AI & SOCIETY
https://doi.org/10.1007/s00146-024-02082-z

MAIN PAPER

Towards trustworthy medical AI ecosystems – a proposal for supporting responsible innovation practices in AI-based medical innovation

Christian Herzog[1] · **Sabrina Blank**[1] · **Bernd Carsten Stahl**[2]

Received: 28 March 2024 / Accepted: 10 September 2024

Abstract
In this article, we explore questions about the culture of trustworthy artificial intelligence (AI) through the lens of ecosystems. We draw on the European Commission's Guidelines for Trustworthy AI and its philosophical underpinnings. Based on the latter, the trustworthiness of an AI ecosystem can be conceived of as being grounded by both the so-called rational-choice and motivation-attributing accounts—i.e., trusting is rational because solution providers deliver expected services reliably, while trust also involves resigning control by attributing one's motivation, and hence, goals, onto another entity. Our research question is: What aspects contribute to a responsible AI ecosystem that can promote justifiable trustworthiness in a healthcare environment? We argue that especially within devising governance and support aspects of a medical AI ecosystem, considering the so-called motivation-attributing account of trust provides fruitful pointers. There can and should be specific ways and governance structures supporting and nurturing trustworthiness beyond mere reliability. After compiling a list of preliminary requirements for this, we describe the emergence of one particular medical AI ecosystem and assess its compliance with and future ways of improving its functioning as a responsible AI ecosystem that promotes trustworthiness.

Keywords Ethics · AI · Healthcare · Ecosystems · Responsible research and innovation · AI governance · Trustworthiness

1 Introduction

A prominent narrative in the discussion about trust and ethics in artificial intelligence (AI), particularly medical AI, can be paraphrased as follows:

> AI promises enormous benefits but simultaneously raises significant ethical concerns (Flick et al. 2020). These ethical concerns are a crucial obstacle to AI adoption. Developers, vendors, and users need to be able to trust the technology to use it. Such trust can only develop if ethical and related problems are successfully addressed—an antecedent that is true across all AI application domains but maybe most prominently visible in medical AI applications (Haque et al. 2020; Iqbal et al. 2016; Topol 2019).

In this view, trust is central, but also largely instrumental to AI innovation, which takes primacy in many political agendas. In this contribution, we want to challenge this view by highlighting trustworthiness as a normative requirement for any AI-based innovation to be desirable in the first place. Rather than perceiving a lack of trust as an obstacle to innovation, by referring to philosophical concepts of trust and trustworthiness (cf., e.g., Baier 1986; Nickel et al. 2010), we want to explore how innovation ecosystem governance structures can support the emergence of innovation that is deserving of trust—particularly in the medical sector.

However, it remains true that addressing ethical issues is a prerequisite for the formation of trust relationships or avoiding a loss of trust. Ethical concerns of medical AI include issues such as biases and fairness (Ricci Lara et al. 2022), transparency and explainability (Kempt et al. 2022) but

✉ Christian Herzog
christian.herzog@uni-luebeck.de

Sabrina Blank
sabrina.blank@uni-luebeck.de

Bernd Carsten Stahl
bernd.stahl@nottingham.ac.uk

[1] Ethical Innovation Hub, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

[2] School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UK

Published online: 16 October 2024

also broader questions of distributive justice (Lehoux et al. 2019). Proposed and often already implemented mitigation options include high-level development of policy, legislation and regulation (CAHAI 2022; OECD 2019), including the EU's AI Act (European Commission 2021b; The European Parliament and the Council of the EU 2024), national policy and legislation (e.g. UK Government 2021) or the creation of new regulatory bodies (Stahl et al. 2022). They furthermore cover existing organisational activities such as risk management (Clarke 2019) or impact assessments, including AI-specific human rights impact assessments (Mantelero and Esposito 2021) or AI impact assessments (Stahl et al. 2023). In addition, there is a plethora of codes of ethics (Fjeld et al. 2020; Jobin et al. 2019) meant to guide individuals and organisations and that in many cases have been translated into more applicable development methodologies (Borenstein et al. 2021; Kazim and Koshiyama 2021), concrete implementation advice (Petersen et al. 2022) and other types of tools (Morley et al. 2021a, b).

This very brief characterisation of the debate about the ethics of AI points to several challenges. As stated before, it is often proposed that open ethical questions prevent the development of trust, which, in turn, prevents the benefits of AI from materialising (UK Government 2023). This pragmatic utility of ethics to engender trust and promote industrial policy may be contentious. However, it points to a fundamental issue of AI ethics, namely its nature of an ethics of socio-technical systems. AI technologies are technical systems that form part of broader socio-technical systems embedded in political, economic, and social systems. Because of its nature as a system, there are rarely straightforward solutions to ethical concerns. There is not just a multiplicity of issues, stakeholders, and possible interventions, but these also typically stand in complex relationships. Consequently, there are frequent references to "AI ecosystems", particularly in interventions that highlight the ethical complexities of AI (see, e.g., Findlay and Seah 2020; Fjeld et al. 2020; UK Government 2021; UNESCO 2020). Some authors refer to such a reorientation of attention towards relations and structures as a "structural turn" (Bolte and Van Wynsberghe 2024) or a 'relational turn' (Branford 2023; Heilinger 2022).

