AI & SOCIETY
https://doi.org/10.1007/s00146-024-01880-9

OPEN FORUM

Challenges of responsible AI in practice: scoping review and recommended actions

Malak Sadek[1] · Emma Kallina[2] · Thomas Bohné[2] · Céline Mougenot[1] · Rafael A. Calvo[1] · Stephen Cave[2]

Received: 18 April 2023 / Accepted: 15 January 2024

Abstract

Responsible AI (RAI) guidelines aim to ensure that AI systems respect democratic values. While a step in the right direction, they currently fail to impact practice. Our work discusses reasons for this lack of impact and clusters them into five areas: (1) the abstract nature of RAI guidelines, (2) the problem of selecting and reconciling values, (3) the difficulty of operationalising RAI success metrics, (4) the fragmentation of the AI pipeline, and (5) the lack of internal advocacy and accountability. Afterwards, we introduce a number of approaches to RAI from a range of disciplines, exploring their potential as solutions to the identified challenges. We anchor these solutions in practice through concrete examples, bridging the gap between the theoretical considerations of RAI and on-the-ground processes that currently shape how AI systems are built. Our work considers the socio-technical nature of RAI limitations and the resulting necessity of producing socio-technical solutions.

Keywords Artificial Intelligence · Responsible AI · Participatory AI · Human-centered AI

Malak Sadek and Emma Kallina have contributed equally to this work.

✉ Malak Sadek
m.sadek21@imperial.ac.uk

Emma Kallina
emk45@cam.ac.uk

Thomas Bohné
tmb35@cam.ac.uk

Céline Mougenot
c.mougenot@imperial.ac.uk

Rafael A. Calvo
r.calvo@imperial.ac.uk

Stephen Cave
sjc53@cam.ac.uk

[1] Dyson School of Design Engineering, Imperial College London, London, UK

[2] Leverhulme Centre for the Future of Intelligence & Cyber-Human Lab, University of Cambridge, Cambridge, UK

Published online: 19 February 2024

1 Introduction

The recent rise of AI systems has been accompanied by public scandals such as privacy breaches[1] and systems amplifying bias.[2] Enhanced by the highly complex nature of AI systems and their prophesied disruptive impact, concerns regarding these negative effects of AI are multiplying. This is further exacerbated by the fact that AI systems are increasingly involved in critical decisions with a significant impact on people's lives. Some recent examples of these decisions include: deciding whether or not to detain criminal defendants (Dressel and Farid 2018); analysing which child protection requests seem credible (Chouldechova et al. 2018; Kawakami et al. 2022); in HR related topics (Mujtaba and Mahapatra 2019); facial recognition (Lohr 2018); detecting hate speech on social media (Modha et al. 2020); and significant decisions such as offering insurance (Ho et al. 2020) or a loan (Abuhusain 2020) to an individual. As a reaction to concerns about the misuse of AI, 'Responsible AI Guidelines' have been published, authored by Big Tech companies, academia and research institutes, as well as governments and NGOs. Responsible AI (in the following abbreviated RAI) describes AI systems that respect human rights and democratic values (OECD 2019). More than 80 of RAI guidelines have been made publicly available to ensure a future in which AI systems hold up against this standard (Jobin et al. 2019), and Gutierrez and Marchant (2021) identified more than 600 soft laws around RAI. RAI guidelines refer to a set of RAI principles which represent the different aspects of RAI, e.g. 'Privacy', 'Fairness and Non-Discrimination' or 'Accountability'. Studies comparing existing guidelines found that they are converging towards the same set of principles, even more in recent times (Jobin et al. 2019; Fjeld et al. 2020). This level of convergence suggests that we are arriving at a set of 'core principles'; which is currently the most favored approach towards principled RAI (Fjeld et al. 2020).

However, this increasing number of RAI guidelines has so far had little significant impact on the AI practice (e.g. McNamara et al. 2018). There is a general consensus regarding a substantial divide between the saturated space of theoretical AI ethics and the practical AI applications being developed today (Jobin et al. 2019; Munn 2023; Morley et al. 2021a). McNamara et al. (2018) showed that the mere presentation of RAI guidelines did not influence decisions of professional software engineers, as well as Computer Science students, across eleven software-related ethical decision scenarios. Remarkably, this gap between theory and practice is even recognised by the practitioners themselves (Ibáñez and Olmeda 2021). The following quote from an AI developer in an interview study by Ibáñez and Olmeda (2021) provides a glimpse into the current situation: "I think we read them all because they are coming out. There are many in the 'stratosphere'. That is when you read the principles and say, how do I translate them in practice? It gets more complicated." (Ibáñez and Olmeda 2021, p. 9).

Responding to this need, the following analysis presents possible reasons for this divide through highlighting shortcomings and gaps within RAI guidelines. By recognising and structuring evidence, this paper points out areas that limit the impact of current approaches on practices. Drawing on practical examples and theoretical suggestions, we cluster the identified problems into five areas: (1) the overly abstract nature of RAI guidelines leaving room for diverging interpretations, (2) the problem of identifying, prioritising, and aligning values, (3) the difficulty of operationalising success and impact metrics, (4) the fragmentation of the AI development process, and finally, (5) the lack of internal advocacy and accountability. In a second step, this paper examines existing recommendations on how to overcome each of these obstacles. Through examining less common approaches to RAI, we are attempting to gain an overview of current attempts across a variety of disciplines, approaches, and schools of thought. This allows us to gain a more comprehensive awareness of potential solutions, as well as to understand the relationships between interventions and existing challenges.

This paper differs from existing works in two ways. First, whilst critique on the principled approach to RAI itself is not novel, this paper summarises and clusters existing insights, including both theoretical suggestions as well as examples of concrete tools and practices. Currently, works discussing the limitations of RAI in general largely tend to focus on one or few RAI values such as explainability (e.g. Varanasi and Goyal (2023)); one aspect, e.g. as the need to move away from high-level approaches (such as RAI guidelines) when considering algorithmic fairness (John-Mathews et al. 2022); or on one domain, such as that of healthcare and bioethics (McCradden et al. 2020).

Second, while multiple works also mention recommendations to address explored limitations of RAI or RAI guidelines (e.g. Lee et al. 2021; Morley et al. 2021b; Chen et al. 2021; Peters et al. 2020), most solutions remain on a theoretical level with little guidance regarding their practical implementation. So far, only limited publications distinguish between higher level principles and concrete practices and focus on a singular domain or intervention (Harbers and Overdiek 2022). Accordingly, our work connects these philosophical critiques with practical examples addressing these limitations within AI-developing organisations. This is accomplished by highlighting examples of existing tools

[1] Such as facial recognition database company Clearview AI Inc compiling 20 billion images of people's faces without their consent.

[2] A recent example is the racist and sexist images created by the DALL-E Mini image generator.

Table 1 Keywords used to explore the discourse around each limitation covered

Limitation of current RAI interventions	Keywords used for searching
Abstract Principles	"RAI Principles" AND ("Abstract" OR "Generic" OR "High-Level" OR "Implement")
Narrow & Contradicting Values	"RAI Principles" AND ("Values" OR "Specific" OR "General" OR "Contradict" OR "Contrast" OR "Clash" OR "Universal" OR "Global" OR "Goal")
Lack of RAI Metrics	"RAI Principles" AND ("Metric" OR "Measure" OR "Impact" OR "Effect" OR "Success")
Fragmentation of the AI Pipeline	"RAI Principles" AND ("Fragmentation" OR "Labor" OR "Pipeline" OR "Workforce")
No Internal Advocacy and Accountability	"RAI Principles" AND ("Advocacy" OR "Internal" OR "Accountability" OR "Compliance" OR "Pressure")

that facilitate the implementation of each recommendation provided. Such an approach responds to the demand of Morley et al. (2021b) for a shift towards a more practice-based, participatory understanding of RAI to tackle existing limitations.

Furthermore, this work is a first attempt to complement the identified clusters with a review of recommended actions from various disciplines. In our work, we are reviewing existing, but hitherto scattered recommendations from different disciplines to arrive at a more comprehensive and multidisciplinary overview of alternative approaches.

The following section discusses the process followed to produce this paper, then rest of the paper outlines five major limitations of the current, principle-based approach to RAI. Afterwards, each limitation is matched with a corresponding set of recommendations from multi-disciplinary perspectives, with each including a practical example of how the recommendation can be operationalised.

1.1 Methods and process

The aim of this paper is to highlight the most salient critiques for RAI principles and begin a multidisciplinary discourse on possible solutions operationalised through concrete examples. As such, we've conducted a scoping review (Munn et al. 2018) given its ability to "identify knowledge gaps, scope a body of literature, clarify concepts [and] to investigate research conduct". We believe this aligns with our goals of compiling a range of recurring opinions and perspectives, understanding their arguments and scope, locating research gaps and limitations and finally identifying possible recommendations and corresponding practical examples. Given the evolving and expanding nature of the RAI space, we favored a more rapid approach to capture the state of the space in a rigorous but timely fashion (Sadek et al. 2023a).

