AI and Ethics (2022) 2:697–711
https://doi.org/10.1007/s43681-021-00127-3

ORIGINAL RESEARCH

From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience

Ilina Georgieva[1] · Claudio Lazo[1] · Tjerk Timan[1] · Anne Fleur van Veenstra[1]

Received: 2 September 2021 / Accepted: 7 December 2021 / Published online: 12 January 2022

Abstract
In the field of AI ethics, after the introduction of ethical frameworks and the evaluation thereof, we seem to have arrived at a third wave in which the operationalisation of ethics is central. Operationalisation is required, since ethics frameworks are often not suited to be used by data scientists in the development of AI-based services or products. Therefore, in this paper, we aim to contribute to this third wave by mapping AI ethical principles onto the lifecycle of an AI-based digital service or product and combining it with an explicit governance model to clarify responsibilities in operationalisation. We then discuss practical, conceptual, and political implications of this analysis to end with key challenges around operationalising AI ethics.

Keywords AI ethics · AI lifecycle · Regulation of AI · Data science · Operationalisation of ethics · Governance of AI

1 Introduction

The promise of Artificial Intelligence (AI) solutions for many societal challenges, including such in education, law enforcement, health care and cybersecurity, comes with an equally well-publicised problematique of unfair treatment, challenges to fundamental rights and the labour market, accountably gaps, mis- or untraceable diagnoses, and adversarial system attacks. Accordingly, next to concerns about global competitiveness and productivity ([9], 1), approaches to AI regulation are highly topical matters. The latter have resulted in the abundance of ethical frameworks for the development of socially beneficial or *human-centred* AI and its deployment (what we call *the first wave* of literature on ethical AI), as well as in efforts to assess the potential normative consensus and convergence [22, 45, 55] of these frameworks (*the second wave*).[1] *The third wave* of scholarship on ethical AI, which builds on the first two by exploring how to turn AI principles into actionable data science practice and governance (see for instance [10, 12, 24, 29, 62]), examines as such the spectrum of applicability of AI ethics in real-world cases [45], highlights issues of standardization in specific sectors [33, 63], and connects questions of enforcement and regulatory control to AI governance [62]. Such studies document further the challenge in integrating AI systems within existing organizational structures, and show that not every AI-facilitated solution brings about the expected added (business) value [20].

This paper contributes to the third wave of scholarship on ethical AI by mapping AI ethical principles onto the lifecycle of an AI-based digital service or product to investigate their applicability for the practice of data science. At

✉ Ilina Georgieva
ilina.georgieva@tno.nl

[1] Strategic Analysis and Policy Department, Toegepast Natuurwetenschappelijk Onderzoek (TNO), Anna van Burenplein 1, 2595 DA Den Haag, The Netherlands

[1] Alongside the typology literature, there are also lively debates about the usefulness and impact of such ‘regulatory’ efforts. The criticism ranges from ethics’ inappropriateness as a soft law instrument [68] and their conflation with law in this context [13], doubts on their effectiveness [53], and implementability by developers or users [62] given their abstractness, to outcries against disregarding inherent questions of infrastructure and power [25], 66 or AI ethics’ use to merely delay actual regulation [43].

present, it is challenging for data scientists and organizations aiming to develop digital services to adhere to principles of AI ethics as a result of a lack of actionable guidelines and governance. Our work with public authorities, for instance, shows that public bodies are aware of AI ethics and checklists on national level, but struggle to single-handedly align them with EU-level guidance and often look for external help in the translation process [30]. Studies further show that organizations have the capacity to deal with only on a limited set of issues and mitigation measures during the rolling out of a service, suggesting the need for a broader stakeholder framework to ensure a comprehensive engagement and coverage of AI ethics [60].

Focusing on the development of actional guidelines for data science, our paper highlights which phases of the AI lifecycle require more attention in the implementation of ethical principles, and more importantly *by whom.* Our ambition is thereby twofold: on one hand, to contribute to the translation of AI ethics to data science practice relevant for services development, and to the wider debate on AI governance on the other. We do so by pointing to existing conceptual gaps, which are pre-conditioned and therefore embedded in current instruments for the development, deployment, and regulation of ethical AI, and which often materialize in systems' ex post-evaluations. By exposing these fallacies, we hope to start a conversation on how to avoid them in the future and to effectively accelerate the engagement of policymakers and professionals with AI institution building.[2]

Our work is inspired from an ongoing project on effective oversight for AI used by public bodies in the Netherlands [48], for which we conducted a preliminary principles-mapping exercise based on the seven requirements for trustworthy AI put forward by the European Commission's High-Level Expert Group on AI (HLEG) [1]. Specifically, we position the (sub)principles and values that substantiate the seven HLEG requirements in the phase of the AI lifecycle to which they pertain (see Table 3). With respect to the latter, we build on the work of Morley [45] and Stix [62], whose studies provide prototype frameworks for implementing high-level AI principles. We add to their work a governing layer using the rationale of Shneiderman's model for human-centred AI[3] that distinguishes between three levels of organizational structures—team, organization, and industry—when thinking about responsibility for an AI system [58]. By combining the results from visualizing the ethical principles on an AI system's continuum and Shneiderman's structures model, we identity conceptual and organizational gaps in the current conversation on trustworthy AI and draw conclusions for effective principles translation and human-centric governance of AI.

The paper proceeds as follows. To set the scene, the next section describes the context and process within which the AI HLEG guidelines emerged, including a definition of AI, and the background of the mapping exercise. The consecutive section then methodologically describes our principles mapping and provides tables to follow the process. A section on results and discussion draws conclusions from our principles mapping in a threefold manner—practically, conceptually, and politically—and discusses implications for the governance of AI. The outlook section raises points about the way forward.

2 Dealing with high-level AI principles: the HLEG ethics guidelines for trustworthy AI

2.1 Framing AI

To frame our work in the complex and vast scope of AI, we start with a definition that will guide the rest of our analysis. Based on a mix of OECD, EC, and AI HLEG definitions, the working definition for AI that we adhere to is the following:

AI applications are machine-based systems that rely on one or more algorithms, often using mathematical optimisation techniques. They are applied in specific settings, fulfilling requirements for a given set of human-defined objectives. They perform human-like cognitive functions automating one or more tasks such as gathering, combining, cleaning, sorting, classifying, and inferring (personal) data, as well as selecting, prioritising, recommending, predicting, and supporting or making decisions. AI systems are designed to operate with varying levels of autonomy, in a way that allows the creation of adaptive services at scale and in (near) real time, influencing real or virtual environments [71, 50].

Defined in this fashion, AI systems are used in perception of the environment, including the consideration of the real-world complexity; information processing, including the interpretation of inputs; decision-making, including reasoning and learning with a certain autonomy; and achievement of specific goals, considered as the ultimate reason of AI systems [56].