Taking this perspective of AI as an ecosystem—or maybe *ecosystems*—changes the perspective on AI ethics from being concerned with individual technological artefacts to wider considerations concerning socio-technical systems. The present article focuses on medical AI and the question under which conditions stakeholders would trust AI systems for good reasons. Therefore, we focus on the research question: What constitutes a responsible AI ecosystem that can promote justifiable trustworthiness in a healthcare environment? We respond to the question via a combination of conceptual and empirical evidence. The conceptual analysis of responsible and trustworthy AI ecosystems and of the specific requirements that arise in a medical context allows us to highlight the characteristics that a responsible AI ecosystem in healthcare would have to display. We then present an example of such a medical AI ecosystem and analyse its current state to determine to which extent it meets the requirements of a responsible and trustworthy AI ecosystem. By framing it this way, we do not want to be mistaken as claiming that the entire ecosystem is guaranteed to allow only trustworthy or responsible conduct and thus can itself be denoted a trustworthy ecosystem. Rather, we suggest to denote a trustworthy ecosystem as a short-hand for stating that there are active steps taken within both centralised and decentralised governance structures that support conduct that lead to justifiable trust relationships. We will elucidate in the sequel, what criteria we adopt to denote justifiable trust—and while our criteria will not be formal, they will go beyond the mere vernacular meaning of trust.

We explicitly limit our discussion to applications of narrow AI in medicine—perhaps some limited multi-purpose AI tools as well. As such, we conceive of systems that can, e.g., intelligently identify tissues (Bockelmann et al. 2022), pose estimation in operating theaters (Hansen et al. 2019), or image classifiers and segmentation in radiology (see, e.g., Topol 2019, and references therein). Although an interesting topic in its own right, we do not consider AI as substitutes for—or even mimicking—medical personnel and its corresponding ethics (cf. Danaher and Nyholm 2024), let alone touch on scenarios in which any AI solution may actually be recognized as developing anything reminiscent of consciousness. Accordingly, we deem the ethical issues discussed in this paper to be concrete, even pressing, but not related to discussions about existential risks humanity may face due to advancements in AI.

The work presented in this article makes academic and practical contributions of significant relevance to several audiences. By further developing the idea of responsible AI ecosystems and describing what would render them trustworthy, the article contributes to the AI ethics literature. This ecosystems approach to the ethics of AI can help understand some of the fundamental limitations of current AI ethics. We realise that the relationship between ethics and systems in general and ecosystems in particular is complex. Ethics tends to focus on individual humans as the subjects of responsibility, whereas systems emphasise the role of structures over individual agency. This potential contradiction needs to be addressed for AI ethics to make progress. In addition to the theoretical interest of the article, it makes an important contribution to the practice of AI ethics, whose debate has long moved beyond purely theoretical interest. Implementing AI systems in many contexts, including healthcare, calls for practical ways of dealing with possible

ethical issues. The ecosystems perspective will provide pointers to how this may be achieved.

The next section provides a conceptual account of trustworthy and responsible AI ecosystems, followed by an account of trustworthiness in medical AI systems. Based on the insights derived from these positions, the article offers an empirical account of one particular medical AI ecosystem, which provides the basis for the discussion and conclusion.

2 Trustworthy and responsible AI ecosystems

The answer to the research question calls for some conceptual clarity concerning the shape and content of AI and its influences on trustworthiness. Therefore, we introduce the concept of AI ecosystems and ask what would constitute responsible AI ecosystems. We follow up with a more specific exploration of issues of trust in medical AI.

2.1 Responsible AI ecosystems

That AI can raise ethical and social concerns has triggered a wide-ranging discourse on AI ethics (Bartneck et al. 2021; Dignum 2019; Dubber et al. 2020; Siau and Wang 2020), including questions about privacy and data protection, fairness, discrimination, transparency, accountability, and reliability, but also broader societal questions related to distributive justice concerning AI benefits and risks, changing nature of work, impact on democracy and power concentration as well as international and global concerns such as the exacerbation of global heating or the changing nature of warfare. In the corresponding application settings, relevant stakeholders include those on the receiving end of AI technologies, such as patients for medical applications (e.g., Cabitza et al. 2017), workers for management applications (e.g., Lane et al. 2023) or citizens in public administration applications (e.g., Hadwick and Lan 2021). In addition to the users of AI, stakeholders also include developers, designers, and managers of AI technologies. These are usually more proficient regarding the technical intricacies of issues, e.g., with biases. For instance, the very goal of devising a bias-free AI solution may be unviable, as different notions of fairness may not be simultaneously enforceable incurring inherent trade-offs (Kleinberg et al. 2017).

There is considerable buzz around the word AI. Quite obviously, many of our insights and suggestions in this paper apply to technology and medical technology, in general. Regarding trustworthiness, the significance of AI over standard technologies, however, is not based on its potential to mimick human behavior, e.g., in chatbot interfaces—even though we agree that there is significant potential for deception. Rather, we follow many authors in characterising AI as an inherently socio-technical technology (e.g., Duenser and Douglas 2023) that relies on relational data and is increasingly capable of automating, standardising, mediating and affecting interactions without acting as a clearly visible technology, but rather with a tendency to remain opaque. In that regard, we identify a common theme of AI's influence on trust relationships: AI's ranges of application and possibilities to affect healthcare practice are so large that it typically adds to the complexity of, involved parties in and challenges to establishing, maintaining and nurturing meaningful trust configurations. We will provide a more in-depth discussion of trust and medical AI in the next section.

Mitigation strategies for ethical issues with AI include international collaboration and agreements on principles, e.g., on the level of the UN (Guterres 2020; UNESCO 2020), the OECD (OECD 2019), the G20 (Jelinek et al. 2020) or subsets of these groups. Such broader international policy agreements can feed into national policies (Ulnicane et al. 2021). These policies tend to focus on promoting the benefits of AI. However, they typically pay attention to risks and the need to address these (e.g. UK Government 2021). The next step of such interventions can include new legislation or regulation, with the most notable example being the proposed EU AI Act (European Commission 2021a), or the application of existing laws, e.g., data protection law, competition law, intellectual property or liability law. While there is much activity on the policy level, many suggestions also address the integration of considering ethical concerns at the organisational level, e.g., through industry bodies but also organisational mechanisms such as data governance (Hall and Pesenti 2017), risk management (Clarke 2019; NIST 2022) or impact assessments (Stahl et al. 2023). Finally, organisations and individuals can make use of an abundance of ethical frameworks, principles, and codes (Fjeld et al. 2020; Jobin et al. 2019), standards (IEEE Computer Society 2021) and professional/technical guidance.

