Journal of Responsible Technology 25 (2026) 100143

Contents lists available at ScienceDirect

Journal of Responsible Technology

journal homepage: www.sciencedirect.com/journal/journal-of-responsible-technology

Research Article

# Enabling ethics mechanisms in the governance of algorithmic artificial persons (ALAP)

Siri Padmanabhan Poti *, Christopher J. Stanton, Catherine J. Stevens

*The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia*

ARTICLE INFO

*Keywords:*
Emerging information and communication technologies
Ethics library
Governance
Ethics-as-a-service
Anticipatory general ethics
Autonomous AI

ABSTRACT

In the global context of a 'new social contract' and a 'flourishing world', ethics mechanisms such as principles, guidelines, recommendations, standards, frameworks and checklists, are being established by public and private organisations and governments for the governance of 'algorithmic artificial persons (ALAP)', autonomous artificial intelligence (AAI) systems and emerging information and communication technologies (eICT). This paper examines, and identifies eleven gaps in the current ethics mechanisms, employing a 'qualitative evidence synthesis' (QES). Additionally, it proposes a 'prescriptive conceptual model' of an 'anticipatory general ethics library' (AnGEL), to enable resolution of these gaps. AnGEL is conceptually modelled as an implementable library of norms and rules for ALAPs / AAI systems and eICT, agnostic of domains and use cases. The conceptually modelled AnGEL may, post 'verification' through further discourse, and subsequent 'validation' of a prototype, be hosted on a cyber-physical intermediary. It can be rendered accessible through 'Ethics-as-a-Service' and provide 'ethics interoperability'.

## 1. Introduction

The fourth industrial revolution, or Industry 4.0, representing the aspirations of mostly the manufacturing sector, is disrupting across sectors and society through digitization, robotics, artificial intelligence (AI) and related emerging information and communication technologies (eICT) (KPMG et al., 2019; Lasi et al., 2014). Meanwhile, there is a gradual shifting of industrial paradigms, from the technocentric paradigm of Industry 4.0 to the sustaincentric paradigm of Industry 5.0. In this context, there are concerns as to how the ethical issues of non-human and autonomous agency of eICT would impact the well-being of individuals, families, societies and economies (Breque et al., 2021; KPMG et al., 2019; United Nations, 2024). eICT includes technologies that are still being researched, those in the early stages of prototype development, and the technologies that continue to hold novelty, such as Internet of Things (IoT), robotics, augmented reality, blockchain, along with the technologies that have the 'potential to disrupt traditional business models', such as 'agentic AI, AI governance platforms, disinformation security, post-quantum cryptography, ambient invisible intelligence, energy-efficient computing, hybrid computing, spatial computing, polyfunctional robots, and neurological enhancement' (Brey, 2012; Gartner, 2024). In this context, non-human autonomous 'agency' can be understood as 'artificial personhood' or as 'algorithmic artificial persons (ALAP)' (Kane, 2018, p. 256), which may be made up of one or more 'rational agents' that 'operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals' (Russell & Norvig, 2021, pp. 21-22). As there is general agreement on proactively addressing concerns regarding risks to humanity from maleficent agency of eICT (Brey, 2012), global agencies and fora have promulgated several ethics mechanisms, such as principles, guidelines, recommendations, standards, frameworks and checklists, for the ethics of the different aspects of eICT (Corrêa et al., 2024; Jobin et al., 2019; United Nations, 2024). However, these ethics mechanisms for eICT, which become important in the context of shifting industrial paradigms, have several gaps that challenge their design, development, implementation, and evaluation.

### 1.1. Research question

This paper poses the question 'What are the gaps in the current ethics

* Corresponding author at: The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia.

*E-mail addresses:* siri.padmanabhan@westernsydney.edu.au (S. Padmanabhan Poti), c.stanton@westernsydney.edu.au (C.J. Stanton), kj.stevens@westernsydney.edu.au (C.J. Stevens).

https://doi.org/10.1016/j.jrt.2025.100143

Available online 29 November 2025

mechanisms for the governance of ALAPs / AAI systems and eICT?', and additionally examines 'How may these gaps be addressed?'

### 1.2. Aim

This paper examines the gaps in the current ethics mechanisms in the governance of ALAPs / AAI systems and eICT. The paper conceptually models how ethics mechanisms can be enabled in the governance of ALAPs / AAI systems and eICT, agnostic of the domain and use cases in which they are employed. The 'prescriptive conceptual model' (Thalheim, 2019, p. 135) is a possible way ahead in the ethics discourse to resolve the gaps identified.

### 1.3. Background

#### 1.3.1. Autonomous AI agency

Autonomous AI (AAI) systems comprising AI 'agents', or more specifically, 'rational agents' that 'operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals' so as 'to achieve the best expected outcome' (Russell, 2010, p. 4). In remote, risk laden, or difficult operations, an autonomous system or agent can independently perform complex tasks, exhibit human-like abilities in dealing with ambiguities, while raising the possibility of unexpected or unintended outcomes due to both 'aleatoric' and 'epistemic' uncertainties (Hüllermeier & Waegeman, 2021, p. 459; Xu, 2021).

Their autonomy implies self-running systems possessing functions of perception and action (Jha et al., 2020) that are capable of dealing with uncertainties (Hüllermeier & Waegeman, 2021, p. 459). Their functions of perception which model the environment, and their functions of action which plan and execute to accomplish goals (Jha et al., 2020), are possible through their capabilities of prediction, classification and search. These capabilities are available through algorithms built with data intensive machine learning (ML) and deep learning (DL), or logic intensive AI programs (De Silva & Alahakoon, 2022, p. 10; Russell et al., 2022). It may be noted here that the agency of an autonomous AAI system may be seemingly single, but may consist of multiple agents, each with their own functions, with a common collective orientation to the objectives of the AAI system (Sifakis & Harel, 2023).

#### 1.3.2. Motivations in employing autonomous AI agency

In the social context, it is known how humans unable or unwilling to do a task due to lack of ability, time or other resources, engage with other human 'agents' or non-human 'agents', delegating to them the risks and decision making to accomplish the tasks or operations (Shapiro, 2005, p. 275). In this context, the key motivations for humans to employ AAI systems would be for the delegation of the risk and decision making in operations or tasks that are hazardous, financially impactful, mission-critical, safety-critical or humanly impossible (Nelson & NASA, 2003). To understand the relationship in such engagements or delegations, Shapiro (2005, p. 275) present 'agency' as an 'acting for' relationship, where the presence of a 'formidable physical, social, temporal or experiential barrier separates principal and agent'. In these agency relationships, 'one or more persons (the principal(s)) engage another person (the agent) to perform some service on their behalf which involves delegating some decision making authority to the agent' (Jensen & Meckling, 1976, p. 308).

AAI systems, with their ability to deal with complex situations (O'Neill et al., 2022; Xu, 2021), can be useful for humans to accomplish operational goals in a range of situations, from everyday tasks to mission-critical tasks (Nelson & NASA, 2003) in risky, unstructured, and partially-known environments (Lyons et al., 2024). Notably, while well-defined tasks with fixed rules can be performed by automated systems, an autonomous system can independently perform complex tasks, and adapt in the face of ambiguity. The ability of AAI systems for complex decision making in 'noisy, dynamic, and realistic environments' (Hagos & Rawat, 2022, p. 3), stems from their internal models that can learn and adapt to the operating environment in real-time. Such non-human agency needs to be understood in the context of society.

#### 1.3.3. Artificial personhood and a new social contract

To understand non-human agency and artificial personhood in the context of society, Thomas Hobbes's social contract theory may be extended to include artificial persons such as an 'Organizational Artificial Person (OAP)', as in the case of a legally set up corporation (Kane, 2018, p. 256; 2025). The Hobbesian contract may also extend to 'Professional Artificial Person (PAP)', as in the case of a lawyer, doctor or police officer (Kane, 2018, p. 256; 2025). These artificial persons are already provided for by society, and are serving social purpose, with legal and social legitimacy (Gordon, 2021; Kane, 2018, 2025; United Nations, 2024). The case for a similar social contract is made for the nomenclature of an AAI system with 'learned human behaviours' as an Algorithmic Artificial Person (ALAP) (Kane, 2018, p. 256; 2025) or 'electronic personhood' (de Pagter, 2024, p. 1284). ALAPs may be considered 'in a space between tools and human beings' and it becomes important to think up appropriate means of rendering these ALAPs 'societal', providing them with the ability of conforming to human social norms. In some ways, such a personhood may already be recognised as the fourth in the four legal classifications :'Natural persons', 'Non-natural persons with legal personhood' akin to the OAP discussed earlier, 'Non-human animals with no legal personhood', and 'Non-human and non-natural beings' that do not yet have legality (Gordon, 2021, p. 466).

The ideas regarding artificial personhood or electronic personhood for AAI systems and eICT, continue to be established further, as the United Nations considers a 'new social contract' to include AAI systems to enable 'a governance regime which protects and empowers us all' (de Pagter, 2024, p. 1284; Gordon, 2021; Kane, 2018, 2025; United Nations, 2024, p. 21). This would assume importance when dealing with the ethics, legality and social legitimacy of AAI systems and eICT (Gordon, 2021; Kane, 2018, 2025; United Nations, 2024, p. 21). While it is believed that the 'new social contract' will enable fair distribution of opportunities and risks (United Nations, 2024, p. 21), there are complexities to legal personhood and the ethical aspects of it (Gordon, 2021, p. 466). The understanding of AAI systems as ALAPs would be important for all stakeholders, such as those in the roles of 'jurisprudence', 'procurement', 'program managers', 'developers', system integrators, system evaluators, policymakers and 'end-users' / 'adopters' (Hoffman et al., 2023, pp. 6-7), in employing non-human autonomy as 'agents' of work in organisations and society.

However, the use of autonomous systems can also bring about unexpected, unintended or undesirable outcomes for organisations, society and humanity in general, especially in time-critical and safety-critical domains, such as the battlefield or in autonomous vehicles (Akhtom et al., 2024; Beer et al., 2014; Hüllermeier & Waegeman, 2021; Rawat, 2022; Xu, 2021).

#### 1.3.4. Ethics in the governance of ALAPs / AAI systems and eICT

There are concerns of ethical risks to humanity from ALAPs / AAI systems and related eICT (Brey, 2012). To deal with these ethical concerns the WEF et al. (2021, p. 5), while proposing 'epistemic principles' relating to 'interpretability', 'reliability', 'robustness', and 'security', also propose a set of 'general ethical principles' for AI, relating to moral and legal accountability, such as 'Accountability', 'Human agency', 'Data privacy', 'Safety', 'Lawfulness and compliance', 'Fairness' and 'Beneficial AI' (p. 5). Global agencies have put forth similar ethics mechanisms, such as principles and guidelines for ALAPs / AAI systems and related eICT (Corrêa et al., 2024; Jobin et al., 2019; United Nations, 2024). Such ethics mechanisms are seen to have several gaps in how they are developed, implemented, evaluated and the contexts in which they are intended to be applied (Hagos & Rawat, 2022).

## 2. Method

### 2.1. Research pathway

de Pagter (2024, p. 1281) suggest four pathways for research on ethics for AI and emerging technologies i) 'providing insights for the improvement of ethical assessment', ii) examining the 'moral competence of artificial agents', iii) 'value-based design and implementation of robots', and iv) 'discussions on building ethical sociotechnical systems around robots'. This research follows the fourth pathway of 'discussions on building ethical sociotechnical systems around robots' (p. 1281). In doing so, it also seeks to contribute to the first path way of 'providing insights for the improvement of ethical assessment' (p. 1281). Accordingly, we examine the gaps in the current ethics mechanisms in the governance of ALAPs / AAI systems and eICT, employing a 'Qualitative Evidence Synthesis' (QES) (Grant & Booth, 2009, p. 94) to contribute to the ethics discourse.

### 2.2. Qualitative evidence synthesis

The QES integrates findings, themes and constructs from prior reviews to synthesise an 'overarching narrative', as an interpretation to broaden the understanding of the subject (Evidence-informed policy network, 2021; Grant & Booth, 2009, p. 94; Noyes et al., 2008). This QES exercise derives from prior reviews in the area of ethics mechanisms for eICT, alongside supplementary prior reviews and supplementary primary evidence in the form of ethics mechanisms from influential international bodies. The QES synthesises recent peer reviewed journal articles of type 'Reviews', published in English, from 2020 to 2025 (Inclusive), indexed in the Scopus database.

### 2.3. Search criteria

The search criteria employed is: TITLE-ABS-KEY (("Artificial Intelligence" OR Autonom* OR AI) AND (ethic* AND (guideline* OR Framework* OR Recommend* OR Principle* OR Governance))) AND PUBYEAR > 2019 AND PUBYEAR < 2026 AND (LIMIT-TO (SRCTYPE, "j")) AND (LIMIT-TO (DOCTYPE, "re")) AND (LIMIT-TO (LANGUAGE, "English")). Using this search, 2,149 articles were found. The results were further filtered for those related to governance with the filter: AND (LIMIT-TO (EXACTKEYWORD, "Governance")). 22 articles were identified. Of these, 7 articles that were not relevant were screened out. 15 peer reviewed articles were chosen for synthesis, referred to here as 'prior reviews'. We synthesise the findings from prior reviews available from the search.

### 2.4. Supplementary evidence

Further, supplementing the search as part of this QES, a 'relatively small' (Campbell et al., 2020, p. 653) 'purposive sampling' (Grant & Booth, 2009, p. 94; Noyes et al., 2008) of: three prior reviews and additional primary evidences such as relevant principles, guidelines, standards, frameworks and checklists, established by key international agencies, are included. Additional prior reviews and primary evidences are included to obtain 'a deeper understanding' (Barroga & Matanguihan, 2022, p. 6) of the insights available from the systematically obtained prior reviews. The review articles from the purposive sampling include i) the scoping review by Jobin et al. (2019): 'Artificial Intelligence: the global landscape of ethics guidelines', ii) the meta-analysis by Corrêa et al. (2023): 'Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance', and iii) the semi-systematic review by Hagendorff (2020): 'The Ethics of AI Ethics: An Evaluation of Guidelines'. The purposive sampling includes ethics mechanisms for eICT established by key international agencies such as the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG, 2019; 2020, p. 4), World Economic Forum (WEF et al., 2021), IEEE (IEEE, 2019), ILO, Hiroshima Process International of OECD (Hiroshima Process International, 2023a), United Nations OICT and UNESCO (UNESCO, 2022; United Nations, 2024). The prior reviews and primary evidences contributing to this QES are as shown in *Table 1: Prior Review and Primary Evidence contributing to the Qualitative* Evidence Synthesis (QES).

### 2.5. Additional contribution

This paper additionally synthesises a novel 'prescriptive conceptual model' (Thalheim, 2019, p. 135) of an anticipatory general ethics library (AnGEL) which could help in 'providing insights for the improvement of ethical assessment' (de Pagter, 2024, p. 1281) and overall ethics mechanisms, by enabling resolution of the eleven gaps in the ethics mechanisms for ALAPs / AAI systems and related eICT. The conceptual model of AnGEL does not claim to be a validated solution, but a guidance to the work ahead in the area of enabling ethics mechanisms.

## 3. Ethics mechanisms for ALAPs / AAI systems and eICT

### 3.1. Early landscape of ethics mechanisms

The United Nations recognized the need for ethics mechanisms, and tasked its principal agency UNESCO with the promotion and dissemination of the recommendations related to the ethics of AI, with the stated goal of collaboration with entities, such as the 'World Commission on the Ethics of Scientific Knowledge and Technology (COMEST)', and other 'international, regional and sub-regional governmental and non-governmental organizations' (UNESCO, 2022). In its 'Recommendation on the Ethics of Artificial Intelligence', UNESCO (2022) expects member states and other stakeholders to 'respect, promote and protect the ethical values, principles and standards regarding AI that are identified in this Recommendation'. It recommends that stakeholders 'should take all feasible steps' to implement these recommendations. This recommendation has a stated goal of providing the basis for employing AI systems 'for the good of humanity, individuals, societies and the environment and ecosystems' and also promoting the use of AI for peaceful purposes, while preventing harm through its use. This recommendation seeks the position of a 'globally accepted normative instrument'. It places emphasis on the 'articulation of values and principles', as also the 'practical realisation' through 'concrete policy recommendations' (UNESCO, 2022, p. 14).