[2] For another take on the necessity of AI institution building, see Enholm et al. [20]. Artificial Intelligence and Business Value: a Literature Review. Information Systems Frontiers, 1–26. The authors discuss among others the role of AI ethics in adding value to the implementation of AI systems, and thereby to their underlying business cases.

[3] Shneiderman's 'human-centred' lens is of particular relevance in the context of the AI requirements by the EC's HLEG, who describes its own approach in developing the AI guidelines as embedded in the foundational values of the EU and thus as 'human-centric'.

2.2 From high-level requirements to national contexts

The Ethics Guidelines for Trustworthy Artificial Intelligence [1] are a natural continuation of the EU's Strategy on AI (COM(2018) 237) and a manifestation of the EC brand—*human-centric AI*—which puts people at the centre of the development of AI [18]. Central to the guidelines is further the notion of AI "made in Europe". The latter means AI development that is grounded in the foundational values of the EU—respect for human dignity, freedom, democracy, equality, the rule of law and human rights—and that pays due regard to ethical and human-centric considerations. In the process, commentators have readily drawn parallels to the case of the General Data Protection Regulation (GDPR), in which the EU used its first-mover advantage to widen the influence of European values on other stakeholders and markets by means of regulating the domain [9]. Accordingly, when it comes to the alignment of AI with the EU's values and the HLEG's framework, the former is seen as preparing the grounds with global standards of design and usability for another wave of the "Brussels effect" [8].

Even though the ethics principles formulated by the AI HLEG have not been left uncriticized,[4] stakeholders of all shapes and sizes have been dedicating time and resources to embed them in their operations. Our work with public authorities documented an ambivalent picture of those efforts. While aware of AI ethics and checklists on national level, public bodies struggle to single-handedly align them with EU-level guidance and would often opt for an 'AI requirements mix'. This means that they simultaneously apply principles from different ethical frameworks, giving priority to the ones that seem most straightforward and those on which they could acquire external guidance. In addition, they would seek the assistance of other specialised authorities such as the National Data Protection Authority.

The willingness of actors to engage with AI ethics has been further confirmed Stahl et al., who put forward that European organizations are not only aware of ethical issues surrounding AI, but that they are also willing to assume responsibility for actively engaging with them [60]. At the same time, organizations are reported to focus only on a limited set of issues and mitigation measures, maintaining that a number of ethical concerns lie beyond the scope of their expertise and remit, and suggesting the need for a broader stakeholder framework to ensure a comprehensive engagement and coverage of AI ethics [60].

Table 1 Classification categories

	AI lifecycle [45]	AI governance [58]
1	Business/use-case development	Safety culture (Organization)
2	Design	Reliable systems (Team)
3	Training & test data procurement	
4	Building	
5	Testing	
6	Deployment	
7	Monitoring	Safety culture (Organization)
8	Outside of the AI lifecycle	Trustworthy certification (Industry)

It is against this layered background that the mapping exercise has been carried out. As most actors that look into the use of AI to optimise their processes, public authorities are struggling to understand which principles apply to which applications, when and how to structure their oversight, and how to fit both principles and oversight in their organizational culture. In the Dutch context, the recent mishap surrounding the use of SyRI (Systeem Risico Indicatie) [49]—an AI system aiding municipalities and other governmental agencies to rank citizens according to the risk of them committing fraud with social benefits or taxes—has prompted stakeholders to think about co-creation of principles and technology.

3 Methodology

As indicated in the previous section, in our mapping exercise, we focus on the HLEG requirements for trustworthy AI—(1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) societal and environmental well-being; and (7) accountability—due to the nature of our ongoing project, conditioned on the needs of Dutch public authorities to engage with the guidelines to deduce actionable points from them.

The HLEG requirements for trustworthy AI are mapped onto the general AI-lifecycle stages, for which we use the AI-lifecycle typology of Morley [45] that guides machine learning (ML)—a specific type of AI—developers through the phases of algorithmic development. We extended their work by adding an additional layer to it—the governance structures of Shneiderman [58]— to visualize the party/entity who is likely to be responsible for overseeing the process of development and (principles) implementation. The governance layer to the model is also a valuable addition to

[4] Due to the composition of the HLEG group—its strong industry representation combined with the lack of proper civil society engagement [66]—the legitimacy of the produced guidelines has been questioned. Next to concerns of 'ethics washing' [41] and the enshrining of a revenue-interested business culture in them [29], the principles are also seen as weak from a procedural perspective ([37], 3).

Table 2 The requirement 'Explainability' in the HLEG Guidelines [1] mapped and coded to the AI lifecycle

Section	Coding	Classification
"Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system)"	(Not coded)	
"Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability)"	Technical explainability	Deployment; monitoring
"Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher)"	Process explainability	Deployment; monitoring
"In addition, explanations of the degree to which an AI system influences and shapes the organizational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency)"	Business explainability	Business/use-case development; monitoring

the discussion section, in which we point to the interplay of governance issues with power imbalances in AI development and deployment. The combination of both models is summarized in Table 1. We would like to note, however, that since Morley et al.'s model is designed for ML researchers and developers, the latter inevitably impacts the focus of our results as well.

Having established our lifecycle framework, we then proceeded to map the content of the seven HLEG requirements to each cycle stage. To do so, we first performed theme coding of the requirements. The latter was accomplished by conducting content analysis of the text that specifies the requirements. The explanatory statements (i.e., the principles that substantiate the requirements) were named and then classified within one or more relevant lifecycle stages or 'outside of the AI lifecycle'. Subsequently, we used the checklists in the revised Assessment List for Trustworthy AI (ALTAI) [2] for an additional reiteration round, as they may provide more concrete advice on how to implement the requirements. Upon going through the checklists, the positioning of some of the principles/sub-requirements was revised or added to additional cycle phases. For instance, we identified from the sections in the HLEG document describing 'explainability' [1] a number of sub-requirements, and coded them as *process explainability, technical explainability,* and *business explainability,* upon which they were individually classified to the AI lifecycle, as illustrated in Table 2. The full list can be found in Appendix 1: coding of requirements in ethical guidelines for trustworthy AI.

Under 'Explainability', the ALTAI checklist poses the question *'Do you continuously survey the users if they understand the decision(s) of the AI system?'* (p. 15), which relates *process explainability* to the continuous surveying of users [2]; thus, the requirement is also meant to apply in the monitoring phase. One could argue and go further in claiming that explainability, when broken down into subcategories of the AI lifecycle, and also, for the AI development process itself, there are more places or instance in which the question of explaining, the 'what', 'for whom' and 'how' are relevant across different stages that are something related to one of the three types in the table above, or sometimes cut across all three [see [4]]. Quality controllers, ML ops process designers, external auditors, front-end developers, or data controllers to name a few are all roles that somehow deal with both technical, process- and business explainability. Moreover, within the technical domain, there is a distinction to be made between types of models, levels of complexity, and moment of explanation (during the development process, or post-model development, for example). 4 discern "between (1) opaque systems, where the mappings from input to output are invisible to the user; (2) interpretable systems, in which users can mathematically analyse the mappings; and (3) comprehensible systems, in which the models should output symbols or rules along with their specific output to aid in the understanding process of the rationale behind the mappings being made" ([4], p. 10, see also [44]).