The reason for reiterating this multitude of issues, stakeholders and mitigation options is to demonstrate the need for a perspective that allows the understanding of the problem complex more holistically. Individual interventions may help address particular problems and developing them in more detail is called for. However, the overall outcome of any individual intervention is difficult to predict because of the complexity of the landscape. Hence, there is increasing prominence of a systems view of AI. As AI represents a rapidly developing landscape that includes many different contributors and participants that interact in complex and often unforeseeable ways, the metaphor of AI as an ecosystem has gained prominence (Digital Catapult 2020; H-LEG on AI of the EC 2019; Nishant et al. 2020; NIST 2022). Using the ecosystems perspective allows moving beyond the technical systems that constitute the artefacts at the heart of AI and the socio-technical systems of which they form a part and

to look at how those socio-technical systems are embedded in economics, society, and politics.

It has been argued that the ecosystems perspective provides a unique vantage point for observing the ethics of AI and the basis for practical interventions in such ecosystems (Stahl 2021, 2022). For the present article, this raises the question of the impact of the ecosystems metaphor on questions of trust in medical AI. A more or less trusting relationship between someone who trusts (a trustor) and someone or something they trust (a trustee) will typically develop within a socio-technical context, thus within the relevant ecosystem in which they operate. We propose that a justified trusting relationship is more likely to develop in an environment that can be described as a responsible ecosystem. By this, we mean an ecosystem geared towards recognising and addressing ethical concerns in a way that promotes the overall good. Previously, the term 'flourishing' has been adopted to represent this overall good (Stahl et al. 2021). The term draws on the ancient Greek tradition of philosophy. However, it resonates with modern liberal and democratic societies in that it emphasises the ability of individuals and groups to achieve their potential and to live a good life according to justified preference (Bynum 2006).

A responsible AI ecosystem is thus one that proactively promotes human flourishing. How exactly this can be achieved is a difficult question to answer. The etymology of the term 'responsibility' points to the ability to respond, to give answers. A responsible AI ecosystem should thus be one that can respond meaningfully to questions from its constituent members but also from external stakeholders. Some conditions for an ecosystem to be able to do so will be that it is clearly defined, has a knowledge base and mechanisms to update this knowledge base, and has an adaptive governance structure (Stahl 2021). These are probably necessary but not sufficient characteristics of responsible AI ecosystems, and there is no guarantee that the ecosystem in question can address ethical concerns or that this impacts the formation of trust. Therefore, this article inquires about the constituents of responsible AI ecosystems that can promote justifiable trustworthiness. We cannot hope to claim that our proposals will guarantee trustworthiness, but hope to give plausible propositions that support trust relationships.

Some of these propositions may be too idealistic. For instance, many of those trust relationships that hopefully incur will lead to some kind of transitive effect, meaning that some entity A acknowledging the trustworthiness of another entity B adds to the grounds of yet another entity C (e.g., a patient) to also trust because C finds A trustworthy. Now C—as the potentially most vulnerable of the entities—must at least have the ability to determine the trustworthiness of A. In addition, it would be better if C must not only depend on its trust relationship to A because A's assessments could occasionally be false. So rather, A should try and facilitate that C forms a kind of trust relationship with B in a more direct sense. Accordingly, many of our proposals will require A to translate some of the characteristics of B (which may be quite technical or involved in terms of the application domain of, say, medical expertise, outcome statistics, organisation, etc.) for C to understand. Other proposals will amount to A enforcing characteristics on B. Both approaches and yet other dedicated ones, may eventually also lead C to form at least some ability to approximate B's trustworthiness in more direct terms. In any case, in a multitude of approaches, it may still occur that A does indeed find B to be trustworthy but still fails to communicate this to C convincingly, while C is incapable of assessing the trustworthiness of B itself. Even with the most appropriate ecosystem structures in place, epistemic processes may be outpaced by technological and organisational advances, such that C is left without devices to assess trustworthiness directly or transitively. We cannot and will not continue within such a level of formal rigor, but would like to take this brief excursion as pointing out the qualitative and propositional nature of our inquiry—as opposed to a formally verifiable one.

To move towards an answer to what qualitative propositions support the formation of trust relationships within an ecosystem, we first discuss trustworthiness in medical AI.

3 The Issue of trustworthiness in medical AI

After having reviewed relevant literature on responsible AI ecosystems, we continue by exploring the issue of trustworthiness, which is particularly pronounced in medicine. Here, those most in need of being able to trust are typically patients in potentially dire and particularly vulnerable circumstances. However, we will also consider the issue of trust from other perspectives, such as that of medical personnel, which is increasingly dependent on technology (Rampton et al. 2022), albeit carrying the brunt of the responsibility (Binkley 2021; Sand et al. 2022). We will first review the concepts of trust in and trustworthiness of technology in more general terms before commenting on their specific significance in medical AI.

3.1 The concepts of trust and trustworthiness in public and academic debate

Trust in and trustworthiness of technology are notions that have spawned a sustained philosophical debate (cf. Andras et al. 2018; Durán and Formanek 2018; Durán and Jongsma 2021; Floridi 2019a, b; Laux et al. 2023; Nickel et al. 2010; Rieder et al. 2020; Ruokonen 2013; Ryan 2020). Trust and trustworthiness are subject to various interpretations, and the use of these terms in high-profile policy, consultancy, and ethics guidelines is widely debated. For instance, in

reference to the European Commission's Ethics Guidelines on Trustworthy AI (H-LEG on AI of the EC 2019), Ryan (2020) writes that "[t]rust is one of the most important and defining activities in human relationships, so proposing that AI should be trusted, is a very serious claim." Ryan's contention appears justified, as, indeed, common political jargon often referring to increasing the acceptance of AI (e.g., The State Chancellery of Schleswig–Holstein, Germany 2021) or building trust (e.g., European Commission 2020) as essential goals in their own right—without ever casting doubt onto whether acceptance or trust would be warranted.