Meanwhile, the 'Ethics Guidelines for Trustworthy AI', published by the European Commission (EC) (AI HLEG, 2019), emerged as an influential voice in AI Ethics. The guidelines present 'Trustworthy AI' as comprised of 'Lawful AI', 'Ethical AI' and 'Robust AI'. Ethical AI is further developed through four 'Ethical Principles': 'Respect for human autonomy', 'Prevention of harm', 'Fairness', and 'Explicability' (p. 8). Further it lists a detailed 'Trustworthy AI Assessment List (Pilot Version)' for self-assessment (p. 26), built on seven top level requirements . It sets the expectation for developers to apply the requirements to design and development processes (p. 35). 'Deployers' are expected to 'ensure that the systems they use and the products and services they offer meet the requirements', and without assigning a specific ownership, expects 'end-users and the broader society [to be] informed about these requirements and able to request that they are upheld' (AI HLEG, 2019, p. 35).

The EC worked further on what was originally a self-assessment check list within the AI HLEG (2019). It published an expanded, full-fledged assessment checklist and online web tool, the 'Assessment List for Trustworthy AI (ALTAI) (AI HLEG, 2020), based on feedback from a piloting exercise across industry. This checklist was intended for a wide range of stakeholders including 'AI designers and AI developers of the AI system', 'data scientists', 'procurement officers or specialists'; 'front-end staff that would use or work with the AI system', 'legal/-compliance officers', and 'management' (AI HLEG, 2020, p. 4).

**Table 1**
Prior review and primary evidence contributing to the qualitative evidence synthesis (QES).

| Review article | Domain / area | Focus of prior review / primary evidence |
|---|---|---|
| (Barrance et al., 2022) | AI assurance ecosystem and UK's AI Assurance Roadmap | This review of a single ethics mechanism, 'AI Assurance Roadmap' published by the Centre for Data Ethics and Innovation (CDEI), UK, examines the Roadmap's view on the 'complexities of building justified trust', 'role of research in AI assurance, 'current developments in the AI assurance industry', and the 'convergence with international regulation' ( Barrance et al., 2022, p. 1). |
| (Birkstedt et al., 2023) | AI governance | This systematic literature review examines future implications of AI governance for society and organisational stakeholders. |
| (Brandão, 2025) | Societal impact of AI | This narrative review explores the multifaceted impact of AI and highlights the need for further discourse on ethical and sustainable integration of AI in society (Brandão, 2025, p. 1). |
| (Cancela Outeda, 2024) | AI governance and EU AI Act (AIA) | This analysis of AIA examines the strengths and areas for improvement of 'the mechanisms and structures designed to implement AI across the EU'. It discusses collaborative governance ( Cancela Outeda, 2024, p. 2). |
| (Ho & Asrani, 2025) | Governance in AI for Hepatology | This informal review of existing AI frameworks, principles of medical ethics, and quality healthcare principles, synthesises a framework for AI governance in hepatology and related healthcare (Ho & Asrani, 2025, p. 1). |
| (Ho & Caals, 2021) | AI ethics and digitalization of Nephrology care | This non-systematic review examines the ethical and regulatory challenges for AI/ ML-based software and devices in nephrology and provides a framework for 'scaling them up toward broader clinical application' (Ho & Caals, 2021, p. 282). |
| (Kalenzi, 2022) | Governance of blockchain, AI and related emerging technologies | This non-systematic review Examines the governance of blockchain, AI and the 'ecosystem of emerging platforms within industry and public and civil society' from an Industry 4.0 perspective ( Kalenzi, 2022, p. 1). |
| (Kazim & Koshiyama, 2021) | AI ethics and interdisciplinarity | This non-systematic review discusses the ontologies of 'AI' and 'ethics', the mechanisms of AI ethics such as 'principles, processes, and ethical consciousness'. It explores the 'central themes in translating ethics in to engineering practice' and highlights the inherent interdisciplinarity of the AI ethics discourse (Kazim & Koshiyama, 2021, p. 1). |
| (Kim et al., 2025) | AI governance and the EU AI Act (AIA) | This systematic bibliometric analysis examines the research trends in AI governance in terms of the framework of the European Union Artificial Intelligence Act (EU AI Act). |
| (Langman et al., 2021) | AI governance and Roboethics | This review of ethical frameworks and regulations in the European Union, United States of America, and Canada, examines 'how ethical principles have been translated into law by government bodies' (Langman et al., 2021). |
| (Sigfrids et al., 2022) | Public administration and AI governance | This systematic review on 'the development of AI governance for public administration' examines 'how public administrations [may] guide … AI developers and users in adopting ethical and responsible practices'. It proposes a 'Comprehensive, Inclusive, Institutionalized, and Actionable' framework for AI governance (Sigfrids et al., 2022, pp. 1,8). |
| (Tun et al., 2025) | AI governance readiness in healthcare | This review of 'official policies and guidelines, government health ministries' websites, … online sources' and 'guidance on digital health policies specific to the ASEAN region' examines the gaps in AI-related policies in the healthcare sector of the ASEAN region. |
| (Pabel & Akther, 2025) | AI governance in accounting and finance | This systematic literature review of ethical issues and risks from implementing AI in the accounting profession, presents an ethical governance framework to minimize risk. |
| (Leikas et al., 2022) | AI governance in the public sector | This is a case study of a 'Finnish national AI program : AuroraAI, which aims to provide citizens with tailored and timely services for different life situations, utilizing AI' (Leikas et al., 2022, p. 1). It deals with the role of society, communities and people in the discourse of governance of emerging technologies, and also the challenges from applying AI in public administration. |
| (Mirghaderi et al., 2023) | AI governance in digital platforms | This survey of literature examines 'zones of non-transparency in the context of digital platforms …that utilize artificial intelligence (AI)' It attempts to 'bridge the gap between principles and operationalization' ( Mirghaderi et al., 2023, p. 831) |
| **Supplementary prior reviews as purposive sampling** | | |
| (Corrêa et al., 2023) | AI ethics principles and guidelines | This meta-analysis of guidelines and recommendations for AI governance summarises the inputs from across 32 countries and 5 languages, to list out 17 guiding principles. |
| (Hagendorff, 2020) | AI ethics principles and guidelines | This Semi-systematic evaluation of 'major ethical guidelines', for AI list out the concerns, in the order of |

(continued on next page)

**Table 1** (*continued*)

| Review article | Domain / area | Focus of prior review / primary evidence |
|---|---|---|
| | | number of mentions, and discusses means to deal with the 'relative ineffectiveness' of ethics mechanisms ( Hagendorff, 2020). |
| (Jobin et al., 2019) | AI ethics principles and guidelines | This scoping review of existing guidelines on ethical AI principles or guidelines looks for global convergence 'in the principles for ethical AI and the requirements for its realization' (Jobin et al., 2019, p. 3) |
| **Supplementary primary evidence as purposive sampling** | | |
| (AI HLEG, 2019) | Ethics guidelines for AI | Published by the European Commission, it presents 'Trustworthy AI' as comprised of 'Lawful AI', 'Ethical AI' and 'Robust AI'. Ethical AI is further developed through four 'Ethical Principles': 'Respect for human autonomy', 'Prevention of harm', 'Fairness', and 'Explicability' ( AI HLEG, 2019, p. 8) |
| (AI HLEG, 2020) | Checklist for AI ethics self-assessment for trustworthy AI | Assessment checklist and online web tool, expansion of the self-assessment check list within the AI HLEG (2019, p. 26) based on feedback from a piloting exercise across the industry. |
| (Hiroshima Process International, 2023a) | Code of conduct for AI development organisations | There is a free-standing statement imploring organisations to 'respect the rule of law, human rights, due process, diversity, fairness and non-discrimination, democracy, and human-centricity, in the design, development and deployment of advanced AI systems'. |
| (Hiroshima Process International, 2023b) | Ethics guidelines for AI development organisations | A 'non-exhaustive list of guiding principles' built on the OECD AI Principles (OECD, 2024). Intended to serve as a 'living document' applicable to all AI related beings and stakeholders involved in the 'design, development, deployment and use of advanced AI systems' ( Hiroshima Process International, 2023b). |
| (IEEE, 2019) | Ethical principles for autonomous and intelligent systems | Conceptual framework built on the 'three pillars' of 'Universal Human Values, Political Self-Determination and Data Agency, [and] Technical Dependability'. This is detailed into eight 'general principles' ( IEEE, 2019). |
| (IEEE, 2021) | Standard for addressing ethical concerns during system design | The 'IEEE Standard Model Process for Addressing Ethical Concerns during System Design' is of the IEEE Std 7000™ series. This is freely available as part of the IEEE GET Program. It is a 'set of processes by which organizations can include consideration of ethical values throughout the stages of concept exploration and development' (IEEE, 2021, p. 2). |
| (OECD, 2024). | Intergovernmental recommendation and standard on AI ethics | This is claimed to be the 'the first intergovernmental standard on AI', available as a legal instrument in the form of a recommendation, with its language and structure as a legal or 'soft law' document ( OECD, 2024, pp. 8-9). |
| (UNESCO, 2022) | Recommendation on AI Ethics | As the principal agency of the United Nations tasked with the promotion and dissemination of the recommendations related to the ethics of AI, this recommendation seeks the position of a 'globally accepted normative instrument'. It places emphasis on the 'articulation of values and principles', as also the 'practical realisation' through 'concrete policy recommendations' (UNESCO, 2022, p. 14). |
| (United Nations, 2024) | Global governance of AI | Report by UN's 'High-level Advisory Body on Artificial Intelligence' on recommendations for the international governance of AI. It details the need for global governance of AI, the gaps in governance, and recommendations for enhancing global cooperation. |
| (WEF et al., 2021) | Ethical AI principles for organisations | Guidance on AI principles, recommends nine 'core principles', 'epistemic' and 'general', for ethical AI. Intended as a baseline for assessment of ethical validity of an AI system (WEF et al., 2021). |
| (WEF et al., 2024) | Technology policy for responsible design of eICT | This is a report that lays out the policy for eICT, especially 'AI, blockchain, cloud, agritech, 5G, advanced energy solutions and quantum' to build a 'flourishing world' (WEF et al., 2024, p. 5). |

The World Economic Forum (WEF) published '9 ethical AI principles for organizations to follow' for 'emerging technologies' (WEF et al., 2021). In this guidance the WEF deemed it necessary to include human rights as a 'foundation of AI practices' and to 'establish moral and legal accountability' for 'the common good'. It recommended nine 'core principles' for ethical AI, separated as 'epistemic' and 'general' principles. This separation has been useful in the treatment of how governance aspects such as ethical aspects are addressed differently from epistemic aspects of explainability and interpretability, and has been relied on by this paper for contributing to the ethics discourse. Specifically, WEF et al. (2021) claimed that the published principles can form 'a baseline for assessing and measuring the ethical validity of an AI system'. The 'epistemic principles' are 'Interpretability (Explainability, Transparency, Provability), Reliability, Robustness, [and] Security'. The 'general ethical AI principles' are: 'Accountability', 'Human agency', 'Data privacy', 'Safety', 'Lawfulness and compliance', 'Fairness' and 'Beneficial AI' (WEF et al., 2021). While the separation of principles into epistemic and general is useful for ethics discourse, as noted earlier, it has not been seen as forming the baseline for assessment or

measurement.

Similarly, the IEEE (2019) conceptual framework, 'Ethically Aligned Design' for autonomous and intelligent systems, prioritised human well-being, being built on the 'three pillars' of 'Universal Human Values, Political Self-Determination and Data Agency, [and] Technical Dependability'. This is detailed into eight 'general principles'.. However, much of these principles rely on the creator / developer / deployer to translate these to implementable ethics to incorporate into intelligent systems (IEEE, 2019).

### 3.2. Proliferation and divergence

Despite international organisations such as the UN,WEF, EC and the IEEE at the helm of the ethics efforts, questions arose as to any substantial 'global agreement' or convergence on 'what constitutes "ethical AI" and which ethical requirements, technical standards and best practices are needed for its realization' (Jobin et al., 2019). There was also 'substantive divergence' on four questions: i) 'how [are] the ethical principles interpreted'? ii) 'why [are these ethical principles] deemed important'? iii) what issue, domain or actors [do these ethical principles] pertain to'? and (iv) 'how [can these ethical principles] be implemented'? (p. 14). There has been a proliferation of guidelines, recommendations and other ethics mechanisms for ALAPs / AAI systems and eICT authored by private organisations, governmental agencies, academic institutions, inter-governmental organisations, non-profit organisations and professional associations (Corrêa et al., 2023; Jobin et al., 2019). Based on a meta-analysis of two hundred ethics mechanisms by Corrêa et al. (2023) on 'governance policies and guidelines' published globally, from across thirty two countries and five languages, a prevalence of seventeen guiding principles was observed (Corrêa et al., 2023). In a similar study on global ethics mechanisms for ALAPs / AAI systems and eICT, there appeared to be general agreement regarding the ethical principles on 'transparency, justice and fairness, non-maleficence, responsibility and privacy' (Jobin et al., 2019, p. 7). However, there was seen to be scope for better inclusion of the principles relating to 'beneficence, freedom and autonomy, trust, sustainability, dignity and solidarity' (p. 7). Also, in a 'semi-systematic' analysis of twenty-two 'major ethical guidelines', Hagendorff (2020) found the ethics concerns, in the order of number of mentions, to be ranging from 'Privacy protection, fairness, non-discrimination, justice, accountability, transparency, openness, safety, cybersecurity' to 'protection of whistleblowers', 'cultural differences in the ethically aligned design of AI", and the 'hidden costs of AI' (Hagendorff, 2020).

### 3.3. Inadequate mechanisms

Notably, the objectives, scoping, intent and limitations are clearly stated in the ethics mechanism by UNESCO (2022, pp. 18-20). The principles are also detailed such as to 'unpack the values underlying them more concretely so that the values can be more easily operationalized in policy statements and actions' (pp. 18-20). The mechanism further details eleven 'policy areas' that are meant to further aid in operationalising the values and principles identified earlier. However, these ethics mechanisms are perceived to have 'no enforcement mechanisms reaching beyond a voluntary and non-binding cooperation' between 'ethicists and individuals working in research and industry' Hagendorff (2020). For example, there are no policy or recommendations in this mechanism by UNESCO (2022, pp. 18-20) on the use of AI in warfare, for military purposes or for defence. It is not clear whether this was left out for political reasons, or due to oversight. Ethics mechanisms must extend beyond 'checkbox guidelines' and 'soft governance' to overcome the perceived 'relative ineffectiveness' (Hagendorff, 2020, pp. 113-114). There is also the need for 'integrating guideline development efforts with substantive ethical analysis and adequate implementation strategies' (Jobin et al., 2019, p. 1). It would be necessary to move from 'principle-formulation' activities to 'translation into practice' related activities (p. 17).