Similarly, the requirement 'avoidance of unfair bias' [1] consists of two different sub-requirements: *avoidance of unfair bias,* which calls for the removal of discriminatory biases in the collection phase and suggests oversight processes; and *diversity* in hiring practices. Upon investigating the corresponding ALTAI checklist [2], we find that for both sub-requirements, the focus is widened across the AI lifecycle. The checklist in this regard is both very general and strangely specific. Surely hiring processes are one important area where bias is augmented through the use of AI-based systems: it surely is not the only one. We see in general that questions of bias and fairness also reside in several steps in the AI-lifecycle process, with similar challenges both in AI technology itself as in the requirements and context around the AI-based system or service that is being developed. More generally, when categorized as '*tech4equality*' versus '*equality4tech*'[5], the former would be aimed at technologies

[5] Terms coined by [70].

Table 3 Heatmap of trustworthy AI requirements across the AI lifecycle, and its corresponding governance structures and responsible stakeholders

	Governance structure	Safety culture	Reliable systems					Safety culture	Trustworthy certification
	Responsibility	Organization	Team					Organi-zation	Industry
	AI lifecycle stage	Business / use case development	Design	Training & test data procurement	Building	Testing	Deploy-ment	Monitor-ing	*(Outside AI lifecycle)*
Requirements for trustworthy AI (HLEG)	Human agency and oversight	0	2	0	0	0	5	5	2
	Technical robustness and safety	0	5	0	1	6	7	1	0
	Privacy and data governance	0	1	4	0	2	2	0	1
	Transparency	1	0	0	0	0	6	4	0
	Diversity, non-discrimination and fairness	2	4	2	2	2	1	2	1
	Societal and environmental well-being	1	0	2	2	0	2	3	0
	Accountability	0	2	0	2	0	2	4	1

Numbers (also represented by colour intensity) indicate the amount of sub-requirements that are engaged with in the corresponding stage of the AI lifecycle (For a description and mapping of all individual sub-requirements, see Appendix 1)

that help alleviate bias within AI systems, whereas the latter is concerned with external factors that can help make AI more fair (for instance, by creating more diverse developer teams or train non-experts in data and AI logics, etc.). The former would fit in Shneiderman's reliable systems-category, where the latter would fit in both the safety culture-category as in the wide industry responsibility-category.

We proceeded in the manner described above with the coding and mapping of all HLEG requirements.

4 Results and discussion

This section presents the results of our principles-mapping exercise in a twofold manner. It visually portrays the comprehensive mapping (Table 3) resulting from the implementation of the methodology described in the previous section, and offers further observations and discussion points on the implications of the current design of the HLEG guidelines for AI governance.

Table 3 shows how the weight of the requirements is distributed across the AI lifecycle presented in a heatmap. The heatmap contains the requirements—vertically grouped per high-level requirement (e.g., 'Transparency'), and the AI lifecycle (in the horizontal direction), containing the phases of *Business/use-case development, Design, Training and test data procurement, Building, Testing, Deployment, and Monitoring*. The table allows us to see during which phases the requirements are mostly engaged with and under whose responsibility. For instance, the 'Transparency' requirement is accounted for during the first, sixth, and seventh phase of the AI lifecycle, leaving a gap from the design phase until the testing phase. Note that the table is an aggregate view of all the coded sub-requirements across the corresponding intersection of their parent requirement, and the phases in which they are required to be accounted for. See Appendix 1: coding of requirements in Ethical Guidelines for Trustworthy AI for the detailed mapping of coded sub-requirements.

4.1 Findings and discussion

In this section, we discuss the results of our mapping exercise. For the sake of overview and comprehensiveness, the results are structured around practical, conceptual, and normative/political observations. These sub-sections address in turn questions of operationalisation of the ethical principles, the human-centeredness of the HLEG requirements, and questions pertaining to the embedding of ethical frameworks in society—the already existing normative structures.

4.2 Practical findings and implications

Our results show that during the AI design phase, not many ethical requirements are actively engaged with, and professionals are not given the tools to do so themselves. While the guidelines are design-oriented towards accessibility, stakeholder participation, and security, the other requirements are mainly engaged with through impact and risk assessments. Exceptions are the documenting of data processing, the testing of data sets, and the creation of diversity in team composition. This leaves the requirements to ‘avoid unfair bias’, for which the ALTAI checklist implicates the algorithm design stage, and the requirement to ‘explicate which trade-offs have been made’. However, the requirement states that “trade-offs should be addressed in a rational and methodological manner within the state of the art” [2]. To that end, when it comes to methods for implementing the requirements, the guidelines give limited guidance other than appointing responsibility to developers, deployers (e.g., managers), and end-users [1].

For managers, whereas the guidelines promote positive corporate governance principles such as *stakeholderism*, it is yet unclear how to translate them into practice, leaving much room for corporations to use AI to further predominantly or only shareholder-related interests [28]. Moreover, AI as an organizational capability requires numerous (less tangible) human and organizational resources such as interdepartmental coordination, organizational chance capacity, and risk proclivity [42], which are by definition prerequisites for the implementation of any human-centric capability within an organization.

For developers, the document calls for the ‘translation’ of ethical requirements to an operational level, either into procedures and constraints (e.g., blacklisting behaviour) or ethics-by-design [1], which afterwards should be tested and validated to establish whether the AI system satisfies the ethical requirements. Only then, we can speak of Trustworthy AI, where the human centricity is addressed in a deployed AI. However, practitioners have major issues in adequately implementing these guidelines [21], let alone correctly validating adherence to the guidelines. Thus, we need to further inspect these methods for translation. The required ‘translation’ of ethical requirements raises two related methodological issues: (1) that ethical requirements are not in alignment with AI development practices, and (2) that there is no widespread knowledge on how ethics may be tractably implemented and validated in software.

First, this desirable structural translation method—in essence engineering—is yet to be introduced in the analytical AI design practices. Software Engineering (SE) structurally specifies what is expected of the system in terms of requirements [Requirements Engineering (RE)], and how the system will satisfy those requirements—the software design [34]. In practice, however, the development of Machine Learning (ML)-based systems relies heavily on data analysis and model-fitting on trial-and-error basis [57]. There are some reference processes, such as the Cross-Industry Standard Process for Data Mining (CRISP-DM), of which the initial step—‘business understanding’—relates to Requirements Engineering. However, “neither the RE nor the ML community […] detailed what these steps demand” [67]. Furthermore, as these AI systems require a vast, complex infrastructure that possibly introduces risks, the lack of software engineering practices may incur long-term costs [57]. As a result, there are many barriers for scaling AI deployments [19], and especially so for considering ethical principles such as fairness [31]. Another challenge posed by AI systems is that over time they may deviate from the ethical requirements embedded in them [64].