The common rationale adopted by some appears to be that trust automatically leads to acceptance, that acceptance is desirable, and that common drivers of trust merely relate to expectations met and risks mitigated (cf. Gillespie et al. 2023; Hengstler et al. 2016). According to a KPMG report (Gillespie et al. 2023), trust in AI also merely equates to the *perceived* trustworthiness of AI, which, in turn, supposedly indicates tractable ways towards acceptance as dependent on the potential trustor's subjective confidence in institutions, exposure to demonstrations of beneficial AI implementations and education on AI technologies. Similarly, the EU's white paper on artificial intelligence even makes explicit mention of an "ecosystem of trust", which should serve the purpose of building trust to "speed up the uptake of the technology" (European Commission 2020, p. 10).

Contrary to this kind of jargon that saturates discussions on the European level, Rieder et al. (2020) have made a compelling case for the existence of deeper philosophical underpinnings of trust in and trustworthiness of technology/AI within the European Commission's Ethics Guidelines on Trustworthy AI (H-LEG on AI of the EC 2019) by summarising and drawing conclusions from work, such as Nickel et al.'s (2010), but also by analysing the guidelines themselves. A central aspect in that latter regard rests on the guidelines' inherent claim to go beyond legal compliance, hence, explicitly dealing with the moral dimension of what *should be done* rather than "what *legally* can be done" (Rieder et al. 2020, p. 3). This relationship between the trustworthiness of technology and the moral domain has been made explicit by Nickel et al. (2010), who distinguish between the *rational-choice* and *motivation-attributing accounts of trust*. We will adopt this conceptualisation, fully aware that there might be competing alternatives that are less or even more intricate. For the purposes of this paper, we pursue the goal to plausibly indicate that Rieder's and Nickel's accounts lead to useful strategies for ecosystem governance that promotes medical AI innovations that cater to the important needs of both patients, relatives as well as medical personnel. This is why we will also comment on the plausibility of the trust accounts in medicine in the sequel.

In the rational-choice account, trust is conceptualised as a rational cost–benefit calculation regarding the effectiveness of relying on something or another person to perform as expected. Hence, the rational-choice account equates trust with reliance and therefore lacks any moral interpretation that is implicit (or explicit) to many philosophical accounts of trust. For instance, Baier (1986, p. 235) references the benevolence (or malevolence) of the human will when denoting trust as "accepted vulnerability to another's possible but not expected ill will (or lack of good will)". Acknowledging the trustee as an autonomous agent with personal motivations warrants a concept of trust beyond the rational-choice account. This motivation-attributing account of trust, then, requires the trustor to attribute a specific motivation to the trustee, such as the motivation to act per the values and goals of the trustor.

When accepting the merit of the motivation-attributing account of trust, it is widely held that technologies cannot be appropriate targets of motivation attributions (see, e.g., Hatherley 2020; Nickel et al. 2010; Rieder et al. 2020; Ryan 2020). Ryan argues that in normative terms, technology lacks agency and cannot be held responsible. Similarly, while more interested in the supposedly adverse effects of trusting (medical) AI on human relationships, Hatherley (2020) states that technology can only be relied upon. From this, Ryan (2020) even dismisses the notion of trustworthy AI as altogether misguided. Rieder et al. (2020) attempt to reconcile the motivation-attributing account of trust with a more vernacular understanding of trustworthy technology. To do so, they suggest that technology can be trustworthy in a derived sense by considering "the network of […] technologies and human agents, who are in various ways involved […] as the unit of analysis" (Rieder et al. 2020, p. 7). Hence, the responsible human agents within the socio-technical ecosystem are essentially tasked with fulfilling the requirements of being trustworthy.

In support of the human role, Montemayor et al. (2022) engage deeper with the issue of empathy—a capacity required in medical and nursing care that the authors argue AI lacks, and which subsequently presents an obstacle to certain applications. Perry (2023) underscores this by conceptualising empathic responses as indicating a motivation to make an effort, among other things. Shteynberg et al. (2024) raise caution to potential issues when the illusion of empathy is simulated. It hence seems that empathy can play a significant role in signalling trustworthiness in the sense of the motivation-attributing account. Empathic humans signal understanding and a willingness to invest the necessary resources to care and potentially also to act.

According to Rieder et al. (2020, p. 7), two general components—a moral and an epistemic—give structure to requirements for trustworthiness: the moral component refers to the requirement of being truthful and trust-responsive. The epistemic component refers to the requirement of being competent both in factual and self-reflective terms. In

other words, to be trustworthy, a human agent must be truthful in communicating intentions (such as adopting some goal or motivation on the trustor's behalf) and being responsive to trust, i.e., acknowledging that the trust conferred upon is reason enough to act accordingly. Furthermore, one must be able to carry out the respective tasks while knowing the extent and limits of one's abilities, how to recognise and potentially reduce deficits in capabilities. To be clear, truthfulness, trust-responsiveness and factual and self-reflective competence may come in differing degrees. Human subjects may fail to disclose some lesser important facts or may not be fully aware of the extent of their abilities. Neither Riedel or Nickel engage in a discussion of the appropriate (and most often only inaccurately assessable) degree by which agents should exhibit both the moral and epistemic component to form a necessary condition for trustworthiness. As our exposition is qualitative in nature, we do not and cannot on any provable threshold conditions, but would rather focus on general characteristics that we deem of value.