Accordingly, this is not an exhaustive review of all existing critiques for RAI principles or their limitations. Instead, we started by using the key phrase "Responsible AI" AND ("Limitations" OR "Problems" OR "Gaps" OR "Issues") across academic databases (ACM Digital Library and IEEE Xplore) and industry-based and regulatory sources using a commercial search engine (Google Search) to systematically explore existing limitations, opinions and critiques. This method was used by similar reviews in the space of socio-technical considerations for AI systems (Sadek et al. 2023b) and yielded 466 results. We then excluded sources which did not mention RAI principles or did not mention an issue/problem/gap/limitation with RAI principles and collected the limitations mentioned across remaining sources (13 limitations). From those, we excluded limitations that were included in less than 10 sources (as our aim was to represent the most prevalent limitations) and those that were not clearly defined. This process resulted in the 5 limitations included in this paper. In order to map out these limitations and corresponding recommendations and practical examples we then used the keywords outlined in Table 1.

2 Limitations of existing RAI principles

AI systems have the potential to ease the burden of decision-making for humans and handle situations with more objectivity and efficiency, but there are several risks involved. While some of the factors are rooted in technical or technological domains (e.g. creating explainable and interpretable AI systems); many are also organisational (such as adhering to transparent practices); social and cultural, (e.g. adhering to ethical responsibilities above management goals); political or legal, (such as providing auditing or regulatory processes); or even span across different domains (Cognilytica 2021). Consequently, it seems crucial to guide the development of such systems. However, a chasm exists between the dominant, principle-based approach to RAI guidance and the practices and cultures of those implementing AI on the ground (Ibáñez and Olmeda 2021; Munn 2023; Jobin et al. 2019). Accordingly, in this section, we will explore five

bottlenecks preventing the seamless implementation of the RAI guidelines on-the-ground.

2.1 Limitation 1: abstract principles

The authoring bodies of responsible AI guidelines are very diverse, culturally as well as in their organisational forms. Thus, guidelines are motivated by various, often conflicting objectives (Jobin et al. 2019). This diversity of authors is even exceeded by the heterogeneity of current and future use cases of AI systems. To develop general-purpose guidelines, encompassing the full range of stakeholder needs and use cases, RAI principles have been pushed to an abstract level (Mittelstadt 2019), formulated as high-level, philosophical concepts such as 'Promotion of Human Values' or 'Fairness'. For example, if we review Microsoft's Responsible AI principles,[3] they state: "AI systems should treat all people fairly.", "People should be accountable for AI systems. " and "AI systems should empower everyone and engage people." Whilst this abstraction seems necessary to include a broad range of AI systems, it creates several complications when translating these principles into practice. Their broad definitions are not directly actionable since they leave room for diverging interpretations whilst offering little guidance for implementation and fulfilment. In fact, they are often criticised as not guiding any action at all (Munn 2023; Ayling and Chapman 2021; Mittelstadt 2019; Morley et al. 2021b; Hagendorff 2020; Krijger 2021; Whittlestone et al. 2019; Harbers and Overdiek 2022). Whilst higher-level constructs such as 'Fairness' are universal, their definitions mean different things to different people and they allow for various correct, but potentially conflicting interpretations (Mittelstadt 2019) or even mathematical definitions (Narayanan 2018). Enhancing this ambiguity, principles are often stated but not defined, leaving it to practitioners' discretion to define and adapt them as they see fit, defeating the purpose of offering guidance (Whittlestone et al. 2019). One of the few empirical studies performed on this topic found that reading ACM's code of RAI ethics had virtually no effect on ethical decisions of developers (McNamara et al. 2018). Similarly, other interventions originating from but downstream of RAI principles, such as design recommendations (e.g. design the system to be transparent), are also being viewed as too abstract and difficult to apply in practical situations (Elshan et al. 2022).

2.1.1 Variations in implementation and fulfilment

The issue of diverging interpretations is amplified given the resulting heterogeneous on-the-ground implementations. Due to the abstract nature of the principles, AI practitioners have to translate them into low-level tasks that are actionable for their specific use case (Mittelstadt 2019). Thereby, they have to consider a variety of social, cultural, legal, and political factors that differ across regions and contexts, as well as between individuals (Jobin et al. 2019; Felzmann et al. 2020; Chazette and Schneider 2020). Examples include the system's specific properties, its audience, the context of its application, the employees' own obligations and incentives, as well as the level of organisational commitment to RAI (Rakova et al. 2021). That imposes decisions with moral consequences onto AI creators (e.g. whether to choose equal opportunities for all individuals vs for groups (have varying baselines of opportunity, thus conflicting with individual fairness)); decision-making that lacks formal procedures and policies (Ibáñez and Olmeda 2021) and that AI creators might not necessarily be trained to make (e.g. in education, see Davies 2017). This shifts the responsibility—and accountability—from organisations to untrained practitioners (Mittelstadt et al. 2016; Morley et al. 2021b). Amplified by the varied professional backgrounds among AI practitioners, each with their own cultural, geographical, and linguistic norms and morals (Mittelstadt 2019; Fjeld et al. 2020), it is not surprising that the implementation of a single RAI principle can manifest in very different, heterogeneous directions. The implementations of RAI principles seem to span a spectrum (e.g. Bibal et al. 2021), whereby especially the private sector might be tempted to lean towards more shallow implementations as opposed to more radical measures in conflict with their business goals (Jobin et al. 2019).

The challenge with diverging interpretations of RAI principles leads us to the next issue: The challenge of whose values are incorporated, and conflicts between values in a specific context.

2.2 Limitation 2: narrow & contradictory values

Christian (2020) extensively discusses the challenges of determining whose values are embedded in an AI system (also summarised by Gabriel (2020)). This includes the decision of who chooses such values, especially in cases where business interests might conflict with the interests of users and other stakeholders. Conflicting schools of thought advocate for different approaches to addressing this question in an ongoing debate; a debate that is hitherto largely ignored by existing RAI guidelines. Nevertheless, values represented in current RAI guidelines are often found to be narrow and contradictory. Highlighting narrowness for example, up to April 2019 not a single guideline was authored by a body located on the African continent (Jobin et al. 2019), reflecting existing economic power hierarchies. In terms of conflicting values across RAI guidelines, an pertinent example is that respecting 'privacy' might prevent the provision of

[3] Accessible at: https://www.microsoft.com/en-us/ai/responsible-ai.

satisfying insights into an algorithm or training data, lowering 'transparency'. Similarly, an algorithm that maximises 'accuracy' is likely to systematically discriminate against minorities, compromising 'fairness' (e.g. Zicari et al. (2021); Whittlestone et al. (2019)).

2.2.1 Narrow set of values in mainstream RAI discourse

The global participation in drafting guidelines for AI systems is deeply imbalanced: The vast majority of guidelines originates or is based upon guidelines authored in the US or Europe (Jobin et al. 2019). As a result, they are tailored to the values and interests of their creators—and not to the values of the communities that are at the highest risk of being disadvantaged. Ingraining such a limited set of values in AI systems risks the creation of guidelines unsuitable for a variety of environments, peoples, and contexts (Sadek et al. 2023c).

2.2.2 Conflict with business goals

To be truly inclusive, the system must incorporate the values and viewpoints of all concerned stakeholders willing to provide input (Christian 2020; Gabriel 2020). However, without clear guidance on process and prioritisation, it is an impossible task to ensure that all stakeholder views or values are included equally. The organisation developing an AI-based system is naturally prioritising its own values and business goals, and even if a company's mission statement includes RAI principles, its implementation is often diluted until conflict with the business interests is minimised or removed (Rakova et al. 2021). The example of Meta's RAI principle 'Transparency & Control' illustrates this point: the company claims to undergo efforts to be transparent towards the user and offer control about data usage and the content that is displayed to them (Pesenti 2021). However, full user control would contradict their business goal of maximising a user's time spent on the app so that more ads can be presented to them. Thus, no actual, explicit controls for the user have been implemented that e.g. would enable them to evade topics that, whilst engaging them, 'suck them in' and thus distract or even compromise their well-being. Instead, 'Transparency & Control' is approached in a less impactful way such as the 'Why am I seeing this?' button (Pesenti 2021), more aligned with the company's business goals than with the user's preferences.

2.2.3 Conflicts between principles

Besides the tensions arising from diverging stakeholder values present in any single AI project, tensions between RAI principles themselves can arise through recommending conflicting actions (Mittelstadt 2019): Whittlestone et al. (2019) identified various pairs of principles that might contain such clashes of conflict. This creates value tensions between stakeholders (Whittlestone et al. 2019), and affects the principles themselves in unknown ways (e.g. for an example of trade-offs with explainability see Chazette and Schneider 2020). Since the RAI space lacks a universal hierarchy of principles that would resolve such conflicts, it is left to the AI creators to make these ethical choices (Mittelstadt 2019). Considering the weight of organisational norms in such decisions, it can be assumed that principles are prioritised in a way that conflicts the least with internal goals.