Second, translation, formulation, and validation of ethical requirements are yet to be thoroughly researched. Given a set of operationalized, tractable, ethical requirements, validation is necessary to ascertain whether the implemented AI system satisfies these requirements and is thus trustworthy. To that end, the Software Engineering community has had multiple calls to action, as there is a lack of methods to account for ethics and values in software design. This is the case specifically with regard to the operationalization of values in software [46], the tractability of ethics in software [5], and software engineering for fairness [11]. Although symbolic methods for representing ethical requirements are being investigated [54], no methods have been produced to date in line with the calls for within the field of Software Engineering.

As a potential way forward, we discern third-wave operationalization efforts within the Design for Values (DfV) community. DfV is a methodological design approach that aims at making values part of the technological design, research, and development [32]. We are already aware that ICT products such as AI systems embed human values [61] and that the explicit translation of human values to design requirements is possible through a ‘value hierarchy’ [65]. In a recent effort to produce an aforementioned design method, Umbrello and Van de Poel [64] propose an iterative process. Their process is based on the AI for social good (AI4SG) values [23], which relate to the ethical principles in the Ethical Guidelines for Trustworthy AI [1]. Similarly, Aizenberg and Van den Hoven [3] make the first step towards an AI design method based on the EU’s foundational values. It seems that the DfV school of thought is laying the groundwork for the operationalization of ethical principles into AI.

4.3 Conceptual findings and implications

Conceptually, we found that in comparison to the open blanks in the business and design stage, the table becomes denser towards the deployment and monitoring phase. Several more specific principles/technical measures are then supposed to ensure proper AI functioning and ethical alignment. This leaves us questioning whether the framework's underlying rationale of promoting human-centric AI is being done justice.

Our doubts in that regard are further substantiated when examining more closely the requirements of human agency, transparency, and accountability, and how those are being dealt with throughout the system's lifecycle. In the case of human agency, it seems to mainly become relevant in the shape of human support, intervention, or external oversight in the post-testing stages. Looking at transparency, the alarming white spaces in the table between the design and test phases, combined with opt-outs and interventions of process and technical explainability during deployment and monitoring, speak for themselves. A similar pattern can be discerned with accountability's implementation, which besides impact assessments in the design and building phases, is also more prominent towards the end of the cycle. These requirements are thus significantly triggered in the (post-)implementation phase of an AI system, when harm is to be averted or has likely occurred. We thus see that the HLEG framework provides space for intervention or damage control, but not for much interaction. Its logic is not that far from technological determinism [15], Marx [39], treating AI as an unstoppable force following a pre-determined path that employs ethical considerations to avoid or mitigate negative impacts [25, 26, 40]. In other words, the framework's regulatory rationale is what we would describe as mostly reactive, often solutionist [40] or *interventionist* but not really holding up to the promise of human-centric.

Our analysis is corroborated also on governance level by Shneiderman's structures model. The latter stipulates that human-centred AI emerges when managers, designers, and software engineers adopt participatory design methods by engaging with diverse stakeholders to cater to and emphasize user experiences [58]. By that Shneiderman means not only shifting the researchers'/developers' focus on measuring human performance and satisfaction, valuing customer and consumer needs, ensuring meaningful human control, etc., but also casting a broader contextual net to take on board considerations of organizational and political culture, and how human experiences across the three governance structures are conditioned by those as well.

Read in this way, Shneiderman's governance framework resonates strongly with other contemporary voices, who question the limited engagement of the ethical AI debate with power structures and the decisions that rest within those, be it on governmental level or on that of private companies [24, 25, 29, 35, 52]. Our conceptual observations thus inevitably lead us to the political ones.

4.4 Normative and political implications

Scholarly work that investigates power relations surrounding the processes and effects of AI often does so by raising questions of accountability and the obscurity of legal and institutional boundaries in which the former exists [52]. Let us begin with the latter. To better explain why we deem it important to bring our attention to legal and institutional boundaries, we should first go back to our mapping results and the visualization of the business/use-case part of the AI lifecycle.

On theoretical level, the requirements are designed to raise questions of the explainability of the business model for review purposes, but requirements expected to have a more direct regulatory impact on AI effects such as human agency (including human dignity and autonomy), accountability, and oversight remain outside of the scope of the business model stage. This is something that we cannot construe as a slip or knowledge shortage on behalf of the stakeholders contributing to the HLEG requirements. After all, the group's composition is known for its industry affiliation [66], and therefore, it is safe to assume a sufficient degree of awareness of business modus operandi.

Instead, the table's blanks in this category are part of what Mittelstadt referred to as "deep political and normative disagreement" [43], and what we put forward as a *normative dissonance* between the bottom–up professional ethics of the fields of AI and data science, and the top–down definitions of what should be regarded as professional in those fields [25] propagated on institutional level. The dissonance is facilitated by focusing the ethical AI debate almost exclusively on data scientists and their data subjects (the latter of which are beginning to receive a more prominent stance in the AI lifecycle due to human-centric considerations [58, 59]), while omitting to extend the same courtesy to business developers and the broader normative structures that allow the AI industry to thrive. However, ethical failures are not accidental but tightly connected to business models [38]. Failing to acknowledge that the AI lifecycle is facilitated by and embedded within other standards, regulations and agents [52], and with that the need for systemic thinking, the debate supports an idealized process, one in which AI is handled as apolitical [17, 52].

We therefore agree with Green [25] and others [13, 53], who put forward that to expect questions on the desired AI impacts and effects, or even going a step further, on what AI solutions are desirable and which ones should be left on the drawing board, to be resolved by engaging data scientists with AI ethics, is misunderstanding the process based on an incomplete theory of social change [69]. AI ethics are not meant to be easy or formulaic, just as principle-based frameworks are not complete systems for ethical decision-making [12]. They are a tool against cognitive biases, a conversation facilitator paving the way for dealing with disagreements on vital ethical questions. Yet, to confuse a priori consensus with a meaningful consensus [43] is a fallacy that would not help ethics achieve their ends as a vital part of the deliberative process. In other words, refusing to acknowledge the politics of AI ethics, and how those reinforce the normative structures within which we deliberate on principled, human-centric AI, will continue setting the debate in the wrong key.