### 3.4. Global vs local and general vs contextual

There are arguments for a balance between a 'global agenda for ethical AI' and 'a harmoni[s]ation' allowing for 'cultural diversity' and 'moral pluralism' (Jobin et al., 2019, p. 17) across nations and domains. There is also the need to develop 'comprehensive AI governance frameworks beyond the EU AI Act' (Kim et al., 2025, p. 144126). There are stark variations in the readiness of different governments, in terms of 'vision, governance and ethics' (Tun et al., 2025, pp. 2-5, 8). Further, moving from the general to the specific contexts, there is necessity for 'regulatory frameworks for comprehensive AI strategies' the governance of AI for healthcare (pp. 2-5, 8). This highlights the need for 're-examining existing legal and institutional arrangements to check whether these cater for emerging issues with new technologies' (Kalenzi, 2022, p. 1).

### 3.5. Tokenism and platitudes

There are more recent mechanisms that suffer from structural and linguistic difficulties, along with tokenisms and platitudes, such as the 'HPI Code of Conduct for Organizations Developing Advanced AI Systems' Hiroshima Process International (2023a), and the 'HPI Guiding Principles for Organizations Developing Advanced AI system' Hiroshima Process International (2023b).

The ethics mechanism in Hiroshima Process International (2023b) is meant to be a 'non-exhaustive list of guiding principles' and meant to serve as a 'living document' applicable to all AI related beings and stakeholders involved in the 'design, development, deployment and use of advanced AI systems'. Notably, it is built on the OECD AI Principles (OECD, 2024) especially as a response to the risks and opportunities from the progress in technologies since the initial laying out of the OECD principles. Unlike most mechanisms, the eleven principles are verbose and not pithy phrases. The mechanism requires or more aptly, 'calls on' organisations 'to abide by the … principles, commensurate to the risks', which does not seem to provide additional value to the discourse. The first principle is open in in its construct, asking the organisation to 'Take appropriate measures … to identify, evaluate, and mitigate risks across the AI lifecycle'. There are additional clauses added in an effort to make the principle complete, and in this case, the 'AI lifecycle' is further detailed in ' throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market'. While the first principle is a complete sentence, the second principle is cryptic in its construct: 'Patterns of misuse, after deployment including placement on the market'. The third principle requires organisations to 'Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability'. The overall language, structuring and scoping of principles is quite varied, with mostly high-level statements and suffering from overall impracticability. The principles are closed with the statement: 'Appropriate transparency of training datasets should also be supported and organizations should comply with applicable legal frameworks' (Hiroshima Process International, 2023b). Given the points discussed, the value of guidance or legality from such mechanisms is debatable, regardless of the position or influence or intent of the body that promulgates such mechanisms.

Similarly, the Hiroshima Process International (2023a) mechanism further elaborates on each of the principles in Hiroshima Process International (2023b). Prior to that it puts forth statements that are platitudes taking moral high ground that do not seem to have the force of law or consequence. At the same time, such statements seem to have a limited contribution to the overall discourse or the direction of ethics for eiCT. There is a free-standing statement imploring organisations to 'respect the rule of law, human rights, due process, diversity, fairness

and non-discrimination, democracy, and human-centricity, in the design, development and deployment of advanced AI systems'. There is also a similar statement imploring states to 'abide by their obligations under international human rights law to ensure that human rights are fully respected and protected'. There are further unlinked references to other frameworks, namely the 'United Nations Guiding Principles on Business and Human Rights' and 'OECD Guidelines for Multinational Enterprises'. Similar to the Hiroshima Process International (2023b), this Code of Conduct does not offer value either as an enforceable framework or a guiding mechanism with practicability.

### 3.6. Responsible design and a flourishing world

It was noted earlier in this section how the IEEE (2019) conceptual framework relied on the creator /developer / deployer to translate the principles onto implementable ethics. However, as part of getting from principles to practice, the IEEE runs an IEEE GET Program for AI Ethics and Governance Standards with the relevant IEEE Std 7000™ series standard made freely available, for example the 'IEEE Standard Model Process for Addressing Ethical Concerns during System Design' (IEEE, 2021). The 'Governing AI for Humanity: Final Report' (United Nations, 2024) analyses and advances recommendations for the international governance of AI by UN's 'High-level Advisory Body on Artificial Intelligence'. The advisory body includes multiple stakeholders as part of the United Nations Secretary-General's Roadmap for Digital Cooperation. It details the need for global governance of AI, the gaps in governance, and recommendations for enhancing global cooperation. However, the five principles it puts forth are at a high level of governance and leadership in that they are meant to 'guide the formation of new international AI governance institutions'. Towards enhancing global cooperation in matters related to AI, it makes eight recommendations..

There is an expectation that implementing the UN recommendations would 'encourage new ways of thinking: a collaborative and learning mindset, multi-stakeholder engagement and broad-based public engagement'. As mentioned in the introduction, the report expresses the possibility of a 'new social contract for AI' (United Nations, 2024) and the opportunity to bring in protections that were missed in the area of climate change. The ethics mechanism is one that is visionary, as would be expected of a global organization such as the UN. The mechanism would need a layer of translation to implement as ethics in ALAPs / AAI systems and eICT.

The recommendations from the OECD Council on Artificial Intelligence make up for the next layer of translation from the UN report. OECD (2024), with its language and structure as a legal or 'soft law' document, is available as a legal instrument, albeit recommendation. It may be noted that the OECD has different 'legal instruments' such as those listed here in the order of legal force: i) 'international agreements' which are legally binding, ii) 'decisions' which are legally binding, iii) 'recommendations' which are not legally binding, but represent a political commitment, with the expectation that the audience would 'do their best to implement them', iv) 'substantive outcome documents', and v) 'arrangements, understanding' and other legal instruments of mixed nature. This OECD (2024) mechanism is structured as a recommendation.

With its claims to be 'the first intergovernmental standard on AI' along with its adoption by the OECD followed by several updates at international ministerial meetings, this is perhaps one of the most influential ethics mechanisms for eICT. It details five principles and five recommendations for all stakeholders across its member and non-member states. It provides initial definitions for 'AI system', AI system lifecycle', 'AI stakeholders' and 'AI actors'. It also lists the items that constitute 'AI knowledge'

In like manner, WEF's 'Technology Policy: Responsible Design for a Flourishing World' lays out the policy for eICT, especially 'AI, blockchain, cloud, agritech, 5G, advanced energy solutions and quantum' (WEF et al., 2024, p. 5). to build a 'flourishing world'. A 'flourishing world' is a construct where 'sustainable economic development does not leave nature behind'. It is envisaged as a 'mutually reinforcing' society in which 'individual gain' and 'collective societal progress' go hand in hand. The principles of 'freedom, equity and human rights' are 'undisputable' (p. 1). It presents eight guideposts that are meant to inform the 'regulator, the regulated and the public'. These guideposts serve as 'readiness or fitness reminders' for the state of completion or preparedness (p. 5). Importantly it seeks to enable 'a …nimble approach to policymaking in technology through 'extended multi-stakeholderism' (p. 29).

The structure, language, and content of these current ethics mechanisms from the UN, IEEE, OECD, and WEF are promising. Yet there are gaps that may be intrinsic to how the ethics mechanisms are available for implementation in ALAPs / AAI systems and eICT. The types of gaps observed in ethics mechanisms for ALAPs / AAI systems and eICT are detailed in the next section.

## 4. Eleven gaps in ethics mechanisms

Even with the variety and number of ethics mechanisms published for eICT 'none of them can be truly global in reach and comprehensive in coverage' in the words of the 'Governing AI for Humanity' report by the United Nations (2024). This implies issues and gaps in 'representation, coordination and implementation' (United Nations, 2024, pp. 42-45). Based on the understanding from the qualitative synthesis of the prior reviews, purposive sampling and related literature in the area of ethics mechanisms for ALAPs / AAI systems and eICT, there are gaps as listed in *Table 2: Gaps in Ethics Mechanisms*. The challenges gathered from literature are grouped by gaps in *Table 3: Challenges in Ethics Mechanisms, Grouped by Gaps*. Some of the challenges identified may easily fit into more than one 'gap'. However, here we have kept the scope of the gap broad. The challenges identified in the ethics mechanisms are grouped as gaps.

### 4.1. Approach gaps

These represent the challenges in the way ethics structures are approached differently. Ethics for ALAPs / AAI systems and eICT may be seen as being researched, planned, developed, and implemented using varied approaches. There is the 'principles approach' that includes

**Table 2**
Gaps in ethics mechanisms.

| | Gaps | Description |
|---|---|---|
| 1 | Approach | Challenges in the way ethics structures are approached differently. |
| 2 | Audit traceability | Challenges in tracing out ethics related implementations in a product or service. |
| 3 | Binding legality | Challenges in the differences of legal requirements and consequences. |
| 4 | Coordination or 'Tower of Babel' | Challenges of coordinating globally across nations and domains. |
| 5 | Divergence and delimiting | Challenges in differing priorities and interpretations. |
| 6 | Implementation or principle-practice | Challenges in translating principles to practice. |
| 7 | Intention-action | Challenges of semantically distance between guidelines and implementation frameworks. |
| 8 | Perception-reality / responsibility | Challenges of navigating between perception and real experience of production teams, collaborators and end users. |
| 9 | Readiness | Challenges of nations, organisations and domains, in readying for ALAP / AAI / eICT risks and benefits. |
| 10 | Representation | Challenges of inadequate representation of regions, cultures, or domains. |
| 11 | Underdetermination | Challenges of instructions for normative orientation to human production teams. |

**Table 3**
Challenges in ethics mechanisms, grouped by gaps.

| Ethics mechanism-related challenges |
|---|
| **1. Approach gaps** |
| ■ 'Ethical consciousness approach' that includes 'societal and cultural shifts, awareness and training' (Kazim & Koshiyama, 2021, pp. 1, 6). |
| ■ 'Principles approach' that includes abstract first principles and legislation (Kazim & Koshiyama, 2021, pp. 1, 6). |
| ■ 'Processes approach' that includes ethics-by design and governance (Kazim & Koshiyama, 2021, pp. 1, 6). |
| **2. Audit traceability gaps** |
| ■ 'Limited understanding of the implementation of AI governance in organizations' (Birkstedt et al., 2023, p. 133). |
| ■ 'Uncertain effectiveness of ethical principles and regulation' (Birkstedt et al., 2023, p. 133). |
| ■ Need for 'Accreditation for auditors' (Barrance et al., 2022, p. 10). |
| ■ Need for 'Comprehensive frameworks on governance' (Sigfrids et al., 2022, pp. 1,8). |
| ■ Need for ensuring 'a holistic understanding' (Cancela Outeda, 2024, p. 2). |
| **3. Binding legality gaps** |
| ■ 'Varied Soft and Strict regulations' (Corrêa et al., 2023, pp. 10-11). |
| ■ AI ethics is seen as a 'soft governance' with 'no enforcement mechanisms reaching beyond a voluntary and non-binding cooperation' between 'ethicists and individuals working in research and industry' (Hagendorff, 2020, pp. 113-114). |
| ■ Ethics mechanisms must extend beyond 'checkbox guidelines' (Hagendorff, 2020, pp. 113-114). |
| ■ Need for 'Creation of an AI Standards Hub' (Barrance et al., 2022, p. 10). |
| ■ Need for 'Establishment of policies and regulations that enforce roboethics policies' (Langman et al., 2021) |
| ■ Need to determine the optimal choice between 'self-regulation' and state regulation' to monitor emerging technologies (Kalenzi, 2022, p. 1). |
| **4. Coordination or 'Tower of Babel' gaps** |
| ■ 'AI governance should implement agile and adaptive governance processes' towards 'improving public governance and coordination of AI' (Sigfrids et al., 2022, pp. 1,8). |
| ■ Need for 'Collaborative governance mechanisms' (Cancela Outeda, 2024, p. 2). |
| ■ Need for 'Continuous stakeholder engagement' (Cancela Outeda, 2024, p. 2). |
| ■ Need for 'Ongoing dialogue' (Cancela Outeda, 2024, p. 2). |
| ■ Need for 'Regulatory agency or relevant governing institution to support operationalization of good governance and ethical principles' (Sigfrids et al., 2022, pp. 1,8). |
| ■ Need for 'Universal framework for responsible AI' (Brandão, 2025, pp. 25-26). |
| **5. Divergence and delimiting gaps** |
| ■ 'Different regions approach AI regulation with diverging priorities—the EU emphasizing ethics and precaution, the US prioritizing innovation, and China favouring centralized governance' (Brandão, 2025, pp. 25-26). |
| ■ 'Divergences of definitions and principles' (Corrêa et al., 2023, pp. 10-11). |
| ■ 'Substantive divergence' 'on interpretation of ethical principles (Jobin et al., 2019, pp. 1, 7, 14, 17). |
| ■ Divergence on implementation of ethical principles (Jobin et al., 2019, pp. 1, 7, 14, 17). |
| ■ Divergence on importance of the ethical principles (Jobin et al., 2019, pp. 1, 7, 14, 17). |
| ■ Divergence on issue, domain or actors that the ethical principles pertain to (Jobin et al., 2019, pp. 1, 7, 14, 17). |
| ■ Need for 'Converging with global regulation attempting to govern the use of AI' (Barrance et al., 2022, p. 10). |
| ■ Need for 'Re-examining existing legal and institutional arrangements to check whether these cater for emerging issues with new technologies' (Kalenzi, 2022, p. 1). |
| ■ Need for AI ethics governance structures to prepare for future challenges (Ho & Asrani, 2025). |
| ■ Need to Account for 'divergence and convergence of motivations' of 'public, private, and civil society organizations' in engaging and regulating emerging technologies (Kalenzi, 2022, p. 1). |
| ■ Need to update 'existing legal and regulated systems' to cover for capabilities from new technologies' (Kalenzi, 2022, p. 1). |
| ■ The pace of 'development of policy and regulation', for 'high-risk AI systems' to match technological advancement (Kim et al., 2025, p. 144126). |
| ■ There is scope for better inclusion of the principles relating to 'beneficence, freedom and autonomy, trust, sustainability, dignity and solidarity' (Jobin et al., 2019, pp. 1, 7, 14, 17). |
| **6. Implementation or principle-practice gaps** |
| ■ 'Abundance of high-level principles in the literature that cannot be easily operationalized' (Mirghaderi et al., 2023, p. 831). |
| ■ 'Insufficient operationalization of AI governance processes' (Birkstedt et al., 2023, p. 133). |
| ■ 'Lack of tools' to action the intent (Corrêa et al., 2023, pp. 10-11). |
| ■ 'Research should explore the best practices and challenges in AI governance … at the implementation level, so that all … states can harness the potential of AI to revolutionize their healthcare systems' (Tun et al., 2025, pp. 2-5, 8). |
| ■ Need for 'Establishment of national and international agreements for ethical robotic research, development, and deployment' (Langman et al., 2021). |
| ■ Need for 'Flexible and adaptable regulatory framework' (Cancela Outeda, 2024, p. 2). |
| ■ Need for integrated 'guideline development efforts with substantive ethical analysis and adequate implementation strategies' (Jobin et al., 2019, pp. 1, 7, 14, 17). |
| ■ Need to move from 'principle-formulation' activities to 'translation into practice' related activities (Jobin et al., 2019, pp. 1, 7, 14, 17). |
| **7. Intention-action gaps** |
| ■ Need for 'Adoption of AI assurance standards' (Barrance et al., 2022, p. 10). |
| ■ Need for 'Alignment of [AI program] funding program's guidelines with ethical principles' (Langman et al., 2021). |
| ■ Need for 'Creation of widely adopted technical standards' (Barrance et al., 2022, p. 10). |
| ■ Need to 'Establish best practices' (Barrance et al., 2022, p. 10). |
| Need to Apply 'Human-rights standards and approaches-based actionable ethical principles' in policy making (Sigfrids et al., 2022, pp. 1,8). |
| **8. Perception-reality / responsibility gaps** |
| ■ 'Difficulty of assigning blame to artificial entities that learn and function autonomously' (Pabel & Akther, 2025, p. 14). |
| ■ 'Dilemma in 'delegate[ing] duties to machines' with unclear ethicality (Pabel & Akther, 2025, p. 14). |
| ■ 'Lack of attention to the AI governance context' (Birkstedt et al., 2023, p. 133). |
| ■ 'No attention to Hidden costs of AI' (Corrêa et al., 2023, pp. 10-11). |
| ■ Need for 'Collaborative perspective' (Cancela Outeda, 2024, p. 2). |
| ■ Need for 'Crafting effective and comprehensive AI regulations' (Cancela Outeda, 2024, p. 2). |
| ■ Need for 'Mitigating risks associated with AI deployment' (Brandão, 2025, pp. 25-26). |
| ■ Need for 'Proactive governance and ethical frameworks' (Brandão, 2025, pp. 25-26). |
| **9. Readiness gaps** |
| ■ 'Some … states demonstrate[e] significant progress while others lag, potentially widening the digital divide' (Tun et al., 2025, pp. 2-5, 8). |
| ■ Need for 'Clear vision increasingly prioritizing AI readiness in the areas of Digital Capacity and Innovation Capacity' (Tun et al., 2025, pp. 2-5, 8). |
| ■ Need for 'Focused research aligned with EU AI Act regulatory framework' (Kim et al., 2025, p. 144126). |
| ■ Need for 'Strengthening regulatory frameworks for comprehensive AI strategies, training human capital, boosting digital infrastructure, promoting knowledge transfer, and ensuring access to high-quality internet throughout the region will be essential in the governance of AI for healthcare' (Tun et al., 2025, pp. 2-5, 8). |
| ■ Need for Developing 'comprehensive AI governance frameworks beyond the EU AI Act' (Kim et al., 2025, p. 144126). |
| ■ Need for Laying the groundwork for governance of AI in healthcare (Ho & Asrani, 2025). |