2.3 Limitation 3: lack of RAI metrics

As discussed above, operationalising abstract principles is a significant challenge. This abstraction of the principles makes it extremely difficult to measure their fulfilment or deficit. While some principles such as fairness can be formally quantified and measured, other principles such as inclusivity are more difficult to assess, including whether they have been achieved (Hagendorff 2020). It is also challenging to evaluate the overall impact of using RAI guidelines (Hagendorff 2020); One of the most common approaches of transforming requirements into checklists has been deemed as insufficient for deciding between public good and commercial interest (Zicari et al. 2021).

2.3.1 Lack of 'Soft' metrics

There is a lack of metrics or indicators that capture whether systems are having the intended impact on those using them and their communities. Mapping from high-level principles portraying desired outcomes to low-level, measurable metrics is highly challenging. For example, the number of clicks and likes on social media might reflect engagement, but how can user well-being be captured? This is exacerbated by the inscrutable nature of many AI systems and the lack of transparency in design practices (Mittelstadt et al. 2016). Thus, companies often fall back on techno-centric metrics, such as system performance or the click-through rate of users. Whilst choosing technical metrics to evaluate a system has the benefit of being more quantifiable and measurable, it overlooks human-centred factors such as the users' mental state and well-being that have socio-technical implications. Through the lack of a holistic measurement to assess a system's effects, insights such as the impact of different design decisions on a user's behaviour and thinking are lost Calvo et al. (2020). Unsurprisingly, many calls have been made to extend such solely technical focus during the assessment of potential harms of AI-based systems to include broader, socio-technical measures of system effects (McCradden et al. 2020; Chen et al. 2021; Harbers and Overdiek 2022).

2.3.2 Prominent metrics counteract RAI

Since RAI guidelines lack actionability and evaluation processes, companies easily fall back on known structures: Most currently used measures of success tend to focus on technical aspects revolving around performance, speed of delivery, and performance—not its underlying ethical standard (Scantamburlo et al. 2020). This is detrimental: Since implementing RAI might require a performance trade-off and slow the development process down, RAI is perceived as a threat to a system's success (Rakova et al. 2021). Thus, even if a company formulated a mission statement including RAI, the individual steps in the creation process are evaluated based on metrics that do not only ignore, but potentially counteract RAI principles. Employees have to choose between completing their step 'successfully' as defined by the company (i.e. fast and accurate), or adhering to the broader RAI mission and their own social responsibility (Rakova et al. 2021; Hagendorff 2020). This is exacerbated by the fact that AI engineers are not educated in including user needs (Peters et al. 2020). The following section analyses other issues causing unclear responsibilities along the RAI development process.

2.4 Limitation 4: fragmentation of the AI pipeline

As discussed above, the RAI guidelines lack a clear implementation process. As a result, the efforts to achieve RAI are distributed and diffused along the pipeline of AI development (Peters et al. 2020). Several teams spread over numerous geolocations can be involved in various activities across the AI pipeline, from data collection, data cleaning, model building, and so on. This pipeline is highly fragmented, rendering it extremely difficult to ensure that RAI guidelines and other standards are uniformly applied across the entire process. Additionally, AI practitioners and their working processes are often poorly integrated with existing business practices and the organisational structures (Peters et al. 2020).

2.4.1 Hidden labour

A main reason behind the fragmentation of the AI-development process is the immense number of humans involved along the AI pipeline (D'ignazio and Klein 2020; Gupta et al. 2022). This contradicts the marketing of technology and algorithms as *labor saving* (D'ignazio and Klein 2020)—created by a single, male genius AI engineer (Drage et al. 2022)—and *autonomous*. These myths are upheld through the invisible nature of human labour in AI development. Termed "hidden labour", these tasks are executed by a human workforce, but remain largely invisible in the public eye; e.g. data collection, cleaning and labelling (D'ignazio and Klein 2020). Whilst these steps are often taken for granted by subsequent data scientists and engineers, they have the potential to contribute to cascading biases and limitations that might be overlooked by the users of such data (D'ignazio and Klein 2020). Often based on freelance or gig work, these steps are opaque, extremely difficult to trace back, and nearly always lack documentation to do so (Morley et al. 2021a). As a result, a large amount of the workforce involved in developing an AI system can neither be contacted, educated, collaborated with, held accountable, or even traced. This structure renders it as unfeasible to ensure the prioritisation of data quality over data volume and to fully mitigate the biases in the data that cascade downstream to a model's eventual behaviour (West et al. 2019; Gupta et al. 2022). As a result, it becomes difficult to ensure that efforts to build RAI begin at these early stages.

2.4.2 Fragmented RAI implementation

Beyond the steps involving hidden labour, the design and development processes show fragmentation as well. They are performed by several individuals and teams whereby the development team is often slightly shielded (Schiff et al. 2020). Thus, it is difficult to assign responsibilities for implementing the RAI guidelines even on an internal level. The individual teams usually lack a clear person or team in charge of ensuring RAI (Rakova et al. 2021) and diffusion of responsibility takes place: employees perceive the responsibility for RAI practice as 'someone else's' task and no one is proactively working towards it (Schiff et al. 2020). Supporting this, Morley et al. (2021b) warns that this lack of clarity in practitioners' roles and responsibility, together with the ambiguous nature of ethical design could foster a culture of turning a blind eye. Demonstrating this, an interview study found that AI engineers across industry, open source, and academia often considered the questions posed in RAI guidelines to be outside of their agency, capability, or responsibility (Widder et al. 2022).

2.4.3 Inefficient task forces

BigTech companies such as Google, Microsoft or IBM created teams that are tasked with creating processes that promote RAI. Parts of their work are accessible on their websites, beautifully presented and coherent (e.g. IBM 2021; Google nown; Corporation 2022). However, the teams usually sit 'outside' of the usual AI pipeline, so their proposed interventions suffer from the same problems as other RAI guidelines: isolated from the overall company, they are not integrated into the development process. Jumping between abstract principles and specific tools (e.g. to measure the fairness of an algorithm) they

remain either diffuse or techno-centric, failing to spark deep reflection on a system's effects. This is enhanced by the fact that these teams often encounter revenue heavy (legacy) systems that cannot be challenged. A prominent example of this unwritten rule is Google's dismissal of Timnit Gebru after she published a paper about the danger of NLP models—an area Google heavily invests in Simonite (2021). This fragmentation of efforts leads to our last cluster of problems: The lack of internal advocacy for RAI.

2.5 Limitation 5: no internal advocacy and accountability

The previous sections described the obstacles practitioners face when attempting to implement RAI for a specific use case. These difficulties are exacerbated by the agile working style common in tech companies, focusing on fast delivery without space for reflection (Peters et al. 2020). There is a clear lack of harmony and embedding between RAI guidelines and existing business structures and practices (Chen et al. 2021; Rakova et al. 2021). As a consequence, there is little evidence that the majority of companies are proactively engaging in RAI (e.g. self-seeking certification or training practitioners) beyond merely reacting to legally binding regulations and technical fixes (Crawford and Calo 2016). These regulations in turn are outpaced by the fast developments of AI systems and are currently nearly exclusively voluntary mechanisms that lack accountability (Gutierrez and Marchant 2021).

2.5.1 Passive reaction to external pressure

This lack of internal advocacy is illustrated by the findings from Rakova et al. (2021) which showed that companies' current efforts in RAI are mainly motivated by external pressure; either regulatory or through fear of scandals and negative public attention (Rakova et al. 2021). This focus on external obligations causes the dangerous misperception that only RAI principles that are reflected by laws are worth implementing. In a global survey, AI practitioners prioritised the same two RAI principles across twelve vastly different use cases (Falk et al. 2020): 'Privacy and Data Rights' and 'Cybersecurity'. Both are reflected in regulations, namely the GDPR. This issue is even increased in scope when AI companies perceive that ensuring RAI is not their own responsibility, but the duty of regulators (see the 'Many Hands Problem' in Schiff et al. 2020). Companies are not used to questioning the social impact of the systems they develop, so there is no feeling of social responsibility—and frankly no legal obligation to do so. Considering the sub-par regulation attempts hitherto, this is detrimental (Morley et al. 2021b; Mittelstadt 2019; Chen et al. 2021).

2.5.2 Unreflected compliance mindset

This passive reaction to external pressure causes a second problem, namely the tendency to fall into a 'compliance ethics' mindset. Including the perception of simply having to abide by externally-set regulations or rules to obtain a 'stamp of approval', genuine considerations of the active respecting and protection of stakeholder values are omitted (DigitalCatapult 2020). This effectively avoids deeper reflections on the issues being highlighted, also referred to as 'deploy and comply' (Crawford and Calo 2016). Since active considerations are missing, it is overlooked that the importance of fulfilling a certain RAI principle is context-dependent and certain pillars are more important than others in a particular use case (Fjeld et al. 2020). For instance, a use case involving minors or mental health data should fulfil the principle 'Privacy' to a higher level than a use case that only includes data from official LinkedIn profiles. Recent studies have shown that practitioners prioritise different RAI principles to policymakers in a given context (Falk et al. 2020) and prioritise different values as compared to a consensus representing the general public (Jakesch et al. 2022). This gap might existing gaps on a broader scale: Whilst policymakers and the general public focus on society as a whole and long-term developments, companies mainly focus on maximising their immediate productivity and revenue whilst complying with external regulations.

