copyright holder Beyond Reach Consulting Limited)

Risk / Lifecycle Stage	Ideation	Data	Design	Development	Deployment	Ongoing use over time	Decommissioning	Third Party

3. Plotting requirements and responsibility
   To help keep accountability at the forefront of AI governance, assign and align every AI governance requirement to manage, mitigate, and manage an AI ethics risk to a responsible person(s) (Table 2)
4. List interventions and strategies to help your organization to manage, mitigate, and monitor risks at each stage of the AI System life cycle (Table 3)

Learning Resources for Students

The following reading material is intended to deepen the knowledge on different instruments that can be used to develop a responsible AI governance framework within organizations. These instruments should cover the different stages of the AI life cycle, from design to deployment, and include context-specific guidelines, standards, and/or certification frameworks.

1. Value-Based Engineering: A Guide to Building Ethical Technology for Humanity (De Gruyter Textbook) | Spiekermann, Sarah | ISBN: 9783110793369
2. iTechlaw's Responsible AI Impact Assessment (RAIIA) tool which can be downloaded from here: https://www.itechlaw.org/ResponsibleAI

3. IEEE ontological frameworks
   Ethical Accountability: https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEE_CertifAIEd_Ontological_Spec-Accountability-2022.pdf?mkt_tok=MjExLUZZTC05NTUAAAGETQHvhqRyJpxehbsTfVHQ3D88oTpizkK-2u0p4IDJF3zbJ2AphqtpsegAVyn4nDEKjPk0H2KzBB2xsikYm4E6Ty1rRyAEumWnb2dvifyEeQ
   Ethical Algorithmic Bias: https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEE%20CertifAIEd%20Ontological%20Spec-Algorithmic%20Bias-2022%20%5BI1.3%5D.pdf?mkt_tok=MjExLUZZTC05NTUAAAGETQHvhWI31Wh8NNeK8rkpq3xDImplIZIV2E_hi3EUhWHL0RzJiSjqTZ_ueYqb0rJ-SKu4_kYgMAWygZyF80qPdxUb_ybwLQIAKOaUGV2JeA
   Ethical Transparency: https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEE%20CertifAIEd%20Ontological%20Spec-Transparency-2022.pdf?mkt_tok=MjExLUZZTC05NTUAAAGETQHvhk2i97UsPFNbzH3-oUDVx_Qk4KdQUdyon6YHLAzDYUx54JOVCY_Oxr2-CwxIAZN7tiaq36aSCV-rKj8pEOG5EPG91AjUBuQBemt5uA
   Ethical Privacy: https://engagestandards.ieee.org/rs/211-FYL-955/images/IEEESTD-2022%20CertifAIEd%20Privacy.pdf?mkt_tok=MjExLUZZTC05NTUAAAGETQHvhbqJ92qFvqon29PnOA4jYmt9VhwjD6oz0WT2NzwiyjUGtBsO8Q5P3TjdT4NwuDIX5E-yRgoUOAadgENoa8mdUn9Fenk3Zb0JV4m-BQ
4. AI Watch: Artificial Intelligence Standardization Landscape Update https://publications.jrc.ec.europa.eu/repository/handle/JRC131155
5. OECD AI Policy Observatory: https://oecd.ai/en/
6. AlgorithmWatch AI Guidelines Global observatory https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/
7. Corporate Digital Responsibility—an international manifesto for businesses: https://corporatedigitalresponsibility.net/cdr-manifesto
8. UK's Digital Catapult AI Ethics Framework https://migarage.digicatapult.org.uk/ethics/ethics-framework/
9. Robotics and AI Laws conf, Standardization and AI 30th May 2022 https://ai-laws.org/2022/08/09/conference-report-4th-rails-conference/?lang=en
10. Data Governance—governing data for sustainable business, Alison Holt, published by BCS, The Chartered Institute for IT https://shop.bcs.org/store/221/detail/WorkGroupByIsbn/9781780173757
11. The AI Book by Fintech Circle, Chapter 7 Trust, Transparency and Ethics - Good Governance of AI by Patricia Shaw https://fintechcircle.com/ai-book/

References

Coded Bias. (2020). *Coded Bias.* Accessed May 15, 2023, from https://www.codedbias.com

Gender Shades. (n.d.). *mit media lab.* Accessed May 15, 2023, from https://www.media.mit.edu/projects/gender-shades/publications/

IEEE 7000™-2021 Standard - Addressing Ethical Concerns During Systems Design. (2023). *IEEE SA - Standards Association.* Accessed May 15, 2023, from https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html

OECD AI Global Principles Overview. (n.d.) *OECD.AI - Policy Observatory.* Accessed May 15, 2023, from https://oecd.ai/en/ai-principles

The IEEE CertifAIEd Framework for AI Ethics Applied to the City of Vienna. (2021). *IEEE SA - Standards Association.* Accessed May 15, 2023, from https://standards.ieee.org/beyond-standards/the-ieee-certifaied-framework-for-ai-ethics-applied-to-the-city-of-vienna/