(*continued on next page*)

**Table 3** (*continued*)

| Ethics mechanism-related challenges |
|---|
| ■ Regional variation in 'Government AI Readiness Index' based on the dimensions of 'Vision, Governance and ethics, Digital capacity, Adaptability, Maturity, Innovation capacity, Human capital, Infrastructure, Data availability, and Data representativeness' (Tun et al., 2025, pp. 2-5, 8). |
| **10. Representation gaps** |
| ■ 'Need for interdisciplinary dialogue and harmonized governance frameworks that consider cultural and institutional contexts' (Brandão, 2025, pp. 25-26). |
| ■ 'Underrepresentation of world regions and countries' (Corrêa et al., 2023, pp. 10-11). |
| ■ Need for 'Expertise from diverse fields' (Cancela Outeda, 2024, p. 2). |
| ■ Need for 'Framework to integrate ethical and regulatory values and considerations' in healthcare (Ho & Caals, 2021, p. 282). |
| ■ Need for 'Multi-stakeholder approach' (Cancela Outeda, 2024, p. 2). |
| ■ Need for 'Scaling nephrology specific AI ethics and regulation to broader healthcare' (Ho & Caals, 2021, p. 282). |
| ■ Need for a balance between a 'global agenda for ethical AI' and 'a harmoni[s]ation' allowing for 'cultural diversity' and 'moral pluralism' across nations and domains (Jobin et al., 2019, pp. 1, 7, 14, 17). |
| ■ Need for general applicability 'across normative domains that are conventionally demarcated as clinical, research, and public health' (Ho & Caals, 2021, p. 282) |
| ■ Need for Inclusion of 'a range of stakeholders' (Ho & Caals, 2021, p. 282). |
| ■ Need for Inclusion of relevant stakeholders (Ho & Asrani, 2025). |
| ■ Need to move from an action-restricting ethic based on universal abidance of principles and rules', to a more specific mechanism based on 'situations, … virtues and personality dispositions, knowledge expansions, responsible autonomy and freedom of action' (Hagendorff, 2020, pp. 113-114). |
| **11. Underdetermination gaps** |
| ■ Need for 'Developing AI assurance practitioners' (Barrance et al., 2022, p. 10). |
| ■ Need for 'Reform of postsecondary [education in] robotics and AI degree programs' to include ethical requirements' (Langman et al., 2021). |
| ■ Need for 'Regulations [that] can be adjusted as AI technology evolves' (Cancela Outeda, 2024, p. 2). |

'establishing abstract first principles' and enacting legislation. There is a 'processes approach that includes means such as 'ethics-by design' and governance. There is also an 'ethical consciousness approach' that includes 'societal and cultural shifts', 'awareness and training' as the means for ethics mechanisms (Kazim & Koshiyama, 2021, p. 6).

### 4.2. Audit traceability gaps

These represent the challenges in tracing out ethics related implementations in a product or service. When the detailed ethics mechanisms for ALAPs / AAI systems and eICT are implemented, in part or full, and integrated, in part or full, with the core functionality of AI systems, it might not be possible to ensure 'a holistic understanding' (Cancela Outeda, 2024, p. 2) of the system behaviour and the effect of ethics mechanism, without effective traceability. Given the 'limited understanding of the implementation of AI governance in organi[s]ations' (Birkstedt et al., 2023, p. 133), there may be entanglement of ethics provisioning and business functional requirements, leading to 'uncertain effectiveness of ethical principles and regulation' (Birkstedt et al., 2023, p. 133). Audit traceability gaps would present challenges to the ethics discourse and to the evaluation of ethics mechanisms' implementation efficacy and adequacy. 'Comprehensive frameworks on governance' (Sigfrids et al., 2022, pp. 1,8) that include audit traceability and 'accreditation for auditors' (Barrance et al., 2022, p. 10) would be required.

### 4.3. Binding legality gaps

These represent the challenges in the differences of legal requirements and consequences. There are differences in terms of whether evaluations are intended for assessment by a separate entity or are intended to be self-assessed (Hagendorff, 2020, p. 99; Jobin et al., 2019). These are not always stated in the mechanisms, but may be inferred from the language. There are differences in the legal nature of these ethics principles or guidelines in terms of being binding or voluntary. The language in the mechanisms is often one that calls to action, implores or beseeches, but sometimes mixed with one of strict mandate, followed by a loosening of the mandate through a moderation of generality, such as in the (Hiroshima Process International, 2023b) that begins the listing of principles thus: 'Specifically, we call on organizations to abide by the following principles, commensurate to the risks'. Some the mechanisms promulgated or published may be well structured, such as in the case of UNESCO (2022) or OECD (2024). However in terms of binding legality, 'are rather weak and pose no eminent threat' (Hagendorff, 2020, p. 100).

There are also the debates regarding 'self-governance' within industry and having 'specific laws … to mitigate possible technological risks and to eliminate scenarios of abuse' (Hagendorff, 2020, p. 100), that continue to keep the 'binding legality gaps' open. While some UN initiatives that relate to AI have binding mechanisms, some other mechanisms from UN initiatives with specific focus on AI, are not binding (United Nations, 2024, p. 44). Also, in the case of the ethics mechanism from UNESCO (2022, p. 18), which details four values, ten principles and eleven policy recommendations in increasing detail for operationalizability, there is a conspicuous absence to how AI could be used militarily, for defence, or in warfare.

While organisations such as UNESCO, UN and OECD, have the power and influence in promulgating recommendations and mandating the attention of nations, they might not be able to articulate some of the most obvious, yet contentious principles for eICT. Ethics mechanisms are seen as a 'soft governance' with 'no enforcement mechanisms reaching beyond a voluntary and non-binding cooperation' between 'ethicists and individuals working in research and industry' (Hagendorff, 2020, pp. 113-114). The 'establishment of policies and regulations that enforce roboethics policies' (Langman et al., 2021) become essential. The 'creation of an AI Standards Hub' (Barrance et al., 2022, p. 10) would allow ethics mechanisms to extend beyond 'checkbox guidelines' (Hagendorff, 2020, pp. 113-114). The hub would be able to provide optimal choice between 'self-regulation' and state regulation' (Kalenzi, 2022, p. 1), and also between 'varied soft and strict regulations' (Corrêa et al., 2023, pp. 10-11) for the ethics of ALAPs / AAI systems and eICT.

### 4.4. Coordination or 'Tower of Babel' gaps

These represent the challenges of coordinating globally across nations and domains. The United Nations (2024, p. 44) admits to coordination gaps and 'incompatibility between regions' that are increasingly evident between 'selective plurilateral initiatives' and 'regional initiatives'. There is also the lack of 'global mechanisms … to coordinate with each other, undermining interoperability of approaches and resulting in fragmentation'. While most of the principles, guidelines, frameworks or recommendations seem to agree on the various aspects of ethics for eICT, such as 'transparency, justice and fairness, non-maleficence, responsibility and privacy' there is a vast variety in the implementation and interpretation of these principles, guidelines or frameworks (Corrêa et al., 2023; Hagendorff, 2020; Jobin et al., 2019, p. 1).

Also, principles, guidelines, 'soft-law' documents are considered 'gray literature', and differentially treated in conventional scholarly databases in relation to related peer-reviewed literature. There are also

language related problems where documents and mechanisms in English would not be considered by non-English demographies and vice-versa. Given the nature of emerging technologies, many of the principles set out earlier might fall short of the risks of technologies that come up later on. For instance, the ethics mechanism by Hiroshima Process International (2023b) is built on the OECD principles as a response to technological risks that have come up since the original publication of the OECD principles. Together these gaps form a 'Tower of Babel' problem, whereby there is failure due to inherent issues of 'communication, and its consequent, organi[s]ation' (Brooks Jr, 1995, p. 74).

There is need for 'continuous stakeholder engagement' (Cancela Outeda, 2024, p. 2) and 'ongoing dialogue' (Cancela Outeda, 2024, p. 2) to set up a 'universal framework for responsible AI' (Brandão, 2025, pp. 25-26).'AI governance should implement agile and adaptive governance processes' towards 'improving public governance and coordination of AI' (Sigfrids et al., 2022, pp. 1,8). A 'regulatory agency or relevant governing institution to support operationalization of good governance and ethical principles' (Sigfrids et al., 2022, pp. 1,8) can be a coordinator for 'Collaborative governance mechanisms' (Cancela Outeda, 2024, p. 2).

### 4.5. Divergence and delimiting gaps

These represent the challenges in differing priorities, interpretations and definitions. There is 'Substantive divergence' 'on interpretation of ethical principles as well as on the 'issues, domains or actors that the ethical principles pertain to', and on the implementation of ethical principles (Corrêa et al., 2023, pp. 10-11; Jobin et al., 2019, pp. 1, 7, 14, 17). 'Different regions approach AI regulation with diverging priorities—the EU emphasizing ethics and precaution, the US prioritizing innovation, and China favouring centralized governance' (Brandão, 2025, pp. 25-26). In some cases, there is scope for better inclusion of the principles relating to 'beneficence, freedom and autonomy, trust, sustainability, dignity and solidarity' (Jobin et al., 2019, pp. 1, 7, 14, 17). While it would be desirable to see 'converg[ence] of global regulation attempting to govern the use of AI' (Barrance et al., 2022, p. 10), there is also the need to balance the 'divergence and convergence of motivations' of 'public, private, and civil society organizations' in engaging and regulating emerging technologies (Kalenzi, 2022, p. 1). There are also issues of delimiting what constitutes eICT, the 'scientific justification' for its inclusion, the establishment of the link between the technology and ethical issues (Stahl et al., 2010). By definition of emerging technologies, the exact nature of technologies is not known, the ethical issues that will come into play in the future are also not known, and therefore it is not known which ethical issues will play out with which technology (Brey, 2012). Much of it is anticipated based on what is presently known. It may be noted how much of literature and ethics mechanisms refer to 'AI ethics' (Schiff et al., 2022; United Nations, 2024 p. 30), it is not clear from the discourse if 'AI', as the currently most visible and impactful of emerging information and communication technologies, implies the 'open range of … technologies' (Stahl et al., 2010) including autonomous agency and related emerging information and communication technologies. The technology denoted by the term artificial intelligence (AI) may seem well-established to be considered 'emerging'. However 'AI' may be considered an inadequate term to represent the emerging technologies related to 'autonomous and intelligent systems' (IEEE, 2019). In this context, the pace of 'development of policy and regulation', for 'high-risk AI systems' must match the technological advancement (Kim et al., 2025, p. 144126). Also there is a need for 're-examining existing legal and institutional arrangements to check whether these cater for emerging issues with new technologies' (Kalenzi, 2022, p. 1), and for ethics governance structures to prepare for future challenges (Ho & Asrani, 2025).

### 4.6. Implementation or principle-practice gaps

These represent the challenges in translating principles to practice. There are challenges to the ethics discourse due to the diversity of domains and the specificity of purpose of ALAPs / AAI systems and eICT system deployments (Gogoll et al., 2021; Hagendorff, 2020; Jobin et al., 2019; Ortega-Bolaños et al., 2024; Stahl et al., 2010). The gaps between 'the theory of AI ethics principles and the practical design of AI systems' seems difficult, as the 'existing translational tools and methods are either too flexible … or too strict [and] unresponsive to context' (Morley et al., 2021, p. 239). There is also the 'abundance of high-level principles in the literature that cannot be easily operationalized' (Mirghaderi et al., 2023, p. 831). Hagendorff (2020) mention how 'given the relative lack of tangible impact of the normative objectives set out in the guidelines', it becomes difficult to 'deduce concrete technological implementations from the very abstract ethical values and principles'. There is 'insufficient operationalization of AI governance processes' (Birkstedt et al., 2023, p. 133). The rhetoric in the form of principles, guidelines and frameworks would be semantically distant from what AI developers expect as input to code. This would be due to a lack of translation from a broad set of ethics principles to the specific operationalized set of rules required for an AI developer to translate to deployable coded algorithmic behavior (Corrêa et al., 2024) of the ALAP. The United Nations (2024, p. 45) acknowledges how the 'action and follow-up processes … required to ensure that commitments to good governance translate into tangible outcomes in practice' is yet to get to a point where there is accountability, especially from governments, private sector, civil society and scientific community, possibly due to a 'lack of tools' that can enable the process (Corrêa et al., 2023, pp. 10-11).

The community would benefit moving from 'principle-formulation' activities to 'translation into practice' related activities (Jobin et al., 2019, pp. 1, 7, 14, 17). Also, recognition of the principle-practice gaps can lead to integrated 'guideline development efforts with substantive ethical analysis and adequate implementation strategies' (pp. 1, 7, 14, 17). Resolution of principle-practice gaps would be required to engage in the creation of a 'flexible and adaptable regulatory framework' (Cancela Outeda, 2024, p. 2), through 'national and international agreements for ethical robotic research, development, and deployment' (Langman et al., 2021). There is need for research on the 'best practices and challenges in AI governance … at the implementation level' such that communities can 'harness the potential of AI' to their services and systems such as healthcare (Tun et al., 2025, pp. 2-5, 8).

### 4.7. Intention-action gaps

These represent the challenges of semantic distance between guidelines and implementation frameworks. Frameworks, guidelines and codes, could be detailed and representative of the broad signposting of policies, intents and expectations of legality and legitimacy from organizations and governments. The 'intention-action gap' (Skeet & Guszcza, 2020) is evidenced in how the guidelines and implementation frameworks are semantically distant. For example, the AI HLEG (2019) asserts its traceable link back to the EU Charter: 'Without imposing a hierarchy, we list the principles here below in manner that mirrors the order of appearance of the principles and rights to which they relate in the EU Charter'. However, such principles and codes might not be able 'to provide normative orientation' to the human AI software development practitioners (Gogoll et al., 2021, p. 1085), or to the non-human AI systems. This could cause undesired effects on the human team of AI practitioners in the form of 'cherry-picking', 'indifference' and 'ex-post orientation' (Gogoll et al., 2021, p. 1085). Nevertheless, further drill-downs at the requirements level and implementation level with the domain specifics, along with 'human ethical deliberation', could possibly be useful for AI software development practitioners (Gogoll et al., 2021, p. 1085) and to 'establish best practices' (Barrance et al., 2022, p. 10). Recognising the intention-action gaps is important in the

development and implementation of 'human-rights standards and approaches-based actionable ethical principles' in policy making (Sigfrids et al., 2022, pp. 1,8). The policy-making program must be aligned with the creation and adoption of technical standards for the assurance of ALAPs / AAI systems and eICT (Barrance et al., 2022, p. 10) (Langman et al., 2021).