2.5.3 Lack of mechanisms for providing accountability

Transparency about a model's creation and operation is an essential mechanism to enable accountability. A lack of transparency prevents accountability since it becomes impossible to identify the human decision that caused harm, or whether to blame the machine itself (Cherubini n.d.; Hemment et al. 2019; Wachter and Mittelstadt 2019). This can harm AI creators themselves: Wagner (2022) speaks of "broken data stories" when a lack of transparency leads to a silo-ing of knowledge, a compartmentalization of data, and a lack of rapid prototyping capabilities, that are significantly stifling innovation and damaging the ability to adapt and respond quickly to emergent trends. In fact, transparency can benefit AI creators through providing clear justifications of the individual steps of system development (e.g. which stakeholder inputs caused which development decision), insulating them from criticism of decisions outside of their responsibility (Antonic 2021; Barker et al. 2023).

The main mechanism for providing transparency—and thus accountability—is documentation (Barker et al. 2023; Custis 2021; FAIR 2021). One way to provide documentation is through initiatives such as algorithmic impact assessments (Reisman et al. 2018; Edwards et al. 2016; Floridi 2018), which serve several other purposes as well; many of which relate to accountability. Algorithmic impact assessments are a tool that aims at assessing possible societal impacts of an AI system, often through a questionnaire about aspects of the system's development and operation, e.g. details about the underlying data set, development practices, or the use case. However, these assessments are very distant or inconsiderate of the real-world impact of different stakeholders and communities (Metcalf et al. 2021). Thus, these assessments can turn into checklist-style obligations that are conducted hastily out of necessity, instead of inspiring genuine considerations (Metcalf et al. 2021). This is exacerbated by the fact that these assessments assign the power to decide the scope of the assessment, whose interests to consider, who to allow to participate, who to give access to the report, and how the assessment informs later work to organisations and decision-makers. This perpetuates unequal power dynamics and thus unbalanced participation patterns, leading to impacts being overlooked or not properly addressed, especially those that can occur during phases different to when the assessment is taking place. Thus, they largely fail to increase organisational accountability for the public (Moss et al. 2021).

3 Recommendations

After discussing salient limitations of RAI guidelines from recent literature, we will now consider recommendations on how to address these. The following section presents remediations for the issues identified above, stemming from various disciplines, such as Data Feminism, Media Studies, Design Engineering, Digital Humanities, Science and Technology Studies, and Human–Computer Interaction. These expand the currently dominant, principle-based approach through other, less commonly applied perspectives. Thereby, we attempt to create a more encompassing perception of possible paths to RAI, including a shift in focus from RAI theory towards its practical implementation, responding to calls of Mittelstadt (2019) and Morley et al. (2021b). Supporting this, we anchor each of the listed recommendations in practice through examples of existing tools and methods that support their implementation for a specific use case.

3.1 Solution 1: participatory interpretation

Highly abstract RAI principles allow for diverging interpretations and implementations, demanding more specific guidance to translate them to a specific use case. However, the diversity of AI applications and their contexts render one-fits-all, 'silver bullet' solutions impossible. Especially the techno-centric toolkits that "hungry methodologists" (term by Keyes et al. (2022)) created to 'solve' RAI, fail to account for the pluralistic requirements of a specific application context (D'ignazio and Klein 2020). For example, a system can be transparent, fair and accountable whilst operating in an ethically questionable domain (Keyes et al. 2019). More technology is unlikely to solve the problems that technology created. Instead, the solution of the abstraction problem might lie in the participation of the affected communities as claimed by the philosopher John Tasioulas (Tasioulas 2021, 2022), the Digital Humanities professor and data feminist Lauren Klein and her colleague Catherine D'Ignazio (D'ignazio and Klein 2020), as well as the technology design researcher Dorian Peters (Peters et al. 2020): Instead of leaving the interpretation of the abstract RAI principles to practitioners, it is essential to collect insights from the users and communities in which an AI system is embedded (e.g. D'ignazio and Klein 2020; Allcott et al. 2020; Gaggioli et al. 2017). Its users and other stakeholders should inform the context-specific meaning, desired level, method, design, and implementation of an RAI principle. Through enabling practitioners to understand the context in which the system will be applied, their focus is directed toward the specific needs and priorities of their stakeholders. This seems especially true for perception-based RAI principles such as 'Transparency & Explainability' or 'Human Control' (versus more regulatory aspects such as 'Accountability'). Such an approach is in line with the goal to create human-centred, ethical systems through fostering empathy—rather than promising a panacea.

Practical example. The 'Unbias Fairness Toolkit' developed by Proboscis (Lane et al. 2018)[4] structures the dialogue with external stakeholders. The toolkit aims at stimulating public dialogue and engagement with the long-term goal of achieving a critical and collective approach to imagining the future of communities (Lane et al. 2018). Through reflective exercises and thought experiments, participants gain awareness about their preferences and values in relation to a specific system. The toolkit contains different tools such as worksheets and cards that facilitate the discussion and explicit statement of the participants' preferences regarding a system, used data types, essential values, and its application context. These exercises help in surfacing the concrete meaning of the abstract RAI concepts for the affected stakeholders. Especially the 'Participant Value Perception Sheet' on which participants can group the cards under different clusters such as 'opportunities', 'costs', or 'motivations' make these very explicit. The gained insights should inform

[4] The toolkit can be viewed and downloaded at: http://proboscis.org.uk/5970/unbias-fairness-toolkit/.

the translation of the abstract RAI principles to the specific use case to ensure that the preferences of affected communities are reflected.

3.2 Solution 2: adopting wider perspectives

Similarly to shifting from abstract values towards a feeling of responsibility and a detailed understanding of the use case, we recommend expanding the exclusive focus on internal interests towards a more holistic assessment of the system's impact on its environment (Crawford and Calo 2016; D'ignazio and Klein 2020). While similar to the first solution in the sense that adopting wider perspectives can benefit from an increase in diverse participation, the key difference here is that for this solution we call for a consideration of the wider socio-technical contexts in which an AI-based system is embedded. All the above describes a shift in the mindset and training of technical practitioners themselves which can be aided by, but does not necessitate, increased stakeholder participation. Such a socio-technical approach includes considering the broader societal contexts in which the system will exist and operate, beyond the individuals directly affecting management goals (D'ignazio and Klein 2020). Such a "society-in-the-loop" mindset requires active, ongoing participation of the communities affected by the system. We are aware that this will add a considerable, but necessary workload. In fact, stakeholder engagement seems to be required to fulfil aspects of upcoming regulations: The EU AI Act, for instance, requires that "due consideration shall be given to the technical knowledge, experience, education, training to be expected by the user and the environment in which the system is intended to be used".[5] Stakeholder engagement is a proven method to understand the context of system deployment as well as user characteristics (Lee et al. 2019; Young et al. 2019). This relationship is acknowledged by emerging standards. The ISO/IEC DIS 42001 standard detailing the management processes around AI development[6] for example, includes stakeholder involvement as a method to fulfil aspects of upcoming regulations. Moreover, various organisations in the AI policy space are calling for stakeholder engagement along the AI lifecycle. Examples are multi-stakeholder organisations such as the Alan Turing Institute (e.g. Leslie 2019, p. 27), HAI Stanford (e.g. Elam and Reich 2022, p. 7), or the ACM Technology Policy Council (e.g. Office 2022, p. 6), as well as civil society organisations (e.g. Amnesty International (e.g. Bacciarelli et al. 2018, p. 8)) or inter-governmental organisations, for instance the WHO (e.g. Organization et al. 2021, p. 16), UNESCO (UNESCO 2021, p. 23), and OECD (Observatory 2018, see Principle 1.1,). Advocating for increased stakeholder engagement reacts to Mittelstadt (2019)'s push for the use of participatory and value-sensitive design outside of academia and into the industry space. Through collaborating with external stakeholders, their priorities, as well as their goals, fears, and wishes in the system's application context can be understood (Gaggioli et al. 2017). These insights expand traditionally considered values with more human-centred aspects, such as pleasure, psychological and physical well-being, autonomy, welfare, connectedness, or privacy (Gaggioli et al. 2017). These dimensions have the potential to vastly improve systems, counteracting the common perception of technical experts that participatory technology design is an act of service or charity (D'ignazio and Klein 2020). Respecting and promoting these values throughout the system's development and design process contributes to a responsible, value-sensitive solution (Rahwan 2018). To reach such a state, it must be recognised that people beyond the traditionally recognised 'AI experts' have valuable inputs and contributions to inform the design of these technologies (Tasioulas 2021; Yang et al. 2020). The values of the entire ecosystem must be considered before any steps are taken—including the hierarchies and intersectionalities involved (Friedman and Hendry 2019; Fuchsberger et al. 2012; Costanza-Chock 2018).