As a noteworthy alternative perspective, Ferrario et al. (2021) propose that the rational-choice account—and, hence, a notion of reliance—is sufficient to conceptualise trust in technology, effectively disavowing the significance of moral requirements. We do not adopt this position. For one, there is no need to equate trust and reliance when a richer notion of trust can emphasise the relevance of stakeholders' values, motivations and agency for trust relationships. Second, as discussed in the sequel, acknowledging trust as something other than mere reliance will give rise to implications for the governance of trustworthy AI ecosystems geared towards a cultural shift beyond the establishment of reliability safeguards. However, our philosophical standpoint is far from disavowing the necessity of strong reliability measures. Neither are we purporting a view that medical technology and AI should simply be trusted. We do, however, maintain that reliability alone—especially in medicine—will rarely be enough and that trust can be earned through appropriate actions. We surmise that the motivation-attributing account's significance and perhaps resilience against disappointments wanes with an increasing social distance between trustor and trustee. Interpersonal relationships should have a firm moral trust footing, while business relationships and relationships with large power gradients should be based on rational decision-making implying checks and bounds, transparency, and interest groups and representatives as intermediaries. As we will discuss, in medicine, however, there is a host of situations, in which rational decision-making is only part what constitutes shared medical decisions.

Later, we will discuss an ecosystems perspective on the requirements for and limits to trustworthiness and the necessary interplay of relevant agents to achieve this. First, however, we will comment on the particular significance of trust and trustworthiness in medical AI. Before we do so, we would like to stress that our exposition is considering a human-centric perspective on medicine. While we do believe that trust and trustworthiness may be significant concepts also in relationships between humans and animals, we do not claim that our endorsement of the rational-choice and motivation-attributing account of trust holds there as well.

3.2 Trust and trustworthiness in medicine and medical AI

Why are trust and trustworthiness especially relevant in medicine and—by extension—in medical AI? The short answer is probably that in the medical domain, asymmetries in terms of both capability and knowledge can become particularly pronounced. Hence, it seems plausible that a lack of competence on, e.g., the patients' behalf necessitates trust. Such a view may lead to an impression that trust in medicine amounts to a mere choice or even a leap of faith—faith in the abilities and medically effective collaboration of all involved in the healthcare sector. We will argue, however, that trust relationships in medicine are better understood in terms of the rational-choice *and* the motivation-attributing account. Consequently, trustworthiness in medicine need not be based on unsubstantiated beliefs but can be gauged in terms of moral and epistemic components.

Quite clearly, patients depend on the knowledgeability and considerable skills of the medical personnel, even though there is a trend toward more informed patients (see, e.g., Alpay et al. 2006; Gardiner 2008; Karnam 2017). Medical AI, however, is directed towards—and is often proclaimed to eventually develop (cf., e.g., Rajpurkar et al. 2022; Topol 2019)—skills superior to humans. Hence, said asymmetries may not only concern the patient-physician relationship but may also become increasingly relevant concerning trust towards medical technology providers on behalf of the medical personnel. When considering direct interactions between patients and AI-based tools, such as diagnostic apps (e.g., Ronicke et al. 2019), medical personnel may even be sidelined as trusted intermediaries, consequently requiring a direct trust relationship between medical AI providers and patients.

During unfamiliar situations of considerable incertitude and vulnerability, patients may have little experience or facts on which a rational choice could be based. Potentially dire situations may make it difficult to form a well-founded expectation of the performance of any medical system. The remaining option is trust in the governance structures and institutions that contribute to making the medical system work. Indeed, consultancies and think tanks have recently stressed the importance of trust in the medical ecosystem (cf., e.g., Edelman GmbH 2023; Read et al. 2021; Sarasohn-Kahn 2022). A great deal of trust in medicine can hence be attributed to, e.g., the reliability of regulatory bodies,

the effective self-regulation of medical associations and fair reimbursement systems to contribute to just, safe, efficient and effective healthcare. Trust in the medical ecosystem thus depends on rational expectations but can, of course, also be diminished by evidence of its (partial) failure.

Still, many suggestions for building trust in the healthcare sector concern personal relationships (see, e.g., AAMC Principles of Trustworthiness 2021). While the significance of individual experiences and encounters with healthcare and medical personnel for the issue of trust can be backed up by evidence (e.g., Read et al. 2021), it is instructive to adopt an ecosystem perspective approach to understanding trust in the medical domain, because it widens considerations to support trustworthiness beyond patient-physician relationships to also account for other actors within the socio-technical assemblage of the ecosystem, such as medical technology providers, users and regulatory bodies (Anoop and Asharaf 2022; Bertelsmann Foundation 2023; Platt and Nong 2023; Ruotsalainen and Blobel 2020, 2022).

More specifically, the ecosystems perspective allows identifying more potential agents responsible for contributing to the overall ecosystem's trustworthiness and asking for the significance of the motivation-attributing account. For instance, pediatric cases, where children may not possess the rational capacity and experience to develop trust towards medical practitioners, reveal the importance of conceiving a triadic trust relationship between the child, parent or guardian, and physician (Sisk and Baker 2019). These trust relationships have not been explicitly considered under the proposed dichotomy of rational-choice and motivation-attributing accounts of trust, but there is good grounds to assume that the model has some merits in cases like these. For instance, Sisk and Baker (2019) emphasize the value of trust-building during non-crisis times. In support of what we call the rational-choice account, they write that "[f]amilies test their clinicians over time against their expectations of what clinicians should do. As the clinical relationship develops, the family's trust becomes increasingly contingent on the clinician's demonstrated actions, which we call relation-based trust". But Sisk and Baker also refer to the moral demands of the moral-attributing account by stating that "[s]tudies suggest that relation-based trust is supported by demonstrations of caring, fidelity, honesty, and competence". Clearly, the aspects of caring, fidelity, and honesty all refer to a relationship in which patients (or guardians, respectively) can rely on physicians adopting morally-salient motivations of those they work for. It is interesting to point out that there are, for sure, short-cuts to 'building trust' based on learned linguistic and body language skills. These could be considered superficial communicative techniques to garner a patient's or parent's trust (e.g., Depraetere et al. 2023). The benefit of adopting our trust model in analysis of such methods would be that these techniques should be grounded in genuine fulfilment of both the epistemic and moral demands of both trust accounts. Only then, could we say that perhaps immediate and life-threatening crises warrant to employ these techniques to build trust more quickly.