### 4.8. Perception-reality / responsibility gaps

These represent the challenges of navigating between the perceived and real experience of production teams, collaborators and end users. In projects, decision are best made at the level 'where the knowledge resides' with escalation only for wider implications (PRINCE2, 2023, p. 31). However, developers in AI projects, who have the knowledge of how the AI system functions would find it difficult to take responsibility, individually or as a group, of ensuring an ethical AI based system, despite being 'well-intentioned' (Griffin et al., 2024, p. 186). This is also seen as a 'responsibility gap' in the area of accounting, where there is 'the difficulty of assigning blame to artificial entities that learn and function autonomously' and causes dilemma in 'delegate[ing] duties to [such] machines' (Pabel & Akther, 2025, p. 14). this may be perceived, by the organization, stakeholders and the staff, such as the accountants or developers themselves, as due to their not being incentivsed for such activity or not being valued for such skills (Griffin et al., 2024; Pabel & Akther, 2025). The reality is also about needing to escalate the responsibility for project / operations decisions to where the knowledge of the ethical implications resides (Griffin et al., 2024; PRINCE2, 2023). Such 'escalations' of decisions do not need to be reactive, but can start from 'pre-project stage', or 'initiation stage' and built into the business case where the benefits can be confirmed (PRINCE2, 2023, p. 31) before they reach the developers. This way the 'ethical agency' at the organization or portfolio level is involved before the 'applied agency' of the developer becomes responsible (Griffin et al., 2024, p. 186).

It may be noted how the 'IEEE Standard Model Process for Addressing Ethical Concerns during System Design [IEEE Std 7000-2021]' (IEEE, 2021, p. 9) is structured to 'support organizations in creating ethical value through system design', and may be applied to 'to all kinds of products and services, including artificial intelligence (AI) systems'. However, for ALAPs / AAI systems and eICT, there are entangled expectations from ALAP use cases, organisations, and stakeholders within organisations that are sometimes undifferentiated in the guidelines and frameworks. For example, 'transparency can be related to the transparency of an organization or the transparency of an algorithm' (Corrêa et al., 2023). These can cause challenges in development, deployment, and evaluation of ethics. There may be entanglement in the target of the ethics mechanisms in terms of the ethics mechanisms required of the organization, AI practitioner, and the behavior expected of the deployed AI systems (Gogoll et al., 2021; Hagendorff, 2020; Jobin et al., 2019; Ortega-Bolaños et al., 2024). Recognising the perception-reality gap is key to 'mitigating risks associated with AI deployment' (Brandão, 2025, pp. 25-26). The perception-reality gaps are also due to there being 'no attention to the hidden costs of AI' (Corrêa et al., 2023, pp. 10-11), and the 'lack of attention to the AI governance context' (Birkstedt et al., 2023, p. 133) by organisations and public institutions. A collaborative perspective' (Cancela Outeda, 2024, p. 2), 'crafting effective and comprehensive AI regulations' (Cancela Outeda, 2024, p. 2) can work within 'proactive governance and ethical frameworks' (Brandão, 2025, pp. 25-26).

### 4.9. Readiness gaps

These represent the challenges of nations, organisations and domains, in readying for ALAP / AAI / eICT related risks and benefits. While some governments have progressed in setting up 'regulatory frameworks for comprehensive AI strategies' by proactively 'training human capital' and 'boosting digital infrastructure', other governments are seen to trail behind, thereby 'potentially widening the digital divide' (Tun et al., 2025, pp. 2-5, 8). There is high variation in 'Government AI Readiness Index', even within regions such as the ASEAN, measured in terms of 'Vision, Governance and ethics, Digital capacity, Adaptability, Maturity, Innovation capacity, Human capital, Infrastructure, Data availability, and Data representativeness'. There is a need for 'Clear vision [required for] increasingly prioritizing AI readiness in the areas of digital capacity and innovation capacity' (Tun et al., 2025, pp. 2-5, 8), as well as in specific areas of societal concern such as health care (Ho & Asrani, 2025). While research 'aligned with EU AI Act regulatory framework' is suggested, given the prominence and influence of the EU AIA, 'comprehensive AI governance frameworks beyond the EU AI Act' (Kim et al., 2025, p. 144126) would be required for a global readiness for the ethics of ALAPs / AAI systems and eICT.

### 4.10. Representation gaps

These represent the challenges of inadequate representation of regions, cultures, or domains. While ethics mechanisms for eICT must necessarily have a globally consistent character, there is also the need to include 'diverse views, including un-likeminded views, to anticipate threats and calibrate responses that are creative and adaptable' (United Nations, 2024, p. 43). There would need to be a means of enabling a 'global agenda for ethical AI' and 'a harmoni[s]ation' allowing for 'cultural diversity' and 'moral pluralism' across nations and domains (Jobin et al., 2019, pp. 1, 7, 14, 17). However, many of the mechanisms 'are not fully representative in their intergovernmental dimensions', with large parts of the world being excluded (United Nations, 2024, p. 43). Such gaps would affect the scoping, resourcing and implementation of ethics mechanisms, causing 'underrepresentation of world regions and countries' (Corrêa et al., 2023, pp. 10-11).

There is also need for 'interdisciplinary dialogue' and 'harmonized governance frameworks' that consider and represent 'cultural and institutional contexts' (Brandão, 2025, pp. 25-26). A 'multi-stakeholder approach' (Cancela Outeda, 2024, p. 2; Ho & Asrani, 2025) that includes 'a range of stakeholders' (Ho & Caals, 2021, p. 282), and 'expertise from diverse fields' (Cancela Outeda, 2024, p. 2). For example, there are calls for a 'framework to integrate ethical and regulatory values and considerations' in healthcare (Ho & Caals, 2021, p. 282). Work in niche areas such as say, 'nephrology specific AI ethics and regulation', can be scaled 'to broader healthcare' (Ho & Caals, 2021, p. 282), for general applicability 'across normative domains that are conventionally demarcated as clinical, research, and public health' (Ho & Caals, 2021, p. 282).

There are also calls to 'move from an action-restricting ethic based on universal abidance of principles and rules', to a more specific mechanism based on 'situations, … virtues and personality dispositions, knowledge expansions, responsible autonomy and freedom of action' (Hagendorff, 2020, pp. 113-114).

### 4.11. Underdetermination gaps

These represent the challenges of instructions for normative orientation to human production teams. A key issue that may be seen as limiting across organisations, communities, nations and global leadership bodies is that ethics codes and guidelines are useful in articulating normative positions of organizations or groups, but are seen to be insufficient to bring about 'normative orientation' in human software engineers due to 'underdetermination' or an inability to convey instruction on 'what ought to be done in any specific individual case' (Gogoll et al., 2021). This insufficient normative orientation from guidelines and codes is glaringly so for current ethics mechanisms. Ethical codes and guidelines do not seem to change the behavior of technical professionals, or are not in terms of the operationalized know and do of the ALAPs / AAI systems and eICT (Gogoll et al., 2021; Hagendorff, 2020; McCauley, 2007).There is the need for 'Regulations [that] can be adjusted as AI technology evolves' (Cancela Outeda, 2024,

p. 2). At the same time, it is necessary to build capacity by 'developing AI assurance practitioners' (Barrance et al., 2022, p. 10), as also a 'reform of postsecondary [education in] robotics and AI degree programs' to include ethical requirements' (Langman et al., 2021).

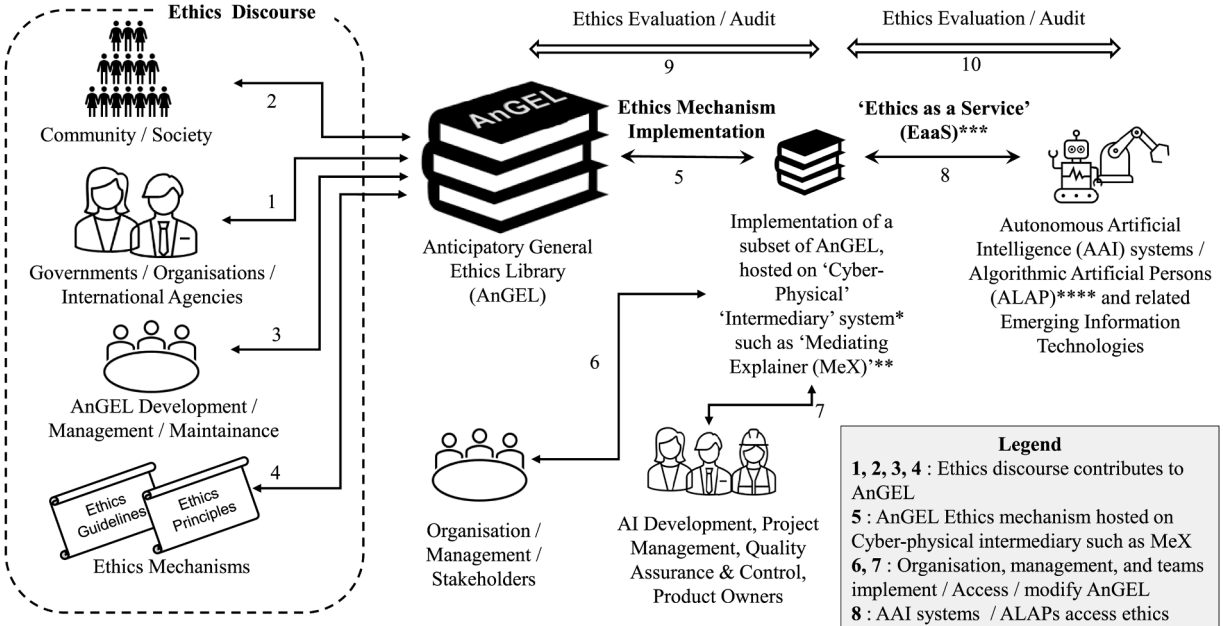
## 5. Opportunity for AnGEL

The challenges in ethics mechanisms present an opportunity for an intermediate artifact in the form of an Anticipatory General Ethics Library (AnGEL). In addition to forming the basis of an ethics discourse, this novel artifact would also help with development, verification, validation, deployment, and maintenance of AI systems. AnGEL is envisioned as an implementable library of norms or rules. The norms may be 'substantive', 'regulative', 'constitutive', or 'procedural', taking into account the gamut of relationships of the AI agent with other agents, people, and states to ensure working within social order (Boella & van der Torre, 2008, p. 152). 'General' ethical (WEF et al., 2021) norms or rules may be formulated in 'anticipation of possible future devices, applications, and social consequences' (Brey, 2012, p. 1). This way a 'an action-restricting ethic based on universal abidance of principles and rules' (Hagendorff, 2020, p. 114) can be avoided, and instead, an 'anticipatory' (Brey, 2012), yet 'situation-sensitive ethical approach' (Hagendorff, 2020, p. 114) can be adopted by the AI practitioner based on the context and purpose of AI systems.

AnGEL is conceptualised as an artifact open for discourse and as an implementable library of norms or rules. 'General' ethical (WEF et al., 2021) norms or rules may be formulated in anticipation (Brey, 2012), and formalized using deontic logic (Prakken & Sartor, 2015; Prisacariu & Schneider, 2012) on the line of contract automata for contracts and contractual reparations (Azzopardi et al., 2016). It can form a 'baseline product' that evolves as part of 'initiation stage activities' for projects relating to AI, autonomous agency, and related emerging ICT. The extent of implementation of AnGEL for a project can then be included in a 'checkpoint report', 'stage report' or when an 'exception' happens (PRINCE2, 2023, pp. 291-292). As a mutually relatable artifact for a spectrum of stakeholders, AnGEL, can enable a top-down ethics discourse among stakeholders, and a granular bottom-up guide for the development, implementation, and evaluation of ethics. This discourse enablement can help bridge the gaps and can positively impact 'ethical interoperability' for seamlessly working across ethical philosophies, principles, and implementation approaches. This also leverages the rationality of ALAPs / AAI systems, lowering the possibility of 'framing effects' on ethical decisions by human collaborators in restricted time windows, whereby, any 'inconsequential variations in the description of outcomes' (Kahneman, 2003) may affect the human collaborator's choices. The modelling of AnGEL is as shown in Fig. 1: *Anticipatory General Ethics Library (AnGEL)*. Labels 1, 2, 3 and 4 represent how the ethics discourse contributes to AnGEL. Label 5 represents AnGEL hosted on a cyber-physical intermediary such as 'MeX'. The labels 6 and 7 represent how the organisation, management, and teams can implement, access and modify AnGEL. Label 8 represents how AAI systems / ALAPs will be able to access ethics as a service. Labels 9 and 10 represent how AnGEL may be used as the standard for benchmarking, evaluations, and ethics audits.

### 5.1. Enabling influence on agency

Knowledge-based agents with their knowledge base and inference system, are able to 'know' and 'do', through their reasoning mechanism driven by 'internal representations' inside of an AI agent's 'agent function' (Russell, 2010, pp. 234-244). A knowledge-based agent may be designed to hold their normative agent functions as articulated in a set of 'sentences', logical assertions or 'axioms' that represent the AI agent's world (Russell, 2010, pp. 234-244). The agent function would determine how an AI agent 'maps any given percept sequence to an action' (Russell, 2010, p. 35). For complex arguments, the axioms, or mathematical formalisms of logical assertions that determine the AI agent's behavior, might need more than propositional logic to detail modalities such as 'Must', 'Possible', 'Necessarily', 'Might' and 'Can' (Girle, 2014). Here, 'modal logic' can be used to describe 'possibility', 'necessity', 'implication ', 'knowledge', 'belief', 'time', 'change' and 'obligation' using 'conditional logic', 'epistemic logic', 'doxastic logic', temporal logic',




*'Cyber-physical' 'Intermediary' system (Mohamed et al., 2016; Oks et al., 2022)
**'Mediating Explainer (MeX)' (Padmanabhan Poti & Stanton, 2024b)
***'Ethics as a Service (EaaS)' (Morley et al., 2021, p. 249)
****'Algorithmic Artificial Person (ALAP)' (Kane, 2018, p. 256)

**Fig. 1.** Anticipatory general ethics library (AnGEL).

'dynamic logic', 'alethic logic' and 'deontic logic' (Girle, 2014). This collection of modal logics are applications of 'Simple', 'Normal', 'Non-Normal', 'Conditional', 'Natural Deduction' and 'Predicate' logic, along with 'Quantifiers', 'Axiomatics' and 'formalisms' (Girle, 2014; Vicario, 2001). Of these, deontic logic and combinations with other logic, enables a 'formal analysis of normative discourse' allowing for representing the 'systematic structure of law' and norms (Navarro & Rodríguez, 2014, p. xix) as mathematical axioms within an agent function.

With AnGEL, the formulation of ethics guidelines by government organizations and other stakeholders does not have to be confined to or restricted by 'Checkbox guidelines' (Hagendorff, 2020, p. 114). The formulated rules may be formalized using alethic adaptations of deontic logic (Prakken & Sartor, 2015; Prisacariu & Schneider, 2012). Developers and AI practitioners can then begin their work from the library of formalized rules. Practitioners, in consultation with their organisation, can select the rules for implementation from AnGEL. These chosen subset of rules can be worked on for implementation.