Data Feminism formulates core principles that emphasise necessary considerations: First, "elevate emotion and embodiment". This emphasises the importance of valuing input from individuals as "living, feeling bodies in the world" (D'ignazio and Klein 2020, p. 2, chapter 3)—beyond their expertise in computer science. Second, "consider context" advocates for examining a system's environment and to understand the values that motivate the system's creation (D'ignazio and Klein 2020). "Embracing pluralism" whilst doing so produces the most complete knowledge from synthesising all perspectives present and advocates for non-dominant forms of information (D'ignazio and Klein 2020). D'ignazio and Klein (2020) recommend investigating power structures in the given use case (principle 'examine power'). Only then, can we challenge these power structures that influence whose values are dominantly advocated for and whose values are typically overlooked, thus ultimately contributing to a more equal inclusion of viewpoints (D'ignazio and Klein 2020). Such process includes reflexivity—the constant self-examination and reflection regarding one's own preconceptions, values, or biases (D'ignazio and Klein 2020).

Practical example. The initiative 'Ethics for Designers'[7] aims at equipping designers with three skills to design

[5] Article 9.4, Page 47 of the EU AI Act: https://artificialintelligenceact.eu/.

[6] The standard can be accessed online at: https://www.iso.org/standard/81230.html.

[7] Full toolkit is accessible at: https://www.ethicsfordesigners.com/tools.

responsible technology that go beyond the traditional, revenue-centric perspective: (1) moral sensitivity to recognise the ethical implications of the project, (2) moral creativity to explore solutions to moral challenges, and (3) moral advocacy to intentionally set their moral position to other stakeholders. A toolkit offers tools for each of these skills. For example, moral sensitivity is supported by canvases that deconstruct the values underlying existing designs and prompt the consideration of the ethical terms for a specific project. Moral creativity is facilitated by a game that motivates a broader consideration of solutions to ethical challenges in the design of a specific system. The tools for moral advocacy include stakeholders to understand their values, how they are prioritised by the impacted communities, and how the system might impact them. An 'ethical contract' guides the negotiation of the overall prioritisation of values to find a common ethical ground. Thus, 'Ethics for Designers' offers tools that prompt teams to consider the socio-technical context beyond traditional aspects through including stakeholders and reflect on values present in the use case. Such considerations can advance our knowledge regarding success metrics that are not merely focused on the technical functioning of the AI system, a topic discussed in the next section.

3.3 Solution 3: operationalizing RAI metrics

This section considers the challenge of translating highly theoretical RAI principles that are difficult to quantify or even observe into measurable outcomes is a considerable challenge (Thomas and Uminsky 2020; Stray 2020; Jacobs 2021). The derived metrics have to be concrete and observable, e.g. the number of clicks or likes. Basing such metrics on the RAI principles is essential since the dominant, techno-centric metrics neither resemble the actual user experience, nor the effects and implications of a system's socio-technical context (Stray 2020; Thomas and Uminsky 2020; Jacobs 2021). As in the previous sections, this challenge could be overcome by engaging relevant communities. The process of doing so for a planned AI system can be outlined in four steps, adapted from Stray (2020):

(1) Identify the relevant stakeholders / affected communities and collaboratively define their values and goals in the context of system use.
(2) Select relevant metrics from the stakeholders' inputs, supported by existing value-sensitive design frameworks.
(3) Harness the selected metrics as managerial performance measures and/or the system's success criteria so that the system's fulfilment of stakeholder values can be assessed.
(4) Evaluate the results based on the chosen metrics, supported by ongoing qualitative data collection; adjust accordingly.

This process shifts the focus in order to establish the establish the relevant metrics more collaboratively with the understanding and approval of external stakeholders. Since it is impossible to initially predict the goals of different users, it is important to avoid setting concrete targets from the outset. Instead, they should be informed by the identified goals of relevant stakeholders (Stray et al. 2021). Thomas and Uminsky (2020) recommends to constantly re-assess and adjust chosen metrics to guarantee that they stir the system towards the desired user experience or societal outcome (Thomas and Uminsky 2020, 2022).

Through engaging stakeholders, practitioners can not only understand which RAI principles are perceived as important in the application context, but also how these are operationalised, evaluated, or 'measured' by the communities themselves. In other words, it advances our understanding of which aspects have to be fulfilled by a specific system so that a specific community defines it as e.g. 'fair' in a specific context. Unsurprisingly, the metrics chosen by external stakeholders reflect their values as discussed above, going beyond techno-centric measurements and consider the system effect on users' well-being or behaviour (Calvo et al. 2020). Through including such 'soft' metrics, various benefits can be unlocked: New use cases, constraints, as well as functionalities might be discovered, informing prevention strategies that avoid risks, as well as promotion strategies that offer protection and support (Calvo et al. 2020). For example, the negative well-being effects of recommender systems used in social media platforms were only discovered through qualitative measurement of the user experience since the quantitative click-rate remained high (Stray et al. 2021). Furthermore, the inclusion of stakeholders increases the transparency of system development—enhancing procedural fairness, trust and credibility—since stakeholders understand where different metrics and decisions originate from (Stray 2020). More recently, tech giants such as Meta and Google started to incorporate well-being metrics into their system assessments (Stray 2020; Gaggioli et al. 2017), but there is still much work to be done in terms of selecting such human-centred AI metrics.

Practical example. While not available as a toolkit as of yet, the IEEE 7000's process (IEEE 7000-2021 2021) could be followed to operationalise value-based metrics. The IEEE 7000 employs a "value-based engineering" or "value by design" approach where it moves away from checklists, impact assessment, and overly generic value lists (Spiekermann and Winkler 2020; Spiekermann 2021) towards a more process-based approach. It expands beyond the risks of neglecting values and modifies the non-functional

requirements engineering process to lead to more value-sensitive and human-centred outcomes (Spiekermann and Winkler 2020). Through supplementing traditional functional requirements with a set of risk-controlled value requirements resulting from initial, abstract stakeholder values, a more holistic design specification can be achieved (Spiekermann and Winkler 2020). These value requirements, akin to traditional design requirements used for building various systems, can then be evaluated and checked against more concretely than trying to use abstract stakeholder values as metrics, allowing for a clearer evaluation of attempts to produce RAI. The process shares multiple similarities to a typical human-centred design process, but is entirely focused on values. Adherence to this process contributes to the identification of ethical "value bearers" (Spiekermann and Winkler 2020): When applied to a practical case study, the process was found to lead to the identification of more stakeholders, more product features, more values across more value classes, and more implications and benefits of the product than a classical product roadmapping process (Bednar and Spiekermann 2021).

3.4 Solution 4: pluralism and transparency

The first step in overcoming the challenges caused by the fragmentation of the AI pipeline is to acknowledge and accept this fragmentation as well as the unsuspectedly large amounts of human labour involved (D'ignazio and Klein 2020). Several projects aim at increasing this awareness: Kate Crawford's work 'Atlas of AI' emphasises the hidden costs, influences, and biases of AI systems (Crawford 2021), the website 'The Anatomy of an AI System'[8] visualises the numerous stages and people involved in the creation of an AI system (including the work of extracting rare minerals for computer chips), and the illustration 'Nooscope' displays the various biases and errors that can arise throughout the entire machine learning pipeline (Pasquinelli and Joler 2021). Such educational resources help teams of AI creators to envision and be conscious of their development pipeline in its entirety. Through emphasising that data sets and AI systems are not neutral, but instead made by human actors, the human assumptions that have been embedded through this process can be identified and questioned (D'ignazio and Klein 2020). These reflections should span the environment in which the data set or model was created, their historical context, as well as the intended use case, each with their underlying biases and hierarchies (D'ignazio and Klein 2020; Koesten et al. 2021; West et al. 2019).

Echoing numerous scholars, we advocate for transparent, and openly communicated data production processes, including who collected it how, where, and when (including the dominant discipline and framework), how representative it is of the world and its potential to reinforce existing biases (D'ignazio and Klein 2020; Koesten et al. 2021). Besides transparent data sets, we require transparent model development processes that clearly state the positionalities of researchers, data-related activities, existing uncertainties and assumptions, as well as the hidden steps and human labour that lay along the AI pipeline. Only then, biases and problematic preconceptions can be challenged, especially by the communities affected by it (D'ignazio and Klein 2020).

This would aid the allocation of resources for reflection and external stakeholder involvement to examine the fit of a data set with the current and desired context of its resulting application. Such sentiment has been echoed by other experts in the field, such as John Tasioulas' call for the three P's of Pluralism, Procedures (focusing on those over outcomes), and Participation (Tasioulas 2022). In more recent years, the field has led to the creation of movements such as Feminist.AI who work on designing interventions and activities to engage broader communities in the design process of AI systems, focusing on those lying at intersectionalities. These interventions can be considered as promoting 'transparency': We require clear communication about the positionalities of researchers, data-related activities, existing uncertainties and assumptions, and the hidden steps and human labour that lay along the AI pipeline. Only then, biases and problematic preconceptions can be challenged, especially by the communities affected by it (D'ignazio and Klein 2020).