Similar considerations apply in obstetrics and gynecology, particularly in high-stakes decisions. This may incur moving from a triadic relationship to a more complex scenario where physicians must consider the potentially conflicting welfare of both child and mother. In times of crises, however, this scenario is even more clearly pointing to the fact that physicians should have learned enough during the formation of a trust relationship with the mother to act in her interests. These could well be that the mother would give priority to saving the child over her own life when it comes to that. It is, indeed, a situation in which reliability may play an entirely inferior role (except maybe in terms of how well any necessary procedure may be performed) to the motivation-attributing account.

Geriatric and mental health situations pose yet further challenges to the trust model. While the motivation-attributing account may continue to be significant, mental illness may be commonly assumed to reduce the capacity of rational decision-making. However, this may, in fact, not hold (cf. Cardella 2020). Hence, it may be hard to say, whether the rational-choice account of trust is not applicable during mental health crises. Even then, qualitative studies show that mental health patients look for more than expertise and reliability in their care professionals (Laugharne et al. 2012). Outside the extreme case described above, this gives an indication that the motivation-attribution account holds merit also in mental health.

Generally, then, if the trust relationship between patient and physician is obstructed, we need to look towards the socio-technical system that is medicine and include other actors, predominantly those that act as proxies, guardians, or intermediaries to the patient. Perhaps in these cases, the trust relationship is even more emphasised. These intermediaries, e.g., parents, guardians, or any other authorized representative of the patient, must clearly be aware of the moral motivations of the patient and act, communicate and discuss with the medical professionals accordingly. In addition, these intermediaries might even have strong and conflicting motivations of their own. For instance, relatives would like their loved ones to live on, while it may be the patient's wish to die peacefully and without struggle–whatever that may mean in concrete situations.

When technology, AI in particular, comes into play, trust should stem from a plausible assumption that AI developers, medical practitioners, etc., generally espouse a patient's personal goals (directly or via the intermediary) to regain health or find acceptable ways of dealing with the disease. In addition, medical professionals require their trust in the medical ecosystem to be warranted. Complex technology,

economic pressures and scarcity of resources—especially time—all contribute to medical personnel frequently taking responsibility for decisions resulting from technologically mediated and often opaque processes (Herzog 2019; van den Eede 2011). Hence, medical personnel need not only rely on, but also trust in medical equipment and AI—or, more precisely, their corresponding providers—but also other institutions, governance structures and entities to support their conception and ideals of humane and caring medicine.

However, we urge caution when attempting to translate the requirements and dynamics of trust and fiduciary relationships from a "traditional" to a heavily technology-supported and -dependent medicine. It remains questionable, e.g., whether one should, in fact, aim for conditions under which patients exert the same levels of trust when the efforts and performance of medical personnel are replaced by—at least in part—technological support systems irrespective of whether these bring along qualitative and quantitative improvements. Hatherley (2020) even seems to object to less direct human-to-human relationships in the medical context per se, which is not without merit, even from a consequentialist standpoint, as less intimate and patient-centric medicine is proven to be less effective, cf. (Bjerring and Busch 2021; Herzog 2022a). We would like to add that when the target of trust shifts from performer to provider of a—presumedly generally reliably performing—technology-based support system (as the technological system itself cannot be an appropriate target of trust), negotiations about a common idea on what medical care should be become entangled in arguably increasingly extrinsic and robust incentive structures, such as interests in profit. Unlike in the more direct interpersonal and interprofessional exchange at the point of care, such negotiations occur in a decentralised and distributed way, remote from considerations required in light of any particular, vulnerable patient's situation.

A trustworthy medical AI ecosystem needs to answer the question of how medical personnel and patients can trust technology providers, institutions and governance structures to support their respective conception of a humane and caring medicine—or at least a broader consensus on this.

4 Responsible and trustworthy innovation ecosystems in AI for health

In the above, we have briefly summarised philosophical viewpoints on trust and epistemic and moral requirements for an agent to be worthy of the merit of trust (trustworthiness). We have also commented on the particular significance of trust and trustworthiness in the medical domain. We now propose concrete implications for the governance and institutionalisation of responsible medical AI ecosystems based on the conceptual criteria for socio-technical AI systems to be trustworthy (Ruokonen 2013)—i.e., epistemic competence and trust-responsiveness. We have already commented that trust and trustworthiness can only be attributed to technology in a derived sense, meaning that only responsible actors within a socio-technical system can be appropriate targets of trust. To discuss what it takes for an ecosystem to be trustworthy, we, therefore, need to discuss the question of which actors and institutions within the socio-technical assemblage are worthy of the merit of trust, i.e., who has (or should have) the epistemic competence to deliver expected results reliably and who responds with due diligence and a moral imperative to account for the trustor's values and interests to the fact that one is being trusted? Similarly to Stahl's (2023) reasoning that a responsible ecosystem must provide "an answer to the question of who is answerable for the uses or consequences of the action of the system", a trustworthy ecosystem must provide an answer to the question of who is worthy of the merit of trust within the system, which—in turn—does not conceive of the ecosystem as comparable to a human being.