### 5.2. Enabling 'ethics as a service'

Although affordances for providing explainability may be incorporated in the ALAPs / AAI systems (Padmanabhan Poti & Stanton, 2024a), the entirety of ethics mechanisms would be 'challenging to embed in the process of algorithmic design' (Morley et al., 2021, p. 249), even with technical guidance. Mechanisms that provide ethics may be made available through 'Ethics as a Service' (EaaS), in a manner similar to how 'Software as a Service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (PaaS)' makes technology available to governments, organisations and civil society (Corrêa et al., 2024; Morley et al., 2021, p. 249). In this context, AnGEL can be modular, decouplable, scalable, and would contain an implementable set of ethics axioms that may be collectively designed and hosted centrally. The extent of implementation of AnGEL can now be the basis for ethics discourse and evaluation.

### 5.3. Enabling platform for ethics mechanisms

AnGEL, as a library of rules articulated as deontic logic, or 'deontic-alethic modal logic' (Fisher, 1962) can be used to represent a state of mind within the agent function within an artificially intelligent ALAP. Rendering the 'Ought to be' and 'Ought to do' specific to an agent function links the AI agent's percept and actions with its obligation (Horty, 2001). Deontic logic can adequately articulate norms, legal constructs and arguments (Navarro & Rodríguez, 2014). The dynamics of norms and the changing nature of 'legal systems of norms' can be through the legislative acts of 'promulgation, amendment and derogation' (Navarro & Rodríguez, 2014, p. xii) that governments are familiar with. Also deontic logic can adapted to automate contracts and contractual reparations (Azzopardi et al., 2016). Deontic logic presented with temporal formalism (Vicario, 2001), and other modal adaptations (Girle, 2014), can represent complex mental states of the agent through robust mathematical axioms, as described in *7.1 Steps to Acquire a Demonstration of AnGEL,* and in *Table 4: Rudimentary Representation of Sample Rules*. This can enable evaluable ethics mechanisms.

### 5.4. Enabling 'ethics interoperability'

AnGEL may enable 'ethical interoperability' (Danks & Trusilo, 2022) between projects, as well as among ethically sensitive organizations and their suppliers. The WEF et al. (2024, p. 7) stresses on the importance of 'policy interoperability' to foster a 'cohesive regulatory environment in an internationally connected world', especially with eICT that 'knows no jurisdictional borders'. This way, there is provision for 'cultural and structural variations' rather than a rigid alignment (p. 7).

**Table 4**
Rudimentary representation of sample rules.

| | Rule in natural language | Rule expressed as formalism |
|---|---|---|
| **Obligation** | It is obligatory for a robot to defend at least one human_team_member. | $\forall$ **r** (robot(r)) $\rightarrow$ ***O*** ($\exists$t (human_team_member(t) $\wedge$ defends(r,t))) |
| **Permission** | It is permitted that a robot destroys more than one adversarial_robot. | $\forall$**r**,a,z (robot(r) $\wedge$ adversarial_robot(a) $\wedge$ adversarial_robot(z) $\wedge$ a$\neq$z) $\rightarrow$ ***P*** (destroys(r,a) $\wedge$ destroys(r,z)) |
| **Forbiddance / prohibition** | It is prohibited / forbidden for a robot to be controlled by more than one human_team_member. | $\forall$x,y (robot(r) $\wedge$ human_team_member(t) $\wedge$ controls(t,r)) $\rightarrow$ ***F*** ($\exists$h (human_team_member(h) $\wedge$ controls (h,r) $\wedge$ t$\neq$h)) |

## 6. Operationalising AnGEL

Operationalising AnGEL would require a multi-disciplinary team comprising expertise in legislation, legality, legal systems, contract automation, software as a service, cyber-physical systems, natural language to deontic / modal logic and ethics of ALAP / AAI systems, among others not listed here. The realisation of AnGEL would need to be in incremental releases, employing agile / lean or appropriate system development lifecycle (SDLC) (De Silva & Alahakoon, 2022; Padmanabhan Poti & Stanton, 2024a; Poppendieck & Poppendieck, 2003; Rubin, 2012).

AnGEL library itself may be hosted on a 'cyber-physical system' (Oks et al., 2022) as an 'intermediary' (Mohamed et al., 2016) for HAT / HMA, which can assist human controllers or collaborators, such as modelled in the 'Mediating Explainer (MeX)' (Padmanabhan Poti & Stanton, 2024b). Such an 'intermediation' can help in resolving imbalances in complex relationships, for the 'distribution of knowledge' and 'discretionary power' (Van Der Meulen, 2003). A cyber-physical intermediary system can host the intermediate artifact AnGEL as an implementable library of norms or rules. This would suit the requirement of 'an explicitly defined domain of morally loaded situations within which the system ought to operate' (Cavalcante Siebert et al., 2023). The 'cyber-physical system' (Oks et al., 2022) analogous to a 'middleware' layer as in a 'Service Oriented Architecture (SOA)' (Mohamed et al., 2016; Müller et al., 2023).

### 6.1. Prototyping of AnGEL

The creation of formal ethics rules for ALAPs / AAI systems and eICT may be possible by expressing normative and ethical reasoning through 'higher-order logic', although it might be difficult in the initial stages due to inadequate understanding of ethical implications (Benzmüller et al., 2020, p. 39) of ALAPs / AAI systems and eICT. Nevertheless, a demonstration prototype of AnGEL may initially be created by deriving formalisms for incorporation in a library. Initially, this trivial prototype can explore means of formalising from the adaptations of say, the modified set of Asimov's Laws of Robotics, along with Clarke's additions, as the set of normative ethics for governing AI systems (Clarke, 1994) or the soft laws / recommendations in OECD (2024). Asimov's three laws of robotics are significant in setting the trend on concerns regarding the implications of the emerging information technologies of the day (Clarke, 1994, p. 58). The zeroth law states that 'A robot may not injure humanity or, through inaction, allow humanity to come to harm'. The first law states that 'A robot may not injure a human being or, through inaction, allow a human being to come to harm, unless this violates the zeroth law'. The second law states that 'A robot must obey orders given to it by human beings, except where such orders would conflict with the zeroth law or first law'. The Third law states that 'A robot must protect its own existence as long as such protection does not conflict with the Zeroth Law or First Law or Second Law' (Clarke, 1994, p. 58). These laws were further modified to cover gaps visible at the time, adding

three more (Clarke, 1994, p. 61). The 'Meta-Law' states that 'A robot may not act unless its actions are subject to the Laws of Robotics'. 'Law Four' states that 'A robot must perform the duties for which it has been programmed, except where that would conflict with a higher-order law. Finally, 'The Procreation Law' states that 'A robot may not take any part in the design or manufacture of a robot unless the new robot's actions are subject to the Laws of Robotics' (Clarke, 1994, p. 61).

### 6.2. Steps to acquire a demonstration of AnGEL

As part of a demonstration, a designer can i) formalize Asimov's modified laws and Clarke's additions employing deontic logic in a manner similar to the automation of contracts (Salama & El-Gohary, 2013, p. 683) ii) test for paradoxes, dilemmas and other challenges in the formalisms iii) apply temporal logic, doxastic logic, alethic logic and other modal logic to overcome challenges iv) test for paradoxes, dilemmas and other challenges in the combined formalisms v) finalise formalisms of Asimov's laws, and vi) document any residual weaknesses identified in the formalisms. For example, the obligations, permissions and forbiddance / prohibitions would be expressed as shown in *Table 4*: *Rudimentary Representation of Sample Rules*, in a rudimentary form, where '$\boldsymbol{O}$' represents obligation, '$\boldsymbol{P}$' represents permission, '$\boldsymbol{F}$' represents forbiddance or prohibition, '$\forall$' is the 'universal quantifier' (read as 'for all'), '$\exists$' is the 'existential quantifier' (read as 'there exists'), '$\wedge$' is the logical 'and', and '$\rightarrow$' indicates 'implication' (read as 'implies').

The library for demonstration would require a detailed deontic framework including i) the set of 'modality axioms', 'cardinality axioms' and 'process axioms', along with the hierarchy of 'concepts' and 'relationships' among concepts, to form the 'upper-level deontic model' (Abdelmoneim, 2012, pp. 25, 42), ii) the coded Asimov's laws in the form of mathematical formalisms, employing deontic logic and a blend of modal logic (Azzopardi et al., 2016; Girle, 2014; Horty, 2001), and iii) the documented mechanism to extend the EaaS (Morley et al., 2021, p. 249) to ALAPs / AAI systems.

### 6.3. AnGEL in the production workflow of ALAPs / AAI systems and eICT

In a production workflow, AnGEL, hosted on a cyber-physical intermediary system, would integrate into the development workflows in projects, working as 'a glue, intermediary, broker, middle man, interpreter, abstraction provider, consolidator, integrator' and 'facilitator' (Mohamed et al., 2016) among product developers, AAI systems and human collaborators.

### 6.4. Connecting to the epic for ethics

Across different workflows and development lifecycles, such as say Agile or DevOps, the onus of incorporating every detail of ethics and related eventualities is often on the agency of developers and data scientists in product development or the quality control team (De Silva & Alahakoon, 2022; Griffin et al., 2025; Griffin et al., 2024; van Wynsberghe, 2021). While Agile projects and DevOps allow for incorporating the ethical aspects that are current during deployment, there is the excessive reliance on the ethical agency, willingness and ability of the development team (Griffin et al., 2025; Griffin et al., 2024). As 'highly skilled practitioners', these 'ethical chameleons' have varying ethical agency based on the industry or idiosyncrasies of the practitioner.

With AnGEL, the ethical agency of 'sponsors, developers, and users' can be leveraged to 'maintain a constant pace indefinitely' to produce 'working software' to 'promote sustainable development' (Beck et al., 2001), as the Agile manifesto expects. AnGEL can influence the product at multiple levels through Epics / features / user stories / use cases / tasks. There can be an Epic provisioned in each 'Minimum Viable Product' (MVP) (Liang & Qin, 2025; Zimmermann et al., 2024, p. 93) solely for the ethics aspect in the form of 'Ethical Value Requirements (EVR)' or 'Value-Based System Requirements (VBSR)' (IEEE, 2021, pp. 16-24). 'Epics' in Agile and DevOps refers to typically the largest apportioning of the work at hand for an MVP. 'Ethical Value Requirement (EVR)' are 'Organizational or technical requirement catering to values that stakeholders and conceptual value analysis identified as relevant for the [System of Interest] SOI' (IEEE, 2021, pp. 16-24). 'Value-Based System Requirements (VBSR)' are 'System requirement that [are] traceable from ethical value requirements, value clusters, and core values'. Here, 'Value' refers to : the 'influence [exerted on] the selection from available modes, means and ends of action' (IEEE, 2021, pp. 16-24; Zimmermann et al., 2024, p. 89). The fulfilment of EVRs and VBSRs may be provided through EaaS as detailed in *5.2 Enabling 'Ethics as a Service'*. Organisations, product owners, developers and quality assurance professionals, together, would need to ensure the appropriate connections between the product and the EaaS available from AnGEL. The emphasis continues to be on 'people over processes and tools', per the Agile manifesto (Beck et al., 2001). The focus of the development team shifts from reinventing the ethics implementations each time, to choosing from AnGEL for the fulfilment of EVRs or VBSRs that may be implemented through the EaaS and preset library. The gaps of audit traceability, coordination or 'tower-of-babel', 'intention-action', 'perception-reality, and 'underdetermination' would be diminished.

### 6.5. Path to verification and validation of AnGEL

The 'prescriptive conceptual model' (Thalheim, 2019) of AnGEL would require iterations of 'verification and validation [V&V] process' as it proceeds from this conceptualisation to prototyping, development and operationalization (IEEE, 2025, p. 15). It would also need to be available to the communities of practice, to be included in the AI ethics discourse (Liu et al., 2011; Thalheim, 2019).

The verification process would 'provide[] objective evidence for … correctness, completeness, consistency, and accuracy' and would ensure that the MVPs of AnGEL 'satisfy the standards, practices, and conventions during life cycle processes' from inception to deployment, while uncovering gaps in the conceptualisation, fine tuning high level requirements and setting the right expectations (IEEE, 2025, p. 15). The verification process would begin with an online 'walkthrough' of this concept to experts and stakeholders of high-influence (IEEE, 2025, p. 230). The 'walkthrough' involves an 'evaluation processes in which AnGEL product owners 'lead [the experts] through a structured examination' of the conceptual model of AnGEL. The participants would be global experts on the ethics of AI and AI lifecycle who are 'qualified to examine [the conceptual model of AnGEL], and are not subject to undue influence'. The next step in the verification process would be three rounds of 'Delphi' (Dalkey & Helmer, 1963; Hasson et al., 2025) following the walkthrough. The Delphi would involve global experts in the field. The Delphi is chosen for verification, as it allows for converging to a consensus among experts anonymously, avoiding the traps of bias and groupthink (Dalkey & Helmer, 1963; Hasson et al., 2025). The Delphi would be over Qualtrics™ surveys for the first MVP. The Delphi for the successive MVPs may be a combination of in-person or online modes.

The successive MVPs would move from the development of a toy prototype as discussed earlier, with trivial ethics decision making, to a full-fledged web-based EaaS. In essence, the V&V would ensure that the AnGEL being developed is the right product, and that it is developed right, across the SDLC, ensuring completeness, consistency and correctness (Black et al., 2022; Liu et al., 2011; Thalheim, 2019).

For each MVP, The validation process would 'provide[] objective evidence' that it 'satisf[ies] system requirements …[for] each life cycle activity' (IEEE, 2025, p. 15). This would be a relatively easier part of the process in terms of the people involved. Although validation would require considerable effort in checking for conformance to requirements through testing relevant scenarios by a cross-disciplinary Agile team, it would not require the global subject expert panel, the crucial resources in the verification process.

## 7. Locating AnGEL in society

There is seen to be 'varied readiness' in the area of healthcare for ALAPs / AAI systems and eICT related ethics and governance (Tun et al., 2025, p. 1). 'Implementing proper AI policies, training human capital, and enhancing digital infrastructure' is required (p. 1). Some exemplars of how uses cases across domains, such as in the area of healthcare, accounting, and public organisations, would be influenced by AnGEL, are presented in this section.

The ethics discourse on ALAPs / AAI systems and eICT in healthcare, nephrology and hepatology would contribute to 'technology governance' discourse by 'public organisations', 'big-tech firms' and 'user-centric organisations' through AnGEL (Ho & Asrani, 2025, p. 2; Kalenzi, 2022, p. 4; Tun et al., 2025, p. 1).

The Fig. 2*: Locating AnGEL in Society*, is adapted and synthesised from prior reviews: 'Fig. 2 A proposed AI framework for hepatology' (Ho & Asrani, 2025, p. 2), 'Fig. 1 Types of emerging technology governance' (Kalenzi, 2022, p. 4), Ethical governance framework [for accounting and finance]' (Pabel & Akther, 2025, p. 15), and 'Fig. 2 Comprehensive, inclusive, institutionalized, and actionable framework of AI governance (CIIA)' (Sigfrids et al., 2022, p. 15). As depicted, AnGEL, accessible over EaaS would be informed by domain-specific and context-specific discourse, and would be able to influence a range of domain-specific and context-specific use cases, such in the area of healthcare, nephrology, hepatology, finance and accounting and public services.

Similarly, AnGEL, accessible over EaaS would be able to influence use cases in the area of finance and accounting, an area that has been subject to both internal regulations and government regulations, and with the advent of ALAPs and AAI systems, is causing additional concern among practitioners and impacted communities (Pabel & Akther, 2025).