Practical example. Gebru et al. (2021) proposed 'datasheets' for data sets; a form of documentation that details the motivation for the creation and composition of a data set, its collection process and recommended use cases. The Data Nutrition Project[9] aims at making this approach more practical by creating a label—similar to a nutrition label—that summarises key facts about a data set. The facts include meta-data and populations, unique or deviating features regarding (demographic) distributions, missing data or variables, comparisons to other data sets, and its intended use case with associated alerts and red flags.

Both tools increase the awareness of the human labour that went into a data set's creation. Since they provide details about the data set's composition and the context of its creation, AI practitioners can use them to ensure a high fit between a data set and their intended use case. We recommend favouring data sets that are accompanied by a datasheet or data nutrition label. Additionally, we hope to motivate the creation of datasheets or the obtaining of a data nutrition label for existing and newly collected data sets. FactSheets describe relevant information at each phase

[8] Accessible at: https://anatomyof.ai/.

[9] For more information, please refer to their website: https://datanutrition.org/.

Table 2 Summary of reviewed RAI interventions' limitations and corresponding recommendations

Limitation of current RAI interventions	Corresponding recommendation
Abstract Principles	Participatory Interpretation
Narrow & Contradicting Values	Adopting Wider Perspectives
Lack of RAI Metrics	Operationalizing RAI Metrics
Fragmentation of the AI Pipeline	Pluralism & Transparency
No Internal Advocacy and Accountability	Emphasising Practitioner Responsibility

of the model's development: pre-training, during training, and post-training. Explainability Fact Sheets summarise key features to make model more explainable. Model Cards describe how a model was developed, including who trained the model, the timeline of the training, which training data was used, and details of model development and performance

Several projects aim at increasing the awareness of the hidden steps in the AI development process. For example, Kate Crawford's work 'Atlas of AI' which emphasises the hidden costs, influences, and biases of AI systems (Crawford 2021). A second example is the website 'The Anatomy of an AI System' that visualises the numerous stages and people involved in the creation of an AI system, including the work of extracting rare minerals for computer chips. Taking this further, the illustration 'Nooscope' displays the various biases and errors that can arise throughout the entire machine learning pipeline. These limitations are framed socio-technically, commenting on training data sets being curated in hidden labour with strong cultural influences and the related misperception of objective automation. Such educational resources can help teams of AI creators to map their pipeline in its entirety and raise awareness of various data provenance issues. This enables them to consider entry points for biases and other ethical risks along the complete process. Only then, mitigation efforts can be targeted at more hidden steps, including the recognition and assignment of responsibilities.

3.5 Solution 5: emphasising practitioner responsibility

Returning to the importance of fostering empathy over compliance, such an approach can also help address the lack of internal advocacy for RAI. A major part of creating human-centred, ethical systems is to foster a feeling of responsibility for the system's consequences in the people creating them, moving away from external regulation and towards intrinsic motivation (Gaggioli et al. 2017). If practitioners would be aware of the real-world impact of their work on society, they are likely to be more inclined to avoid the creation of harm. Such awareness might increase their motivation to engage in discourse and practices which ensure better outcomes for all stakeholders. For example, Yildirim et al. (2023) found that practitioners used a resource advising on RAI practices (the Google + AI Guidebook) not only to build RAI, but also as an educational tool, to communicate issues to elicit buy-in, as well as to develop their own resources.

Organisations that explicitly assign such responsibility internally can equip practitioners with an ever-present toolkit of active reflection and empathy, counteracting unhelpful practitioner cultures that include "rejecting practices or downplaying the importance of [stakeholders'] values or the possible threats of ignoring them" (Manders-Huits and Zimmer 2009). Design interventions are capable of changing cultures (Ozkaramanli et al. 2016) and can be utilised for helping practitioners to shift away from mindless compliance and towards an internal sense of responsibility or "value absorption" (Garst et al. 2022). The shift away from checklists and guidelines and towards more participatory methods with increased exposure to stakeholders is likely to embed such a feeling of empathy in practitioners across some domains (Holden 2018).

Practical example. AIxDesign's AI Meets Design Toolkit[10] seamlessly combines design activities with ML/AI knowledge to integrate both processes across various stages of the AI pipeline. It contributes to internal advocacy by involving a variety of company-internal stakeholders from various disciplines and departments. To enable the participation of non-technical roles, a crash course on AI/ML is provided. It's numerous tools ensure that multi-disciplinary factors beyond management goals—such as user needs, relevant research, and data availability—are considered from the outset. To anchore these considerations throughout the process, several templates prompt the team to evaluate system ideations based on desirability, feasibility, value proposition and value polarity. This helps the team to identify tensions, whilst an additional tool aids the anticipation of unintended consequences. The resulting considerations are formalised in plots focusing on objectives, inputs, features (factors), and outputs (labels) to ensure that technical and non-technical stakeholders have matching mental models. Overall, this toolkit ensures that the entire team is aware of the ethical risks and tensions involved in the context of the system development. Such awareness increases the internal advocacy for and championing of measures to mitigate

[10] Toolkit is fully downloadable at: https://www.aixdesign.co/toolkit.

these risks (Yildirim et al. 2023). Internal advocacy is a step towards better aligning business goals with RAI outcomes and facilitating the needed attitude and cultural shifts away from compliance and towards empathy.

3.6 Conclusions and discussion

Table 2 highlights all mentioned RAI intervention challenges and their corresponding recommendations.

This work provides a review on two levels: First, we discuss limitations of current attempts to embed ethics into AI systems. This perspective steps away from more prevalent surveys of these attempts, often with a philosophical stance, and instead focuses on several socio-technical challenges of their implementation in practice. Second, we recommend counter-actions to the issues identified stemming from various disciplines, illustrated with examples of practical tools. While by no means exhaustive, these recommendations aim to open up a dialog regarding potential solutions for the reviewed issues, paving the way for future work by highlighting opportunities to address the challenges mentioned.

Current attempts to embed ethics in AI-based systems were found to be mostly too abstract, leaving room for interpretation in several instances. This gap could be tackled through participatory approaches to create shared interpretations, lending acceptance and legitimacy by involving key communities and stakeholders. The lacking clarity about whose values have to be included in current approaches to RAI leads to a prioritisation of values that align with business goals. Therefore, an adaptation of various socio-technical perspectives is required to harmonise between attempts to embed ethics in AI systems and the values and considerations of those affected by them. Furthermore, the fulfilment of aspects of RAI cannot be measured. To overcome this, success metrics have to be operationalised with stakeholders, contributing to more holistic and impactful interventions. Current attempts also fail to account for organisational challenges, such as a fragmented pipeline for AI delivery and the lack of internal advocacy for RAI interventions. To tackle these challenges, a culture shift is required from a passive compliance mindset and towards pluralism, transparency, and empathy. While we provide practical examples of tools and methods to operationalise recommendations, we must acknowledge that these examples themselves largely lack evaluative studies and so their effectiveness has not been formally assessed. Instead, the sentiment is that the practical operationalisation of recommendations to overcome the established limitations of RAI guidelines (of which a significant one is the lack of practical implementability of these guidelines) is in itself a helpful step in the right direction.

While this study is not empirical in nature and cannot claim to be exhaustive, it aims to inspire future work to frame RAI limitations in a more socio-technical light, looking beyond technical, mono-disciplinary aspects and towards more holistic, practical, and participatory interventions. It must be noted that increasing participation in the space of AI is difficult and lacks well-defined process guidance. Participatory design is challenging on its own, and the AI space adds further obstacles: Users surrender due to technical complexity, hype generates unrealistic expectations, as well as the lack of process and methods to define RAI requirements with external stakeholders. Further studies are required to make human-centred design more accessible to practitioners.

Previous works have shown that several AI practitioner needs are unmet due the shortcomings of RAI guidelines (Morley et al. 2021b; Yildirim et al. 2023). By providing recommended actions and suggesting longer-term shifts in perception, we aim to mitigate these shortcomings. Despite the plethora of existing definitions of what RAI should entail, given our investigations we conclude that:

- **RAI should be value-sensitive.** Learning from the field of Value-Sensitive Design Friedman et al. (2002), RAI should consider the values and priorities of key stakeholders who own, use, and are affected by these systems. These values should guide the system design from the outset.
- **RAI should be inclusive and participatory.** In order to be truly responsible, AI systems need to account for their diverse range of stakeholders. Their participation is essential to ensure that they have a meaningful role and decision-making power throughout the process of designing, developing and implementing AI systems (D'ignazio and Klein 2020; for Coordination 2019; Leslie 2019). This requires an environment in which diverse stakeholders can feel empowered to make changes and actualize their visions (Delgado et al. 2023, 2021; D'ignazio and Klein 2020).
- **RAI should be human-centred.** Responsible AI has to take human factors into account to ensure that resulting systems affect the socio-technical space positively and meaningfully. This includes the acknowledgement that technologies are not neutral artefacts, but instead influence, and are influenced, by the communities around them (D'ignazio and Klein 2020).