The latter brings us to our more concrete observations on responsibility and accountability. As portrayed by our mapping exercise, accountability is often dealt with in terms of responsibility for impact assessments, redress, and negative impact minimization, and is mostly prominent towards the end of the AI lifecycle. The governance layer added to our table, and inspired by Shneiderman, helps us to then easily determine whether the impact assessment and other related steps would fall under the responsibility of the team, the organization, or of that of external (industry or governmental) actors. Therewith, we obtain a certain level of transparency regarding who would play a pivotal role in the algorithm's development.

However, while such a result remedies some of the responsibility grievances—especially those that caution against equating authorship of technical operations with questions on who or what should be made to answer for those operations [52]—as it allows us to visualize the distributed nature of AI accountability processes, it seems to overburden professionals (the team), while marginally implicating the bodies (organization, industry, and governmental actors) who have set the parameters ([51], 7). The exercise further reaffirms the by now well-known mantra—that attributing responsibility for (omissions in) an AI lifecycle is not necessary the same as attributing responsibility for the implementation of algorithmic decisions. The latter cannot be individually located, absolute, or finite [51].

When looking at accountability, a similar lifecycle challenge occurs: it is hard to project accountability in an AI-service lifecycle on the different responsibility roles or frameworks, which, as showed above, often focus inwards (procedural 'housekeeping') with the idea that this instigates a positive trickle-down effect in society at large. We have seen examples of such an effect in particular ICT-technology companies taking a leading role in privacy protection, and thereby 'leading the way'.[6]

If we take the division as introduced above, at the team level, accountability can be viewed in relation to keeping up with the state-of-the-art, passing on knowledge and insights within and among teams, contributing to online discussion fora or following key virtuous influencers in the field [27], for example. Moreover, to internal stakeholders, accountability is often seen as remaining within time and budget constraints and being able to deliver what a marketing department has already sold [36]. On organizational level, we have seen the development of novel roles and responsibilities such as data protection compliance through the widespread instatement of privacy officers; something similar might happen for AI, be it voluntary or even enforced via law in the EU. Accountability for AI here could involve broader and more fundamental questions on the goal and added value of making systems relying on AI that at some point becomes hard or impossible to explain [44], and on the legal and organizational safeguards that need to be in place to allow for responsible development of such systems (which will be highly contextual, and linked to core questions of liability in AI-based products or systems [16]). On industry level, adoption and consensus-making around accountability in novel ICTs can take place via standards and certification or labels (self-regulation) or via sector-specific or generic regulation. The issue with the former is that this form of self-regulation needs widespread voluntary adoption or market forces to actually work, while the issue with the latter is that enforcing such regulation needs competent auditing institutions (who are not all ready to actually audit different manifestations of AI and automated decision-making, although on top of the agenda for many), and the strengthening of the legal pillar with timely jurisprudence for the sake of effective redress. So far, regarding AI accountability, recent cases involving fairness in AI show that the only potential legal avenues rely on either data protection, anti-discrimination legislation, or liability law, and none of these three are particularly well suited to account for treating AI- harms [7].

The question of accountability is closely related to that of explicability of not only the chosen algorithm or the used datasets to train a particular model [61], but it should also be embedded on organizational and industry level. Operationalizing AI ethical guidelines is something that is currently taking place in pilots and experiments (often after something has gone wrong); we are also contributing to this process through an action-research approach [6], in which the aim is to connect ethics and politics of local governmental organization with their data science departments and industry standards on questions of transparency, accountability, and explicability of AI-based public services. One of the key elements for operationalization is having an applicable and effective governance framework. 'Effective' in this case stands for taking into account both the design process as well as the organizational reality and industry development, and allowing for critical business-case evaluation of the goals, risks and potential consequences for the company, system

[6] See for instance, [47]. Privacy becomes a selling point for tech, with Apple and Microsoft leading the way. Retrieved from: https://www.geekwire.com/2019/privacy-becomes-selling-point-tech-companies-apple-microsoft-leading-way/.

end-users, and the implicated actors of developing AI-based services [14].

5 Conclusion

In this paper, our aim was to contribute to a third wave of AI ethics research by looking into the challenges around operationalisation of AI ethics. This contribution consists of analysing current work in defining guidelines on what European and human-centric AI could look like, most notably the AI HLEG, the work by Morley et al. and Shneiderman, and connecting these through a current case study in which the authors work with a local Dutch municipality that aims to develop an AI-based public service based on ethical guidelines. Through plotting the different guidelines and steps onto the data-service AI lifecycle, a table emerged that allowed for gap analysis. The table provides at a glance which ethical principles are most prominent in the different parts of the AI lifecycle and what is already in place in terms of standards or guidelines at each cross-section. Our intention was thus not to fill the entire table, but rather to show how certain values or ethical principles are taken up and by whom, and how responsibility and accountability are currently 'plotted'. We continue by providing practical, conceptual, and political findings, and implications of the mapping exercise.

Whereas at the time of writing, this is ongoing research, already early on manifestations of the need to develop the system in a transparent, explainable, and auditable way show that translating these points in the design process of the digital service leads to some key challenges of methodological and practical uncertainty. The practical implications evolve around translation work from principles into software engineering standards and requirements, which at the moment are lacking in AI engineering. One conceptual implication of the mapping exercise is that in comparison to the open blanks in the business and design stage, the table becomes denser towards the deployment and monitoring phase—the human as a subject, end-user, or auditor of the AI system is far less prominent in the development stages. This begs the question whether, how, and to what extent humans can or should be there in the entire 'loop' and if so, how to design meaningful interactions with, for instance, choices of AI models or outcomes of model-training sessions. Here, much work is needed if we truly want to democratize AI. Connecting the latter point to the political or normative implications, the development of ethical AI is also highly dependent on the political and institutional landscape. Novel forms of democratic oversight might be needed on top of novel forms of citizen engagement and feedback mechanisms when we are interacting with AI-based systems. As argued elsewhere, the solution lies not on one single framework that offers a simplified form of ethics ("oversimplification of complex and difficult ethical debates hidden in the details of the application of principles" [43]), or in merely establishing a better alignment of ethical principles with AI design practices. The analysis offered in this paper shows that we will need an entire landscape of methods, standards and procedures if we want to develop AI-based services for 'good'. This point is further strengthened by the fact that we started out with a fairly practical problem statement—one stemming from the translation of principles to data science—but our findings on normative/political implications outweigh a mere data science or software engineering focus. We aim to contribute to the cause of 'AI for good' through a carefully designed real-world and multi-stakeholder project in which the AI HLEG principles and guideless are operationalized, ideally leading to filling in some of the blanks in this landscape.

Appendix 1: Coding of requirements in ethical guidelines for trustworthy AI

The rows in the following table should be read as 'sub-requirement *x* is accounted for during phases *y, z, …* of the AI lifecycle'. For instance, the sub-requirement *Stop button*, which is described by the Assessment List [2], refers to a 'stop button or procedure to safely abort an operation when needed' (p. 8)—thus mapped to the *Deployment* phase in the AI lifecycle as it refers to a mechanism or procedure for aborting an operational AI system.