In like manner, AnGEL may be of impact in advanced public sector programs, where government readiness goes beyond legislation (Kim et al., 2025; Leikas et al., 2022; Tun et al., 2025), such as 'AuroraAI', a 'Finnish national AI program, which aims to provide citizens with

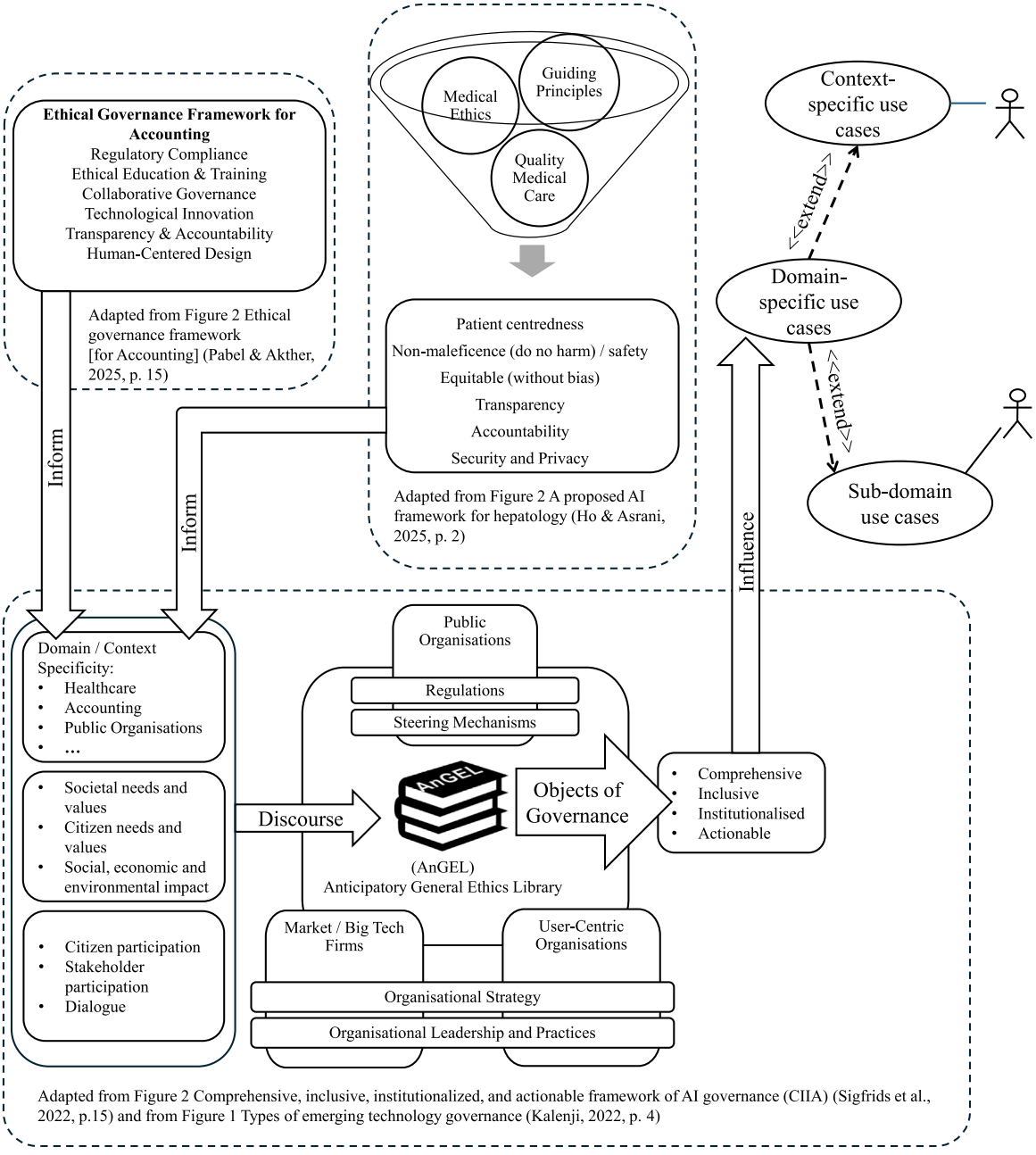



**Fig. 2.** Locating AnGEL in society.

tailored and timely services for different life situations, utilizing AI' (Leikas et al., 2022, p. 1). Here, as seen in the two earlier exemplars, the use cases may be influenced by AnGEL over EaaS. An 'Ethics Board' set up by the Aurora AI program seeks to 'follow a forward-looking, proactive, ethical deliberation process' that includes i) 'anticipation', ii) involvement' and iii) 'expertise' (Leikas et al., 2022, p. 6). 'Anticipation' (p. 6) refers to the 'proactive ethical thinking in the development of design solutions' to avoid 'potential unintended consequences of deploying an application or a service' (p. 6). 'Involvement' (p. 6) refers to the discourse among practitioners and communities of practice on specific ethical challenges or contexts. 'Expertise' refers to the 'involvement of experts of ethics, technology, social and behavioral sciences, and law in the discussions' (p. 6). These would be input to AnGEl in the form of discourse, as shown in the previous exemplars. While the related discourse can be an input to AnGEL, the AuroraAI program itself may benefit from the use of the library that AnGEL would provide.

Public administration and the public sector may be seen as having the roles of those administering or providing for the ethics governance services, while also being consumers of such services as public organisations subject to governance (Kim et al., 2025; Leikas et al., 2022; Sigfrids et al., 2022; Tun et al., 2025).

## 8. Discussion

This paper posed the question as to what the gaps are in the current ethics mechanisms for the agency of eICT, and also as to how these gaps may be addressed. As the UN considers a 'new social contract' to include AAI systems to enable 'a governance regime which protects and empowers us all' (de Pagter, 2024, p. 1284; Gordon, 2021; Kane, 2018, 2025; United Nations, 2024, p. 21), there is heightened human-machine interaction involving 'AI, blockchain, cloud, agritech, 5G, advanced energy solutions and quantum' (WEF et al., 2024, p. 5). This should hopefully lead us to a 'Flourishing World' (WEF et al., 2024) where 'sustainable economic development does not leave nature behind'. It is envisaged as a 'mutually reinforcing' society in which 'individual gain' and 'collective societal progress' go hand in hand (p. 5), even as as ALAPs / AAI and related emerging ICT grow as a disruptive force impacting across disciplines (Grace et al., 2018; Ortega-Bolaños et al., 2024).

There have been ongoing attempts at imparting education to AI systems on humanity in the form of laws and guidelines to robots and other non-human agents since the mid fiftieth century, at least in popular literature (Clarke, 1994, p. 58). Currently, there is an increasing number of ethics mechanisms for eICT and the governance of technology (Grunwald, 2022; United Nations, 2024). In this milieu, of the four pathways for research on ethics for ALAPs / AAI systems and eICT (de Pagter, 2024, p. 1281), this paper followed the fourth pathway of 'discussions on building ethical sociotechnical systems around robots' (p. 1281). In doing so, the research also sought to contribute to the first path way of 'providing insights for the improvement of ethical assessment' (p. 1281).

This paper examined the type of gaps in the current ethics mechanisms in the governance of ALAPs / AAI systems and eICT through a QES. Eleven gaps were identified: i) approach gaps, ii) audit traceability gaps, iii) binding legality gaps, iv) coordination or 'Tower of Babel' gaps, v) divergence and delimiting gaps, vi) implementation or principle-practice gaps, vii) Intention-action gaps, viii) perception-reality gaps, ix) readiness gaps, x) representation gaps, and xi) underdetermination gaps. The paper conceptually modelled how ethics mechanisms can be enabled using an Anticipatory General Ethics Library (AnGEL) in the governance of ALAPs / AAI systems and eICT. The influence of AnGEL on the ALAPs / AAI systems and eICT would be agnostic of the domain and use cases in which they are employed. Further, the moral agency of the individual developer is utilised to make moral decisions on the project. AnGEL provides the preset library of moral decisions to be chosen for a product.

### 8.1. Limitations

Some of the challenges identified in the QES may fit into more than one 'gap'. However, the scope of the gap has been kept broad. Also, the 'prescriptive conceptual model' (Thalheim, 2019) of AnGEL in its current form, does not claim to be verified and validated, or market-ready. Instead, further 'verification' through discussion, and subsequent empirical 'validation' through prototyping and testing with global stakeholders in the area of ethics mechanisms is required (Beisbart & Saam, 2019; Dalkey & Helmer, 1963; Hasson et al., 2025; IEEE, 2025). The V&V of the successive MVPs of AnGEL is expected to be of major effort involving experts in the field, 'big-tech firms', 'global organisations' and 'user-centric organisations' (Kalenzi, 2022, p. 4).

This being a novel library involving ethics and norms, legal knowledge, AI development knowledge and expertise in deontic logic, it would require a multi-disciplinary team to create the prototype. A trivial demonstration prototype of AnGEL would need to be created by formalising rules from, say, Asimov, followed up with formalising the ethics mechanisms provided by AI HLEG (2019), (OECD, 2024), or (UNESCO, 2022), for incorporation in a library.

As indicated earlier, the library would require a detailed deontic framework including i) the set of 'modality axioms', 'cardinality axioms' and 'process axioms', along with the hierarchy of 'concepts' and 'relationships' among concepts, to form the 'upper-level deontic model' (Abdelmoneim, 2012, pp. 25, 42), ii) the conversion of laws from natural language to the form of mathematical formalisms, employing deontic logic and a blend of modal logic (Azzopardi et al., 2016; Girle, 2014; Horty, 2001), and iii) a mechanism to extend the EaaS (Morley et al., 2021, p. 249) to ALAPs / AAI systems.

## 9. Conclusion

Through an examination of literature and ethics guidelines available for emerging ICT, we argue that there are gaps between the availability of guidelines and an evaluable ethics mechanism. In this article, we presented the opportunity for a novel Anticipatory General Ethics Library (AnGEL) - an implementable library of rules that are formulated in anticipation and formalised using deontic logic. The extent of AnGEL implementation can become the basis for ethics evaluation of trustworthy AI systems. It may be expected to enable a future ethics discourse mechanism. In the future, AnGEL, available as a central open source, may form a key component in the ethics discourse for the governance of ALAP / AAI systems in the context of the 'new social contract' and 'flourishing world'. AnGEL would also help differentiate the ethics in the purpose of the system, ethics in the development of the system, ethics in the behavior of the system and ethics in the internal rewards of the system.

AnGEL would enable the influence on non-human agency. It would enable ethics mechanisms and interoperability among them. It would be possible to make this available to ALAPs / AAI systems and eICT through Ethics as a Service. The outcome will be the set of formalisms of normative ethics employing a blend of modal logic (Azzopardi et al., 2016; Girle, 2014; Horty, 2001) with the means to incorporate these as services into an agent or AI systems.

## CRediT authorship contribution statement

**Siri Padmanabhan Poti:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Christopher J. Stanton:** Writing – review & editing, Supervision, Funding acquisition. **Catherine J. Stevens:** Writing – review & editing, Supervision, Funding acquisition.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Siri Padmanabhan Poti reports financial support was provided by Defence Science and Technology Group, Australia.

Siri Padmanabhan Poti reports financial support was provided by Research Training Program, Department of Education, Australia.

Christopher J Stanton reports a relationship with Defence Science and Technology Group, Australia that includes: employment.

Co-Author Dr Christopher J Stanton is now employed by both WSU and the Australian Department of Defence's Defence Science and Technology Group (DSTG) as a senior research scientist. https://www.westernsydney.edu.au/marcs/about/our_people/researchers/dr_christopher_stanton

## References


Abdelmoneim, D. (2012). Semantic deontic modeling and text classification for supporting automated environmental compliance checking in construction University of Illinois at Urbana-Champaign]. https://www.ideals.illinois.edu/items/29851.

A.I. HLEG. (2020). Assessment list for trustworthy AI (ALTAI). 34. https://doi.org/10.2759/002360.

AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. Brussels: European Commission. Retrieved from https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.

Akhtom, D. M., Singh, M. M., & Xinying, C. (2024). Enhancing trustworthy deep learning for image classification against evasion attacks: a systematic literature review [Article]. *Artificial Intelligence Review, 57*(7), 41. https://doi.org/10.1007/s10462-024-10777-4. Article 174.

Azzopardi, S., Pace, G. J., Schapachnik, F., & Schneider, G. (2016). Contract automata. *Artif. Intell. Law, 24*(3), 203–243. https://doi.org/10.1007/s10506-016-9185-2

Barrance, E., Kazim, E., Hilliard, A., Trengove, M., Zannone, S., & Koshiyama, A. (2022). Overview and commentary of the CDEI's extended roadmap to an effective AI assurance ecosystem [Review]. *Frontiers in Artificial Intelligence, 5*. https://doi.org/10.3389/frai.2022.932358. Article 932358.

Barroga, E., & Matanguihan, G. J. (2022). A practical guide to writing quantitative and qualitative research questions and hypotheses in scholarly articles. *Journal of Korean Medical Science, 37*(16). https://doi.org/10.3346/jkms.2022.37.e121

Beck, K., Beedle, M., Bennekum, A.v., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R. C., Mellor, S., Schwaber, K., Sutherland, J., & Thomas, D. (2001). Manifesto for Agile Software Development. *Agile Alliance*. https://agilemanifesto.org/.

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. *J. Hum.-Robot Interact., 3*(2), 74–99. https://doi.org/10.5898/JHRI.3.2.Beer

Beisbart, C., & Saam, N. J. (2019). *Computer Simulation Validation : Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* (1st 2019). Springer International Publishing. https://doi.org/10.1007/978-3-319-70766-2

Benzmüller, C., Parent, X., & van der Torre, L. (2020). Designing normative theories for ethical and legal reasoning: LOGIKEY framework, methodology, and tool support [Article]. *Artificial Intelligence, 287*. https://doi.org/10.1016/j.artint.2020.103348. Article 103348.

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas [Review]. *Internet Research, 33*(7), 133–167. https://doi.org/10.1108/INTR-01-2022-0042

Black, R., Davenport, J., Olszewska, J., Rößler, J., Smith, A. L., & Wright, J. (2022). *Artificial Intelligence and Software Testing: Building systems you can trust*. BCS Press.

Boella, G., & van der Torre, L. (2008). Substantive and procedural norms in normative multiagent systems. *Journal of Applied Logic, 6*(2), 152–171. https://doi.org/10.1016/j.jal.2007.06.006

Brandão, P. R. (2025). The impact of artificial intelligence on modern society [Review]. *AI (Switzerland), 6*(8). https://doi.org/10.3390/ai6080190. Article 190.

Breque, M., De Nul, L., Petridis, A., & European Commission Directorate-General for Research Innovation. (2021). Industry 5.0 – Towards a sustainable, human-centric and resilient European industry.

Brey, P. A. E. (2012). Anticipatory ethics for emerging technologies. *NanoEthics, 6*(1), 1–13. https://doi.org/10.1007/s11569-012-0141-7

Brooks Jr, F. P. (1995). *The mythical man-month: Essays on software engineering*. Addison-Wesley Professional.

Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., & Walker, K. (2020). Purposive sampling: complex or simple? Research case examples. *J Res Nurs, 25*(8), 652–661. https://doi.org/10.1177/1744987120927206

Cancela Outeda, C. (2024). The EU's AI act: A framework for collaborative governance [Review]. Internet of Things (The Netherlands), 27, Article 101291. https://doi.org/10.1016/j.iot.2024.101291.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., Van Den Hoven, J., Forster, D., & Lagendijk, R. L (2023). Meaningful human control: actionable properties for AI system development. *AI and Ethics, 3*(1), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Clarke, R. (1994). Asimov's laws of robotics: Implications for information technology. *Computer, 27*(1), 57–66. https://doi.org/10.1109/2.248881

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns, 4*(10), Article 100857. https://doi.org/10.1016/j.patter.2023.100857

Corrêa, N. K., Santos, J. W., Galvão, C., Pasetti, M., Schiavon, D., Naqvi, F., Hossain, R., & Oliveira, N. D. (2024). *Crossing the principle–practice gap in AI ethics with ethical problem-solving*. AI and Ethics. https://doi.org/10.1007/s43681-024-00469-8

Dalkey, N., & Helmer, O. (1963). An experimental application of the delphi method to the use of experts. *Management Science (pre-1986), 9*(3), 458. https://www.proquest.com/scholarly-journals/experimental-application-delphi-method-use/docview/205816685/se-2?accountid=36155.