Consequently, while our discussion will focus on the ecosystem level, we will, wherever possible, comment on what this might entail for individual entities as part of the ecosystem. Our propositions are not meant to guarantee the ecosystem-wide establishment of trustworthiness. Since we have previously rejected the idea that trust can be generated, we will instead propose mechanisms, governance structures and action patterns that would facilitate that members from within, or other stakeholders from outside the ecosystem, form justified levels of trust towards entities of the ecosystem and—to some degree—towards the ecosystem as a whole, based either on a rational choice or on motivation attributions.

We structure our suggestions by discussing general governance structures and then explicating specific means of addressing ethical challenges and potentials through epistemic and methodological resources to be shared within an ecosystem.

4.1 Ecosystem governance structures and reporting

As summarised by Minkkinen et al. (2021), EU documents make wide-arching statements on, e.g., the sustainable development goals and adopting a human rights approach as value propositions congruent with global market dynamics and competitiveness. We suggest that responsible and trustworthy innovation ecosystems in AI for health connect to these global goals but certainly need to provide a more detailed account of what this should entail for actual socio-technical innovations. We further propose that the idea of cooperation, synergies and—in the style of the ecological origin of the ecosystem metaphor (cf. Moore 1993)—symbioses inherent to ecosystems is best supported by platforms. According to

(Tsujimoto et al. 2018, p. 53), "platform ecosystems […] are composed of industry-wide networks based on complex correlations between firms". Gawer and Cusumano (2014) define external platforms as "products, services, or technologies that act as a foundation upon which external innovators, organised as an innovative business ecosystem, can develop their own complementary products, technologies, or services", which matches our interests. In the following, we will set out potential roles such as platform aspects of a responsible and trustworthy innovation ecosystem in AI for health can be adopted.

4.1.1 Linking local and global ecosystem levels

The maintenance of active connections, negotiations and even socio-ethical deliberations between local (e.g., state or regional) and global (e.g., European) ecosystem levels requires governance structures and platform-based management in addition to an assemblage of essentially separate and predominantly business-oriented entities that are in part symbiotic, part competing relationships with no clear governance level (Tsujimoto et al. 2018). Such a platform-based governance structure can support the formulation of a more detailed, more domain-specific and local stakeholder-related value proposition, which agents can find their personal goals in alignment with and, hence, may eventually deem worthy of their trust. For instance, concerning frictions between data protection and continuous data acquisition and use, a platform-based governance should propose and implement oversight mechanisms, as well as acquisition, anonymisation and security standards that alleviate some ethical concerns while allowing the development of data-based algorithms with proven usefulness (cf., e.g., Vayena & Blasimme 2018). On the local level, such a platform aspect would gather experience with the complexities of data work (e.g., Berg and Goorman 1999) specific to the local data generation and storage modalities and aim to operationalise at least part of the tacit knowledge (cf. Markus 2001) and regulatory expertise necessary to put data to use. Such a governance structure would also act as a relatable voice in the more global debate while simultaneously delineating the ecosystem's boundaries to define whom it speaks for. Especially since, e.g., in the long term, a bias-reduced medical AI cannot be based on data acquisition approaches confined to particular localities, data protection must first be guaranteed locally. However, successful approaches will need active promotion to achieve more global adoption.

4.1.2 Managing and facilitating stakeholder inclusion and interaction

Governance structures and local platform aspects would also maintain active exchange between societal, industrial and academic players. Different platform aspects of a responsible and trustworthy innovation ecosystem should be composed of actors from all of the above stakeholder groups. These ecosystem platform aspects can be transparent contact points with relatable and diverse individuals. They could also function as organising entities for knowledge bases as well as for the proactive identification, formation and maintenance of synergies. London (2022) has also discussed knowledge bases in terms of data use and access as well as joint development of standards. Here, we are more concerned with ethically managing the stakeholder interaction process. Aiming for stakeholder diversity and a sense of power balance between them would support the legitimate and proportionate allocation of resources, (largely) impartial reporting, narration and advertisements of the ecosystem performance and representation of interests. Local platform aspects can become appropriate targets of the potential trust that could form towards an ecosystem in a derived sense, i.e., in terms of their relatable key actors. Platform managers and supervisory boards assume responsibility for the processes geared towards assessing an accurate image of both aggregate as well as more differentiated expectations towards the ecosystem's innovations.

4.1.3 Balancing cooperation With competition

A key task of the governance level and platform aspects would consist of providing a balance between ecosystem evolution as well as governed internal incentive structures and ecosystem standards. For instance, a study by Martinho et al. (2021) has shown that physicians both trust, but also deeply distrust technology companies and demand regulation. These sentiments have to be taken seriously, indicating a balancing of the rational-choice and motivation-attributing accounts with a tendency towards demands for providing proof of the reliability of medical technology. In that regard, an early adoption of standards on the ecosystem level by providing platform-based guidance and incentives could constitute a proactive stance towards regulation while aligning with the value sets of important stakeholders, such as physicians.

Platform-level aspects of a responsible AI ecosystem have to host, promote and contest shared epistemic resources on issues such as regulation, ethics and collaboration. A cooperative advantage within an ecosystem can consist in the ability to consult knowledge bases and experts. Knowledge bases and tools that, e.g., provide information on typical socio-ethical challenges as well as potential socio-technical solutions (cf., e.g., Petersen et al. 2022), facilitate a self-performed ethical reflection (Ayling and Chapman 2022; Manzeschke 2015; Morley, Floridi et al. 2021a, b; Reijers et al. 2016), which may not only strengthen the ecosystem's overall reputation but

may spark fruitful competition. Explicit or implicit value statements of medical technology providers are but one form of differentiating products in competition. Providing developers with the means to actually reflect on how they can provide the medical technology that suits the users' conception of good medical practice may turn out to be a decisive factor for adoption (cf. Birch et al. 2022; Bjerring and Busch 2021; McDougall 2019; Petkovic et al. 2020; Verbeek 2006). In effect, allowing users to transparently choose between solutions that best suit their values is an attempt to realise the motivation-attributing account for trustworthy socio-technical systems.