This study highlighted the obstacles to the implementation of RAI guidelines and pointed out recommended actions. It is crucial that further research sheds light on methods and processes that increase the accessibility of these recommendations in practice. Nevertheless, it remains an essential step that practitioners take responsibility and harness these

recommendations to broaden their problem-solving skills beyond the technical spheres and into wider, socio-technical spaces in which they have arguably more profound influence.

Funding This work is partially funded by the Leverhulme Trust through the Leverhulme Centre for the Future of Intelligence [Award Number: RC-2015-067]. This work was also supported by The Alan Turing Institute's Enrichment Scheme.

Data availability Data sharing not applicable to this article as no data sets were generated or analysed during the current study.

Declarations

Conflicts of interest The authors declare that there are no conflicts of interest present while producing this work.

References

Abuhusain M (2020) The role of artificial intelligence and big data on loan decisions. Accounting 6(7):1291–1296

Allcott H, Braghieri L, Eichmeyer S, Gentzkow M (2020) The welfare effects of social media. Am Econ Rev 110(3):629–76. https://doi.org/10.1257/aer.20190658

Antonic J (2021) How to foster advocacy for digital transformations through collaboration. Presentation at World Interaction Design Day (IxDD) and can be accessed at: https://vimeo.com/619232039

Ayling J, Chapman A (2021) Putting ai ethics to work: are the tools fit for purpose? AI Ethics 2(3):405–29

Bacciarelli A, Westby J, Massé E, Mitnick D, Hidvegi F, Adegoke B, Kaltheuner F, Jayaram M, Córdova Y, Barocas S, Isaac W (2018) The toronto declaration: Protecting the rights to equality and non-discrimination in machine learning systems - amnesty international. https://www.amnesty.org/en/documents/pol30/8447/2018/en/. Accessed on 18 Aug 2023

Barker M, Kallina E, Ashok D, Collins K, Casovan A, Weller A, Talwalkar A, Chen V, Bhatt U (2023) Feedbacklogs: Recording and incorporating stakeholder feedback into machine learning pipelines. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp 1–15

Bednar K, Spiekermann S (2021) On the power of ethics: how value-based thinking fosters creative and sustainable it innovation. WU Vienna University of Economics and Business (Working Papers/Institute for IS & Society)

Bibal A, Lognoul M, De Streel A, Frénay B (2021) Legal requirements on explainability in machine learning. Artif Intell Law 29(2):149–169

Calvo RA, Peters D, Cave S (2020) Advancing impact assessment for intelligent systems. Nat Mach Intell 2(2):89–91

Chazette L, Schneider K (2020) Explainability as a non-functional requirement: challenges and recommendations. Requir Eng 25(4):493–514

Chen J, Storchan V, Kurshan E (2021) Beyond fairness metrics: roadblocks and challenges for ethical ai in practice. arXiv preprint arXiv:2108.06217

Cherubini M (n.d.) Ethical autonomous algorithms. Retrieved from https://medium.com/@mchrbn/ethical-autonomous-algorithms-5ad07c311bcc

Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: Conference on Fairness, Accountability and Transparency, pp 134–148

Christian B (2020) The alignment problem - machine learning and human values. W. W. Norton & Company

Cognilytica (2021) Comprehensive ethical ai framework. (Tech. Rep.), Cognilytica

Corporation M (2022) Responsible ai principles from microsoft. https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6. Accessed 02 Aug 2022

Costanza-Chock S (2018) Design justice: Towards an intersectional feminist framework for design theory and practice. In: Proceedings of the Design Research Society

Crawford K (2021) The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press

Crawford K, Calo R (2016) There is a blind spot in ai research. Nature 538(7625):311–313

Custis C (2021) Operationalizing ai ethics through documentation: About ml in 2020 and beyond. Retrieved from https://partnershiponai.org/about-ml-2021/

Davies W (2017) How statistics lost their power - and why we should fear what comes next. The Guardian

Delgado F, Yang S, Madaio M, Yang Q (2021) Stakeholder participation in ai: Beyond" add diverse stakeholders and stir". arXiv preprint arXiv:2111.01122

Delgado F, Yang S, Madaio M, Yang Q (2023) The participatory turn in ai design: theoretical foundations and the current state of practice. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp 1–23

DigitalCatapult (2020) Lessons in practical ai ethics. Tech Rep, Digital Catapult

D'ignazio C, Klein LF (2020) Data feminism. MIT press

Drage E, Kanta D, Stephen C, Kerry M (2022) Who makes AI? gender and portrayals of ai scientists in popular film 1920–2020

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580

Edwards L, McAuley D, Diver L (2016) From privacy impact assessment to social impact assessment. In: Proceedings of the IEEE Computer Society Conference

Elam M, Reich R (2022) Stanford hai artificial intelligence bill of rights. https://hai.stanford.edu/white-paper-stanford-hai-artificial-intelligence-bill-rights. Accessed on 30 Nov 2023

Elshan E, Zierau N, Engel C, Janson A, Leimeister JM (2022) Understanding the design elements affecting user acceptance of intelligent agents: past, present and future. Inf Syst Front 24(3):699–730

Fair G (2021) Data based science: fair becomes the new normal. Retrieved from https://www.go-fair.org/2021/01/21/data-based-science-fair-becomes-the-new-normal/

Falk B, Gautam J, Srinivasan P, Alanoca S, Bora A, Jain A, Lannquist Y (2020) Ey & the future society report: Bridging ai's trust gap. (Tech. Rep.), EY. Accessed 13 Apr 2022

Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A (2020) Towards transparency by design for artificial intelligence. Sci Eng Ethics 26(6):3333–3361

Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication 1(2020-1)

Floridi L (2018) Soft ethics and the governance of the digital. Philos Technol 31:1–8

For Coordination CEB (2019) A united nations system-wide strategic approach and road map for supporting capacity development on ai. https://unsceb.org/sites/default/files/2020-09/CEB_2019_1_Add-3-EN_0.pdf. Accessed 18 Aug 2023

Friedman B, Hendry D (2019) Value sensitive design: shaping technology with moral imagination. MIT Press

Friedman B, Kahn P, Borning A (2002) Value sensitive design: theory and methods. University of Washington. Tech Rep 2:12

Fuchsberger V, Moser C, Tscheligi M (2012) Values in action (via): Combining usability, user experience and user acceptance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)

Gabriel I (2020) Artificial intelligence, values, and alignment. Minds Mach 30(3):411–437

Gaggioli A, Riva G, Peters D, Calvo RA (2017) Positive technology, computing, and design: shaping a future in which technology promotes psychological well-being. Emotions and affect in human factors and human-computer interaction. Elsevier, pp 477–502

Garst J, Blok V, Jansen L, Omta O (2022) From value sensitive design to values absorption - building an instrument to analyze organizational capabilities for value-sensitive innovation. J Responsible Innov 9(2):196–223

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, Crawford K (2021) Datasheets for datasets. Commun ACM 64(12):86–92

Google (unknown). Responsible ai practices - google ai. https://ai.google/responsibilities/responsible-ai-practices/. Accessed 02 Aug 2022

Gupta A, Wright C, Ganapini MB, Sweidan M, Butalid R (2022) State of AI ethics report (volume 6, february 2022). arXiv preprint arXiv:2202.07435

Gutierrez CI, Marchant GE (2021) A global perspective of soft law programs for the governance of artificial intelligence. Available at SSRN 3855171

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Minds Mach 30(1):99–120

Harbers M, Overdiek A (2022) Towards a living lab for responsible applied ai. In: Proceedings of the DRS 2022. Retrieved from https://doi.org/10.21606/drs.2022.422

Hemment D, Aylett R, Belle V, Murray-Rust D, Luger E, Hillston J, Rovatsos M, Broz F (2019) Experiential AI. Computing Research Repository (CoRR). Retrieved from arXiv:1908.02619

Ho CW, Ali J, Caals K (2020) Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. Bull World Health Org 98(4):263

Holden J (2018) Improving nursing student empathy with experiential learning. In: Proceedings of the Nursing Education Research Conference

Ibáñez JC, Olmeda MV (2021) Operationalising ai ethics: how are companies bridging the gap between practice and principles? an exploratory study. AI Soc 37(4):1663–87

IBM (2021) AI ethics IBM. https://www.ibm.com/cloud/learn/ai-ethics. Accessed 02 Aug 2022

IEEE standard model process for addressing ethical concerns during system design (Standard) (2021) The Institute of Electrical and Electronics Engineers

Jacobs AZ (2021) Measurement as governance in and for responsible AI. arXiv preprint arXiv:2109.05658

Jakesch M, Buçinca Z, Amershi S, Olteanu A (2022) How different groups prioritize ethical values for responsible AI. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399

John-Mathews J-M, Cardon D, Balagué C (2022) From reality to world. A critical perspective on ai fairness. J Bus Ethics 178(4):945–59

Kawakami A, Sivaraman V, Cheng H-F, Stapleton L, Cheng Y, Qing D, Perer A, Wu ZS, Zhu H, Holstein K (2022) Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In: CHI Conference on Human Factors in Computing Systems, pp 1–18

Keyes O, Hutson J, Durbin M (2019) A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In: Extended abstracts of the 2019 CHI conference on human factors in computing systems, pp 1–11