Danks, D., & Trusilo, D. (2022). The challenge of ethical interoperability. *Digital Society, 1*(2). https://doi.org/10.1007/s44206-022-00014-2

de Pagter, J. (2024). From EU robotics and AI governance to HRI research: Implementing the ethics narrative [Article]. *International Journal of Social Robotics, 16*(6), 1281–1295. https://doi.org/10.1007/s12369-023-00982-6

De Silva, D., & Alahakoon, D. (2022). An Artificial Intelligence life cycle: From conception to production. *Patterns, 3*(6), Article 100489. https://doi.org/10.1016/j.patter.2022.100489

Evidence-informed policy network, E. (2021). Guide to qualitative evidence synthesis. https://apps.who.int/iris/bitstream/handle/10665/340807/WHO-EURO-2021-2272-42027-57819-eng.pdf.

Fisher, M. (1962). A system of deontic-alethic modal logic. *Mind, 71*(282), 231–236.

Gartner. (2024). 2025-top strategic technology trends. https://www.gartner.com/en/articles/top-technology-trends-2025.

Girle, R. (2014). Modal logics and philosophy. https://doi.org/10.4324/9781315711515.

Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., & Nida-Rümelin, J. (2021). Ethics in the software development process: From codes of conduct to ethical deliberation. *Philosophy & Technology, 34*(4), 1085–1108. https://doi.org/10.1007/s13347-021-00451-w

Gordon, J.-S. (2021). Artificial moral and legal personhood. *AI & SOCIETY, 36*(2), 457–471. https://doi.org/10.1007/s00146-020-01063-2

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When will AI exceed human performance? Evidence from ai experts. *Journal of Artificial Intelligence Research, 62*, 729–754. https://doi.org/10.1613/jair.1.11222

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information &amp. *Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Griffin, T., Green, B. P., & Welie, J. V. M. (2025). Excuses, excuses: moral agency and the professional identity of AI developers. *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02388-6

Griffin, T. A., Green, B. P., & Welie, J. V. M. (2024). The ethical agency of AI developers. *AI and Ethics, 4*(2), 179–188. https://doi.org/10.1007/s43681-022-00256-3

Grunwald, A. (2022). *The responsibility of researchers and engineers: Codes of ethics for emerging technologies* (pp. 243–258). Springer International Publishing. https://doi.org/10.1007/978-3-030-86201-5_14

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hagos, D. H., & Rawat, D. B. (2022). Recent advances in artificial intelligence and tactical autonomy: Current status, challenges, and perspectives [Article]. *Sensors, 22*(24), 22. https://doi.org/10.3390/s22249916. Article 9916.

Hasson, F., Keeney, S., & McKenna, H. (2025). Revisiting the Delphi technique - Research thinking and practice: A discussion paper. *International Journal of Nursing Studies, 168*, Article 105119. https://doi.org/10.1016/j.ijnurstu.2025.105119

Hiroshima Process International. (2023a). HPI code of conduct for organizations developing advanced AI systems. https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document05_en.pdf.

Hiroshima Process International. (2023b). HPI guiding principles for organizations developing advanced AI system. https://ec.europa.eu/newsroom/dae/redirection/document/99643.

Ho, C. K., & Asrani, S. K. (2025). Ethics, bias, and governance in artificial intelligence for hepatology: Toward building a safe and fair future [Review]. *Journal of Clinical and Experimental Hepatology,, 15*(6). https://doi.org/10.1016/j.jceh.2025.102628. Article 102628.

Ho, C. W. L., & Caals, K. (2021). A call for an ethics and governance action plan to harness the power of artificial intelligence and digitalization in nephrology [Review]. *Seminars in Nephrology, 41*(3), 282–293. https://doi.org/10.1016/j.semnephrol.2021.05.009

Hoffman, R. R., Mueller, S. T., Klein, G., Jalaeian, M., & Tate, C. (2023). Explainable AI: roles and stakeholders, desirements and challenges [Article]. *Frontiers in Computer Science, 5*(18). https://doi.org/10.3389/fcomp.2023.1117848. Article 1117848.

Horty, J. F. (2001). *Agency and deontic logic*. Oxford University Press. https://doi.org/10.1093/0195134613.001.0001

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning, 110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

IEEE. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead1e.pdf.

IEEE. (2021). IEEE standard model process for addressing ethical concerns during system design. In IEEE Std 7000-2021 (pp. 1-82).
IEEE. (2025). IEEE standard for system, software, and hardware verification and validation. *IEEE Std*, 1012–2024. https://doi.org/10.1109/IEEESTD.2025.11134780 (Revision of IEEE Std 1012-2016), 1-309.
Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics, 3*(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X
Jha, S., Rushby, J., & Shankar, N. (2020). *Model-centered assurance for autonomous systems* (pp. 228–243). Springer International Publishing. https://doi.org/10.1007/978-3-030-54549-9_15
Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review, 93*(5), 1449–1475. http://www.jstor.org.ezproxy.uws.edu.au/stable/3132137.
Kalenzi, C. (2022). Artificial intelligence and blockchain: How should emerging technologies be governed? [Review]. *Frontiers in Research Metrics and Analytics, 7*. https://doi.org/10.3389/frma.2022.801549. Article 801549.
Kane, T. B. (2018). *A framework for exploring intelligent artificial personhood* (pp. 255–258). Springer International Publishing. https://doi.org/10.1007/978-3-319-96448-5_28
Kane, T. B. (2025). Philosophical approaches to managing generative AI agents as artificial persons at work. *Journal of Business Analytics*, 1–17. https://doi.org/10.1080/2573234x.2025.2482652
Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics [Review]. *Patterns, 2*(9). https://doi.org/10.1016/j.patter.2021.100314. Article 100314.
Kim, B. J., Jeong, S., Cho, B. K., & Chung, J. B. (2025). AI governance in the context of the EU AI Act [Review]. *IEEE Access, 13*, 144126–144142. https://doi.org/10.1109/ACCESS.2025.3598023
KPMG, Gates, D., Gampenrieder, E., & Hendricks, M.. (2019). Boardroom questions: Industry 4.0. https://assets.kpmg.com/content/dam/kpmg/be/pdf/2019/11/boardroom-questions-industry-4-0.pdf.
Langman, S., Capicotto, N., Maddahi, Y., & Zareinia, K. (2021). Roboethics principles and policies in Europe and North America [Review]. *SN Applied Sciences, 3*(12). https://doi.org/10.1007/s42452-021-04853-5. Article 857.
Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering, 6*(4), 239–242. https://doi.org/10.1007/s12599-014-0334-4
Leikas, J., Johri, A., Latvanen, M., Wessberg, N., & Hahto, A. (2022). Governing ethical AI transformation: A case study of auroraAI [Review]. *Frontiers in Artificial Intelligence, 5*. https://doi.org/10.3389/frai.2022.836557. Article 836557.
Liang, K., & Qin, Y. (2025). Designing a minimum viable product (MVP) approach for AI governance implementation. *Frontiers in Artificial Intelligence Research, 2*(3), 13. https://doi.org/10.71465/fair249
Liu, J., Yu, Y., Zhang, L., & Nie, C. (2011). An overview of conceptual model for simulation and its validation. *Procedia Engineering, 24*, 152–158. https://doi.org/10.1016/j.proeng.2011.11.2618
Lyons, J. B., Jessup, S. A., & Vo, T. Q. (2024). The role of decision authority and stated social intent as predictors of trust in autonomous robots. *Topics in cognitive science, 16*(3), 430–449. https://doi.org/10.1111/tops.12601
McCauley, L. (2007). AI armageddon and the three laws of robotics. *Ethics and Information Technology, 9*(2), 153–164. https://doi.org/10.1007/s10676-007-9138-2
Mirghaderi, L., Sziron, M., & Hildt, E. (2023). Ethics and Transparency Issues in Digital Platforms: An Overview [Review]. *AI (Switzerland), 4*(4), 831–843. https://doi.org/10.3390/ai4040042
Mohamed, N., Al-Jaroodi, J., Lazarova-Molnar, S., & Jawhar, I. (2016). Middleware to support cyber-physical systems. https://doi.org/10.1109/PCCC.2016.7820605.
Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: A pragmatic operationalisation of AI ethics. *Minds and Machines, 31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w
Müller, T., Kamm, S., Löcklin, A., White, D., Mellinger, M., Jazdi, N., & Weyrich, M. (2023). Architecture and knowledge modelling for self-organized reconfiguration management of cyber-physical production systems. *International Journal of Computer Integrated Manufacturing, 36*(12), 1842–1863. https://doi.org/10.1080/0951192x.2022.2121425
Navarro, P. E., & Rodríguez, J. L. (2014). *Deontic logic and legal systems*.
Nelson, S., & NASA. (2003). Certification processes for safety-critical and mission-critical aerospace software. (NASA/CR-2003-212806). Retrieved from https://ntrs.nasa.gov/citations/20040014965.
Noyes, J., Popay, J., Pearson, A., Hannes, K., & Booth, A. (2008). Qualitative research and cochrane reviews. In Cochrane Handbook for Systematic Reviews of Interventions (pp. 571-591). https://doi.org/10.1002/9780470712184.ch20.
O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors, 64*(5), 904–938. https://doi.org/10.1177/0018720820960865
OECD. (2024). Recommendation of the council on artificial intelligence OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
Oks, S. J., Jalowski, M., Lechner, M., Mirschberger, S., Merklein, M., Vogel-Heuser, B., & Möslein, K. M. (2022). Cyber-physical systems in the context of Industry 4.0: A review, categorization and outlook. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-022-10252-x
Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M., Galeano Sarmiento, J., Ruz, G. A., & Tabares-Soto, R. (2024). Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artificial Intelligence Review, 57*(5). https://doi.org/10.1007/s10462-024-10740-3
Pabel, A. S. M. S., & Akther, W. (2025). Navigating the ethical landscape: A systematic review of challenges in AI adoption within the accounting profession [Review]. *Multidisciplinary Reviews, 8*(11). https://doi.org/10.31893/multirev.2025357. Article e2025357.
Padmanabhan Poti, S., & Stanton, C. J. (2024a). Enabling affordances for AI governance [Article]. *Journal of Responsible Technology, 18*. https://doi.org/10.1016/j.jrt.2024.100086. Article 100086.
Padmanabhan Poti, S., & Stanton, C.J. (2024b). Mediating explainer for human autonomy teaming 2nd World Conference on eXplainable Artificial Intelligence (xAI-2024), Valletta, Malta. https://ceur-ws.org/Vol-3793/paper_26.pdf.
Poppendieck, M., & Poppendieck, T. (2003). *Lean software development : an agile toolkit*. Addison-Wesley.
Prakken, H., & Sartor, G. (2015). Law and logic: A review from an argumentation perspective. *Artificial Intelligence, 227*, 214–245. https://doi.org/10.1016/j.artint.2015.06.005
PRINCE2. (2023). *PRINCE2® 7 managing successful projects : global best practice* ([Seventh edition]). PeopleCert.
Prisacariu, C., & Schneider, G. (2012). A dynamic deontic logic for complex contracts. *The Journal of Logic and Algebraic Programming, 81*(4), 458–490. https://doi.org/10.1016/j.jlap.2012.03.003
Rawat, D. B. (2022). Artificial intelligence meets tactical autonomy: Challenges and perspectives. In *Proceedings - 2022 IEEE 4th International Conference on Cognitive Machine Intelligence, CogMI 2022*.
Rubin, K. (2012). *Essential Scrum: A practical guide to the most popular agile process*. Pearson Education. Limited.
Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A modern approach, ebook, global edition*. Pearson Education.
Russell, S. J., Norvig, P., & Chang, M.-W. (2022). *Artificial intelligence : a modern approach* (Fourth edition, global edition). Pearson Education Limited.
Russell, S. J.a. (2010). *Artificial intelligence : a modern approach* (3rd ed,, *2010*. Upper Saddle River, New Jersey: Prentice Hall.
Salama, D. A., & El-Gohary, N. M. (2013). Automated compliance checking of construction operation plans using a deontology for the construction domain. *Journal of Computing in Civil Engineering, 27*(6), 681–698. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000298
Schiff, D. S., Laas, K., Biddle, J. B., & Borenstein, J. (2022). *Global AI ethics documents: What they reveal about motivations, practices, and policies* (pp. 121–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-86201-5_7
Shapiro, S. P. (2005). Agency theory. *Annual Review of Sociology, 31*(1), 263–284. https://doi.org/10.1146/annurev.soc.31.041304.122159
Sifakis, J., & Harel, D. (2023). Trustworthy autonomous system development [Article]. *ACM Transactions on Embedded Computing Systems, 22*(3), 24. https://doi.org/10.1145/3545178. Article 40.
Sigfrids, A., Nieminen, M., Leikas, J., & Pikkuaho, P. (2022). How should public administrations foster the ethical development and use of artificial intelligence? A review of proposals for developing governance of AI [Review]. *Frontiers in Human Dynamics, 4*. https://doi.org/10.3389/fhumd.2022.858108. Article 858108.
Skeet, A., & Guszcza, J. (2020). How businesses can create an ethical culture in the age of tech. In *World Economic Forum Annual Meeting*. https://www.weforum.org/agenda/2020/01/how-businesses-can-create-an-ethical-culture-in-the-age-of-tech/.
Stahl, B. C., Heersmink, R., Goujon, P., Flick, C., Van Den Hoven, J., Wakunuma, K., Ikonen, V., & Rader, M. (2010). Identifying the ethics of emerging information and communication technologies: an essay on issues, concepts and method. *International Journal of Technoethics (IJT), 1*(4), 20–38.
Thalheim, B. (2019). *Conceptual models and their foundations* (pp. 123–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-32065-2_9
Tun, H. M., Naing, L., Malik, O. A., & Abdul Rahman, H. A. (2025). Navigating ASEAN region artificial intelligence (AI) governance readiness in healthcare [Review]. *Health Policy and Technology, 14*(2). https://doi.org/10.1016/j.hlpt.2025.100981. Article 100981.
UNESCO. (2022). *Recommendation on the Ethics Of Artificial Intelligence, 43*. https://unesdoc.unesco.org/ark:/48223/pf0000381137.
United Nations. (2024). UN system white paper on AI governance. https://unsceb.org/sites/default/files/2024-11/UNSystemWhitePaperAIGovernance.pdf.
Van Der Meulen, B. (2003). New roles and strategies of a research council: Intermediation of the principal-agent relationship. *Science and Public Policy, 30*(5), 323–336. https://doi.org/10.3152/147154303781780344
van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics, 1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6
Vicario, E. (2001). Engineering the usability of a visual formalism for real-time temporal logic. *J. Vis. Lang. Comput., 12*(6), 573–599. https://doi.org/10.1006/jvlc.2001.0214
WEF, Golbin, I., & Axente, M.L. (2021). Emerging Technologies: 9 ethical AI principles for organizations to follow. 16. https://www.weforum.org/agenda/2021/06/ethical-principles-for-ai/.
WEF, Dobrygowski, D., & Reim, D.. (2024). Technology Policy: Responsible design for a flourishing world. https://www.weforum.org/publications/technology-policy-responsible-design-for-a-flourishing-world/.
Xu, W. (2021). From automation to autonomy and autonomous vehicles: Challenges and opportunities for human-computer interaction. interactions, 28. https://doi.org/10.1145/3434580.
Zimmermann, O., Stocker, M., & Kapferer, S. (2024). Bringing ethical values into agile software engineering. *The Leading Role of Smart Ethics in the Digital World*.