High-level requirement	Coding	Sub-requirement text	Maps to AI lifecycle
Human agency and oversight	Fundamental rights assessment	In situations where [fundamental rights] risks exist, a fundamental rights impact assessment should be undertaken. This should be done prior to the system's development and include an evaluation of whether those risks can be reduced or justified as necessary in a democratic society to respect the rights and freedoms of others	Design
	Fundamental rights mechanism	Moreover, mechanisms should be put into place to receive external feedback regarding AI systems that potentially infringe on fundamental rights	Monitoring
	Human agency goal	Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system	Outside
	Human support	AI systems should support individuals in making better, more informed choices in accordance with their goals	Deployment
	No unwarranted influence	AI systems can sometimes be deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning, all of which may threaten individual autonomy	Deployment
	Human autonomy	The overall principle of user autonomy must be central to the system's functionality. Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them	Deployment
	Human oversight	Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects	Monitoring
	Oversight training	Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?	Outside
	Stop button	Did you ensure a 'stop button' or procedure to safely abort an operation when needed? (ALTAI, p.8)	Deployment
	Human-in-the-loop (HITL)	Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable	Deployment
	Human-on-the-loop (HOTL)	HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation	Design; monitoring
	Human-in-control (HIC)	HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system	monitoring
	External oversight	Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate. Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required	Monitoring
Technical robustness and safety	Technical robustness	Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured	Design
	Resilience	AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, e.g. hacking	Design; testing; Deployment

High-level requirement	Coding	Sub-requirement text	Maps to AI lifecycle
	Security	For AI systems to be considered secure, possible unintended applications of the AI system (e.g. dual-use applications) and potential abuse of the system by malicious actors should be taken into account, and steps should be taken to prevent and mitigate these	Design; testing; Deployment
	Fallback plan	AI systems should have safeguards that enable a fallback plan in case of problems. This can mean that AI systems switch from a statistical to rule-based procedure, or that they ask for a human operator before continuing their action	Deployment; monitoring
	General safety	It must be ensured that the system will do what it is supposed to do without harming living beings or the environment. This includes the minimisation of unintended consequences and errors	Design; testing; deployment
	Safety risk assessment	In addition, processes to clarify and assess potential risks associated with the use of AI systems, across various application areas, should be established	Design; testing; deployment
	Accuracy	An explicit and well-formed development and evaluation process can support, mitigate and correct unintended risks from inaccurate predictions. When occasional inaccurate predictions cannot be avoided, it is important that the system can indicate how likely these errors are. A high level of accuracy is especially crucial in situations where the AI system directly affects human lives	Building; testing; deployment
	Reliability	It is critical that the results of AI systems are reproducible, as well as reliable. A reliable AI system is one that works properly with a range of inputs and in a range of situations	Deployment
	Reproducibility	Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions. This enables scientists and policy makers to accurately describe what AI systems do. Replication files can facilitate the process of testing and reproducing behaviours	Testing
Privacy and data governance	Privacy	AI systems must guarantee privacy and data protection throughout a system's entire lifecycle. This includes the information initially provided by the user, as well as the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI system generated for specific users or how users responded to particular recommendations). Digital records of human behaviour may allow AI systems to infer not only individuals' preferences, but also their sexual orientation, age, gender, religious or political views. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them	Training and test data procurement; testing; deployment
	Data quality	The quality of the data sets used is paramount to the performance of AI systems. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set	Training and test data procurement
	Data integrity	In addition, the integrity of the data must be ensured. Feeding malicious data into an AI system may change its behaviour, particularly with self-learning systems	Training and test data procurement
	Data processing testing and documentation	Processes and data sets used must be tested and documented at each step such as planning, training, testing and deployment. This should also apply to AI systems that were not developed in-house but acquired elsewhere	Design; training and test data procurement; testing; deployment
	Data access	In any given organization that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place. These protocols should outline who can access data and under which circumstances. Only duly qualified personnel with the competence and need to access individual's data should be allowed to do so	Outside
Transparency	Traceability	The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability	Deployment; monitoring

High-level requirement	Coding	Sub-requirement text	Maps to AI lifecycle
	Process explainability	Whenever an AI system has a significant impact on people’s lives, it should be possible to demand a suitable explanation of the AI system’s decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher)	Deployment; Monitoring
	Technical explainability	Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings	Deployment; monitoring
	Business explainability	In addition, explanations of the degree to which an AI system influences and shapes the organizational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency)	business/use-case development; monitoring
	Representation	AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such	Deployment
	Opt-out	In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights	Deployment
	Communication	Beyond this, the AI system’s capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations	Deployment
Diversity, non-discrimination and fairness	Avoidance of unfair bias	Identifiable and discriminatory bias should be removed in the collection phase where possible. The way in which AI systems are developed (e.g. algorithms’ programming) may also suffer from unfair bias. This could be counteracted by putting in place oversight processes to analyse and address the system’s purpose, constraints, requirements and decisions in a clear and transparent manner	Design; training and test data procurement; building; testing
	Diversity	Moreover, hiring from diverse backgrounds, cultures, and disciplines can ensure diversity of opinions and should be encouraged	Business/use-case development; design; training and test data procurement; building; testing; monitoring; outside
	Accessibility	Particularly in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards. This will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities and with regard to assistive technologies	Business/use-case development; design; deployment
	Stakeholder participation	In order to develop AI systems that are trustworthy, it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its lifecycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organizations	Design; monitoring
Societal and environmental well-being	Sustainable AI	Sustainable and environmentally friendly AI: AI systems promise to help tackling some of the most pressing societal concerns, yet it must be ensured that this occurs in the most environmentally friendly way possible. The system’s development, deployment and use process, as well as its entire supply chain, should be assessed in this regard, e.g., via a critical examination of the resource usage and energy consumption during training, opting for less harmful choices. Measures securing the environmental friendliness of AI systems’ entire supply chain should be encouraged	Training and test data procurement; building; deployment; monitoring

High-level requirement	Coding	Sub-requirement text	Maps to AI lifecycle
	Social impact	Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could also affect people's physical and mental well-being. The effects of these systems must therefore be carefully monitored and considered	Business/use-case development; monitoring
	Society and democracy	Beyond assessing the impact of an AI system's development, deployment and use on individuals, this impact should also be assessed from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts	Training and test data procurement; building; deployment; monitoring
Accountability	Auditability	Auditability entails the enablement of the assessment of algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety–critical applications, AI systems should be able to be independently audited	monitoring
	Minimisation and reporting of negative impacts	Both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the **potential** negative impacts of AI systems is especially crucial for those (in) directly affected	Outside
	Protection	Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system	Monitoring
	Impact assessments	The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI systems can be helpful to minimise negative impact. These assessments must be proportionate to the risk that the AI systems pose	Design; building; deployment; monitoring
	Trade-offs	When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a rational and methodological manner within the state of the art. This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights	Design; building
	Redress	When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress. Knowing that redress is possible when things go wrong is key to ensure trust. Particular attention should be paid to vulnerable persons or groups	Deployment; monitoring

Funding This study was funded by TNO—Program 'Accelerating Trustworthy AI' and Dutch AI Coalition.