Keyes O, Drage E, Mackereth K (2022) Podcast 'the good robot' - os keyes on avoiding universalism and 'silver bullets' in tech design. https://thegoodrobotpodcast.buzzsprout.com/1786427/10692660-os-keyes-on-avoiding-universalism-and-silver-bullets-in-tech-design. Accessed 18 Aug 2022

Koesten L, Gregory K, Groth P, Simperl E (2021) Talking datasets-understanding data sensemaking behaviours. Int J Hum-Comput Stud 146:102562

Krijger J (2021) Enter the metrics: critical theory and organizational operationalization of ai ethics. AI Soc 37(4):1427–37

Lane G, Angus A, Murdoch A (2018) Unbias fairness toolkit. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.2667808

Lee MK, Kusbit D, Kahng A, Kim JT, Yuan X, Chan A, See D, Noothigattu R, Lee S, Psomas A, et al (2019) Webuildai: Participatory framework for algorithmic governance. In: Proceedings of the ACM on Human-Computer Interaction 3(CSCW):1–35

Lee MSA, Floridi L, Singh J (2021) Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI Ethics 1(4):529–544

Leslie D (2019) Understanding artificial intelligence ethics and safety. https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf. Accessed 23 Aug 2023

Lohr S (2018) Facial recognition is accurate, if you're a white guy. Ethics data and analytics. Auerbach Publications, NY, pp 143–147

Manders-Huits N, Zimmer M (2009) Values and pragmatic action: the challenges of introducing ethical intelligence in technical design communities. Int Rev Inf Ethics 10:37–44

McCradden M, Mazwi M, Joshi S, Anderson JA (2020) When your only tool is a hammer: ethical limitations of algorithmic fairness solutions in healthcare machine learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 109–109

McNamara A, Smith J, Murphy-Hill E (2018) Does acm's code of ethics change ethical decision making in software development? In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2018, New York, NY, USA. Association for Computing Machinery. pp 729-733. Retrieved from https://doi.org/10.1145/3236024.3264833

Metcalf J, Moss E, Watkins E, Singh R, Elish M (2021) Algorithmic impact assessments and accountability: the co-construction of impacts. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)
Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1(11):501–507
Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2):2053951716679679
Modha S, Majumder P, Mandl T, Mandalia C (2020) Detecting and visualizing hate speech in social media: a cyber watchdog for surveillance. Expert Syst Appl 161:113725
Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L (2021) Ethics as a service: a pragmatic operationalisation of ai ethics. Minds Mach 31(2):239–256
Morley J, Kinsey L, Elhalal A, Garcia F, Ziosi M, Floridi L (2021) Operationalising ai ethics: barriers, enablers and next steps. AI Soc 38:411–423
Moss E, Watkins E, Metcalf J, Singh R, Elish M (2021) Governing with algorithmic impact assessments: six observations. In: Proceedings of the ACM Conference on Artificial Intelligence, Ethics and Society (AIES)
Mujtaba DF, Mahapatra NR (2019) Ethical considerations in ai-based recruitment. In: 2019 IEEE International Symposium on Technology and Society (ISTAS), pp 1–7
Munn L (2023) The uselessness of AI ethics. AI Ethics 3(3):869–877
Munn Z, Peters M, Stern C (2018) Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. https://doi.org/10.1186/s12874-018-0611-x
Narayanan A (2018) Translation tutorial: 21 fairness definitions and their politics. In: Proc. conf. fairness accountability transp., New York, USA, vol. 1170,p 3
Observatory OP (2018) Inclusive growth, sustainable development and well-being (OECD AI principle). https://oecd.ai/en/dashboards/ai-principles/P5. Accessed 18 Aug 2023
OECD (2019) Recommendation of the council on artificial intelligence - oecd/legal/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Accessed 11 July 2022
Office ATP (2022) Statement on prinicples for responsible algorithmic systems. https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf. Accessed 23 Aug 2023
Organization WH et al (2021) Ethics and governance of artificial intelligence for health: WHO guidance
Ozkaramanli D, Desmet P, Özcan E (2016) Beyond resolving dilemmas: three design directions for addressing intrapersonal concern conflicts. Des Issues 32:78–91. https://doi.org/10.1162/DESI_a_00401
Pasquinelli M, Joler V (2021) The nooscope manifested: AI as instrument of knowledge extractivism. AI Soc 36:1263–1280
Pesenti J (2021) Facebook's five pillars of responsible AI. https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/. Accessed 02 Aug 2022
Peters D, Vold K, Robinson D, Calvo RA (2020) Responsible ai-two frameworks for ethical design practice. IEEE Trans Technol Soc 1(1):34–47
Rahwan I (2018) Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol 20(1):5–14
Rakova B, Yang J, Cramer H, Chowdhury R (2021) Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. In: Proceedings of the ACM on Human-Computer Interaction 5(CSCW1):1–23
Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic impact assessments: a practical framework for public agency accountability
Sadek M, Calvo R, Mougenot C (2023) Co-designing conversational agents: a comprehensive review and recommendations for best practices. Des Stud 89:101230
Sadek M, Calvo R, Mougenot C (2023) Designing value-sensitive ai: a critical review and recommendations for socio-technical design processes. AI Ethics. https://doi.org/10.1007/s43681-023-00373-7
Sadek M, Calvo R, Mougenot C (2023c) Why codesigning ai is different and difficult. ACM Interactions. Retrieved from https://interactions.acm.org/blog/view/why-codesigning-ai-is-differentand-difficult
Scantamburlo T, Cortés A, Schacht M (2020) Progressing towards responsible ai. arXiv preprint arXiv:2008.07326
Schiff D, Rakova B, Ayesh A, Fanti A, Lennon M (2020) Principles to practices for responsible AI: closing the gap. arXiv preprint arXiv:2006.04707
Simonite T (2021) What really happened when google ousted timnit gebru - wired. https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/. Accessed 04 Apr 2023
Spiekermann S (2021) From value-lists to value-based engineering with IEEE 7000™. In: IEEE International Symposium on Technology and Society (ISTAS)
Spiekermann S, Winkler T (2020) Value-based engineering for ethics by design. Computing Research Repository (CoRR)
Stray J (2020) Aligning ai optimization to community well-being. Int J Commun Well-Being 3(4):443–463
Stray J, Vendrov I, Nixon J, Adler S, Hadfield-Menell D (2021) What are you optimizing for? aligning recommender systems with human values. arXiv preprint arXiv:2107.10939
Tasioulas J (2021) The role of the arts and humanities in thinking about artificial intelligence (ai) ada lovelace institute. https://www.adalovelaceinstitute.org/blog/role-arts-humanities-thinking-artificial-intelligence-ai/. Accessed 04 Aug 2022
Tasioulas J (2022) Artificial intelligence, humanistic ethics. Daedalus 151(2):232–243. https://doi.org/10.1162/daed_a_01912
Thomas R, Uminsky D (2020) The problem with metrics is a fundamental problem for ai. arXiv preprint arXiv:2002.08512
Thomas RL, Uminsky D (2022) Reliance on metrics is a fundamental challenge for ai. Patterns 3(5):100476
UNESCO (2021) Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455. Accessed 18 Aug 2023
Varanasi R, Goyal N (2023) It is currently hodgepodge": Examining ai/ml practitioners' challenges during co-production of responsible ai values. In: Proceedings of the CHI Conference on Human Factors in Computing Systems
Wachter S, Mittelstadt B (2019) A right to reasonable inferences: rethinking data protection law in the age of big data and ai. Colum Bus L Rev, p 494
Wagner S (2022) Ai in government featuring stuart wagner, chief digital transformation officer, us air force & us space force. Presentation at Cognilytica. AI and can be accessed at: https://www.cognilytica.com/session/july-2022-ai-in-government/?hash=62dfc4b9b79db
West S, Whittaker M, Crawford K (2019) Discriminating systems: Gender, race, and power. Tech Rep. AI Now Institute
Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The role and limits of principles in ai ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp 195–200
Widder DG, Nafus D, Dabbish L, Herbsleb J (2022) Limits and possibilities for "ethical ai" in open source: A study of deepfakes. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, New York, NY, USA. Association for Computing Machinery. pp 2035–2046. Retrieved from https://doi.org/10.1145/3531146.3533779

Yang Q, Steinfeld A, Rosé C, Zimmerman J (2020) Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In: Proceedings of the 2020 chi conference on human factors in computing systems, pp 1–13

Yildirim N, Pushkarna M, Goyal N, Wattenberg M, Viégas F (2023) Investigating how practitioners use human-ai guidelines: a case study on the people + ai guidebook. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, New York, NY, USA. Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3544548.3580900

Young M, Magassa L, Friedman B (2019) Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. Ethics Inf Technol 12:89–103

Zicari RV, Brodersen J, Brusseau J, Düdder B, Eichhorn T, Ivanov T, Kararigas G, Kringen P, McCullough M, Möslein F et al (2021) Z-inspection®: a process to assess trustworthy ai. IEEE Trans Technol Soc 2(2):83–97