References

1. AI HLEG: Ethics guidelines for trustworthy AI. AI HLEG (2019)
2. AI HLEG.: Assessment list for trustworthy Artificial Intelligence. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (2020)
3. Aizenberg, E., Van den Hoven, J.: Designing for human rights in AI. Big Data Soc. (2020). https://doi.org/10.1177/2053951720949566
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, **58**, 82–115
5. Aydemir, F. B., Dalpiaz, F.: A roadmap for ethics-aware software engineering. In: 2018 IEEE/ACM international workshop on software fairness (FairWare). IEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8452915 (2018)
6. Baskerville, R. L.: Investigating information systems with action research. Communications of the association for information systems. **2** (1999): 19
7. Borgesius, F.J.Z.: Strengthening legal protection against discrimination by algorithms and artificial intelligence. Int. J. Hum. Rights

24(10), 1572–1593 (2020). https://doi.org/10.1080/13642987.2020.1743976
8. Bradford, A.: The Brussels effect: how the European Union Rules the World. Oxford University Press, Oxford (2019)
9. Brattberg, E., Csernatoni, R., Rugova, V.: Europe and AI: leading, lagging behind, or carving its own way? Carnegie Endowment for International Peace, Washington (2020)
10. Brown, S., Davidovic, J., Hasan, A.: The algorithm audit: scoring the algorithms that score us. Big Data Soc. (2021). https://doi.org/10.1177/2053951720983865
11. Brun, Y., Meliou, A.: Software fairness. In: Proceedings of the 26th ACM joint European software engineering conference and sympo-sium on the foundations of software engineering (ESEC/FSE '18), pp. 754–59. (2018). https://doi.org/10.1145/3236024.3264838
12. Canca, C.: Computing ethics: operationalizing AI ethics principles. Commun. ACM **63**(12), 18–21 (2020). https://doi.org/10.1145/3430368
13. Carillo, M.R.: Artificial Intelligence: from ethics to law. Telecommun. Policy **44**(6), 101937 (2020)
14. Clarke, A.E., Star, S.L.: The social worlds framework: a theory/methods package. In: Hackett, E.J., Amsterdamska, O., Lynch, M., Wajcman, J. (eds.) The handbook of science and technology studies, 3rd edn., pp. 113–137. MIT Press, Cambridge (2008)
15. Dafoe, A.: On technological determinism: a typology, scope conditions, and a mechanism. Sci. Technol. Hum. Values **40**(6), 1047–1076 (2015). https://doi.org/10.1177/0162243915579283
16. Dheu, O.: EU report on AI, new technologies and liability: key take-aways and limitations. Leuven. https://lirias.kuleuven.be/2937017 (2020)
17. Donovan, J., Caplan, R., Matthews, J., Hanson, L. Algorithmic accountability: a primer. In: Data and society tech algorithm briefing: how algorithms perpetuate racial bias and inequality. Washington, DC. https://ssrn.com/abstract=3518482 (2018). Accessed 30 Nov 2021
18. EC.: Building trust in human-centric artificial intelligence. Brussels. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58496 (2019). Accessed 30 Nov 2021
19. Eitel-Porter, R.: Beyond the promise: implementing ethical AI. AI Ethics **1**(1), 73–80 (2021). https://doi.org/10.1007/s43681-020-00011-6
20. Enholm, I.M., Papagiannidis, E., Mikalef, P., Krogstie, J.: Artificial intelligence and business value: a literature review. Inf. Syst. Front. (2021). https://doi.org/10.1007/S10796-021-10186-W/TABLES/8
21. Figueras, C., Verhagen, H., Pargman, T. C.: Trustworthy AI for the people? In: AIES 2021—proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society, pp. 269–70. (2021). https://doi.org/10.1145/3461702.3462470
22. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., Srikumar, M. Principled Artificial Intelligence: mapping consensus in ethical and rights-based approaches to principles for A. No. 2020-1. The Berkman Klein Center for internet and society research publication series. https://ssrn.com/abstract=3518482 (2020). Accessed 30 Nov 2021
23. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
24. Gabriel, I.: Artificial intelligence, values, and alignment. Mind. Mach. **30**, 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2
25. Green, B. Data science as political action: grounding data science in a politics of justice. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3658431 (2020). Accessed 30 Nov 2021
26. Greene, D., Hoffmann, A. L., Stark, L. Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Proceedings of the 52th Hawaii international conference on system sciences, pp. 2122–31. https://hdl.handle.net/10125/59651 (2019)
27. Hagendorff, T.: AI virtues—the missing link in putting AI ethics into practice. arXiv e-prints (2020): arXiv-2011
28. Hickman, E., Petrin, M.: Trustworthy AI and Corporate Governance: the EU's ethics guidelines for trustworthy artificial intelligence from a company law perspective. Springer (2021). https://doi.org/10.1007/s40804-021-00224-0
29. Hickok, M.: Lessons learned from AI ethics principles for future actions. AI Ethics **1**(1), 41–47 (2021). https://doi.org/10.1007/s43681-020-00008-1
30. Hoekstra, M., Chideock, C., Van Veenstra, A. F.: Quick scan AI in Publieke Dienstverlening II (2021). https://www.rijksoverheid.nl/documenten/rapporten/2021/05/20/quickscan-ai-in-publieke-dienstverlening-ii
31. Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M., Wallach, H.: Improving fairness in machine learning systems: what do industry practitioners need?" In: Conference on human factors in computing systems—proceedings. (2019). https://doi.org/10.1145/3290605.3300830
32. van den Hoven, M. J., Vermaas, P. E., van de Poel, I.: Handbook of ethics, values, and technological design—sources, theory, values and application domains. In: van den Hoven, M. J., Vermaas, P. E., van de Poel, I. (eds.), Springer. https://www.springer.com/gp/book/9789400769694 (2015). Accessed 30 Nov 2021
33. Huang, J.Y., Gupta, A., Youn, M.: Survey of EU ethical guidelines for commercial AI: case studies in financial services. AI Ethics **1**(4), 569–577 (2021). https://doi.org/10.1007/s43681-021-00048-1
34. IEEE Computer Society.: Guide to the software engineering body of knowledge (SWEBOK). In: Bourque, P. Fairley, R. E. (eds.), 3rd edn. www.swebok.org (2014). Accessed 30 Nov 2021
35. Kalluri, P.: Don't ask if AI is good or fair, ask how it shifts power. Nature **583**, 169–169 (2020)
36. Kiran, A.H., Oudshoorn, N., Verbeek, P.-P.: Beyond checklists: toward an ethical-constructive technology assessment. J. Responsib. Innov. **2**(1), 5–19 (2015). https://doi.org/10.1080/23299460.2014.992769
37. Larsson, S.: On the Governance of artificial intelligence through ethics guidelines. Asian Journal of Law and Society 7, no.3 (2020): 437-451
38. Lauer, D.: Facebook's ethical failures are not accidental; they are part of the business model. AI Ethics **1**(4), 395–403 (2021). https://doi.org/10.1007/S43681-021-00068-X
39. Marx, M.R.S.L.: Does technology drive history?: The dilemma of technological determinism. MIT Press (1994)
40. Metcalf, J., Moss, E., Boyd, D.: Owning ethics: corporate logics, silicon valley, and the institutionalization of ethics. Soc. Res. Int. Q. **86**(2), 449–476 (2019)
41. Metzinger T.: Ethics washing made in Europe. Der Tagesspiegel 8 (2019)
42. Mikalef, P., Gupta, M.: Artificial intelligence capability: conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. Inf. Manag. **58**(3), 103434 (2021). https://doi.org/10.1016/J.IM.2021.103434
43. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
44. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency, pp. 279–88. (2019). https://doi.org/10.1145/3287560.3287574

45. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**(4), 2141–2168 (2020). https://doi.org/10.1007/s11948-019-00165-5
46. Mougouei, D., Perera, H., Hussain, W., Shams, R., Whittle, J.: Operationalizing human values in software: a research roadmap. In: Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering (2018). Accessed 30 Nov 2021
47. Nickelsburg, M.: Privacy becomes a selling point for tech, with Apple and Microsoft leading the way (2019). Retrieved from: https://www.geekwire.com/2019/privacy-becomes-selling-point-tech-companies-apple-microsoft-leading-way/. Accessed 26 Mar 2019
48. NL AI Coalitie.: AI Oversight Lab: Ontwikkelen van Betrouwbare AI-Algoritmen Voor Publieke Instanties. https://nlaic.com/use-cases/ai-oversight-lab-ontwikkelen-van-betrouwbare-ai-algoritmen-voor-publieke-instanties/ (2020). Accessed 30 Nov 2021
49. NOS.: Anti-Fraudesysteem SyRI Moet van Tafel, Overheid Maakt Inbreuk Op Privéleven. 05-02-2020. https://nos.nl/artikel/2321704-anti-fraudesysteem-syri-moet-van-tafel-overheid-maakt-inbreuk-op-priveleven.html (2020). Accessed 30 Nov 2021
50. OECD.: Artificial Intelligence in society. Paris. https://ictlogy.net/bibliography/reports/projects.php?idp=3874&lang=en (2019). Accessed 30 Nov 2021
51. Orr, W., Davis, J.L.: Attributions of ethical responsibility by artificial intelligence practitioners. Inf. Commun. Soc. **23**(5), 719–735 (2020)
52. Reddy, E., Cakici, B., Ballestero, A.: Beyond mystery: putting algorithmic accountability in context. Big Data Soc (2019). https://doi.org/10.1177/2053951719826856
53. Rességuier, A., Rodrigues, R.: AI ethics should not remain toothless ! A call to bring back the teeth of ethics. Big Data Soc. (2020). https://doi.org/10.1177/2053951720942541
54. Rossi, F., Loreggia, A.: Preferences and ethical pri-orities: thinking fast and slow in AI. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems (AAMAS 2019). Montreal: AAMAS. http://www.ifaamas.org (2019). Accessed 30 Nov 2021
55. Ryan, M., Stahl, B.C.: Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J. Inf. Commun. Ethics Soc. (2020). https://doi.org/10.1108/JICES-12-2019-0138
56. Samoili, S., Cobo, M. L., Gómez, G., De Prato, E., Martínez-Plumed, F., Delipetrev, B.: AI watch—defining artificial intelligence. Towards an operational definition and taxonomy of artificial intelligence. In: Joint Research Centre (European Commission). Luxembourg: EUR 30117 EN, Publications Office of the European Union. (2020). https://doi.org/10.2760/382730
57. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., Dennison, D.: Hidden technical debt in machine learning systems. Adv. Neural. Inf. Process. Syst. **28**, 2503–2511 (2015)
58. Shneiderman, B.: Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans. Interact. Intell. Syst. **10**(4), 1–31 (2020). https://doi.org/10.1145/3419764
59. Shneiderman, B.: Human-centered artificial intelligence: three fresh ideas. AIS Trans. Hum. Comput. Interact. **12**(3), 109–124 (2020). https://doi.org/10.1775/1thci.00131
60. Stahl, B.C., Antoniou, J., Ryan, M., Macnish, K., Jiya, T.: Organisational responses to the ethical issues of artificial intelligence. AI Soc. (2021). https://doi.org/10.1007/s00146-021-01148-6
61. Steen, M., Van de Poel, I.: Making values explicit during the design process. IEEE Technol. Soc. Mag. **31**(4), 63–72 (2012). https://doi.org/10.1109/MTS.2012.2225671
62. Stix, C.: Actionable principles for artificial intelligence policy: three pathways. Sci. Eng. Ethics **27**(1), 1–17 (2021). https://doi.org/10.1007/s11948-020-00277-3
63. Trocin, C., Mikalef, P., Papamitsiou, Z., Conboy, K.: Responsible AI for digital health: a synthesis and a research agenda. Inf. Syst. Front. (2021). https://doi.org/10.1007/s10796-021-10146-4
64. Umbrello, S., van de Poel, I.: Mapping value sensitive design onto AI for social good principles. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00038-3
65. Van de Poel, I.:Translating values into design requirements. In Philosophy and engineering: Reflections on practice, principles and process pp. 253–266. Springer, Dordrecht (2013)
66. Veale, M.: A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. European Journal of Risk Regulation (2020): 1–10
67. Vogelsang, A., Borg, M.: Requirements engineering for machine learning: perspectives from data scientists. In: 2019 IEEE 27th International requirements engineering conference workshops (REW). IEEE (2019)
68. Wagner, B.: Ethics as an escape from regulation: from ethics-washing to ethics-shopping. In Being Profiled, pp. 84–89. Amsterdam University Press, 2018
69. Wood, G., Rimmer, M.: Codes of ethics: what are they really and what should they be? Int. J. Value Based Manag. **16** (2003)
70. Xenidis, R. Informal conversation during a team meeting, (2021)
71. Yeung, K.: A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework DRAFT. MSU-AUT (2018) 5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.