To avoid dogma or stagnation, especially in the moral domain, guidance and expert knowledge should not compromise any individual actors' strive to surpass others' 'performance', as ethics is not a check-list, but rather an open-ended endeavour (Rességuier and Rodrigues 2020).

4.1.4 Promoting a transparent reporting culture

Incentives could be set, e.g., for promoting the uptake of a wide-arching reporting and project management culture geared towards trustworthiness. In relation to epistemic requirements, the ecosystems' entities could demonstrate competency and acknowledge current competence limits by reporting on targets and achievements in terms of improved health outcomes, the ethical values pursued and the resources necessary, *as well as* even laying open business models and development statuses. Reporting measurable indicators could support individual actors in exhibiting justified reliance, while value statements and qualitative—but verifiable—reporting on the ecosystems' actors' commitment to these, would support justified motivation attributions. Reports should address all stakeholders from society, industry and academia in ways adjusted to the respective levels of expertise.

Systems engineering approaches to framing and contextualising ethical requirements on, e.g., the business model, stakeholder and implementation level (Gillespie 2019; Walden et al. 2015) could help auditing and quality management, at the very least to maintain consistency, but also in terms of critical review. Developing companies could begin by providing explicit mission and value statements, such as the so-called responsible research and innovation vision (RRI vision) mentioned in the responsibility-by-design standard (CEN CWA 17796:2021 2021) and derived from the RRI-PRISMA project on bringing RRI practice into the industry (Porcari et al. 2019). Again, not only can stakeholders check whether their personal motivations and goals align they could also contest the vision statements themselves, or adjust expectations by observing how well deeds and proclaimed motives of ecosystems' actors match.

4.2 Recognising and addressing ethical concerns and potentials

Beyond reporting or governing for ethical alignment within an ecosystem, we suggest that a platform approach to ecosystems provides the conceptual framework to establish activities that amount to a culture of trustworthy innovation. Ethical governance has generally been recognised as essential to trustworthy AI systems (e.g. Winfield and Jirotka 2018). Winfield & Jirotka, however, contend that public trust is built on standards, safety verification and validation and regulation alone. In contrast, Morley et al. (2021a, b) have recently proposed the framework "ethics as a service" or "platform as a service", which—perhaps contrary to intuition—is not about outsourcing ethical reflection and assessment but rather about distributing ethical responsibility within an ecosystem. We contend that this is a promising avenue to promote trustworthy AI innovation.

4.2.1 Distributing ethical responsibility

A responsible and trustworthy innovation ecosystem in AI for health should demonstrate the epistemic competence to assess and address ethical issues as well as identify and *prioritise* potentials. A platform structure can help to tune the balance between devolved and centralised responsibility.

An issue with auditing is that the developing company typically produces all data, while external auditors may lack the resources to properly assess the data as well as the mandate to also account for possible adverse future unintended effects (Morley et al. 2021a, b). As a remedy, Morley et al. suggest to distribute responsibility over multiple actors (individuals and companies) in the ecosystem to alleviate the dangers of a lack of accountability and ethics washing. More specifically, they suggest an independent multi-disciplinary ethics board, a collaboratively developed ethical code as well as responsible AI practises internal to any company.

The ethics board would develop the ecosystem's ethical code by managing a truly participatory and inclusive discourse as well as triangulating discursive, empirical and normative accounts of the ecosystem's ethics. The code will need to be updated on a regular basis, informed by practical insights and a continuous exchange of opinions. The code will also need to be accompanied by a process that facilitates its adoption at the company level. Morley et al. (2021a, b) identify core tenets of such a process to consist of contextualisation and translation: companies need to identify the meaning of the principles of the ethical code in the immediate context of their innovation themselves, while also given guidance on effective tools (algorithms or socio-technical practices) that allow to translate principles into practice. That guidance, however, requires further translation into practice that only the developing company can provide

and document—for which it will ultimately also be accountable for. We suggest, however, that the maintenance of a shared knowledge base on factual and process expertise on operationalisation should be the responsibility of a platform aspect within the ecosystem, i.e., separate from the ethics board, that works towards an active and shared—and visible—practice of ethical reflection and conduct.

Morley et al. (2021a, b) also suggest that the ethics board should also conduct audits on ecosystem entities (companies), stressing the need for an independent actor that guarantees transparent and truthful reporting on process adherence and contextually and ethically justifiable conduct and outcomes. In our view, we would rather consider the platform aspects that work towards responsible and trustworthy AI within the ecosystems as synergistic actors, working in close collaboration with the companies on shared epistemic resources. We therefore propose that audits—even on internal responsible innovation practices—be carried out by an entity fully independent from the ecosystem and platform aspects.

4.2.2 Platform aspect supporting responsible innovation conduct

The principles-to-practise gap in AI ethics is widely accepted to be real (e.g., Floridi 2019b; Hallensleben et al. 2020; Ibáñez and Olmeda 2022; Morley et al. 2021a, b; Schiff et al. 2020). Principles are often deemed to be too vague, lacking specific guidance. As described above, one way to address this is to demand increasing contextualisation as the discussion moves from ecosystem-wide ethical code to concrete implementation within a product development process. However, even then, Morley et al. identify at least three remaining problems: (i) during the use of translational tools, the implicit or explicit understanding of an ethical principle is not validated against the understanding(s) dominant in society, (ii) translational tools may only diagnose ethical issues but do not offer support for remedies or assigning responsibilities, and (iii) translational tools are geared towards compliance tests, as opposed to establishing a regular culture of responsible development conduct.

We contend that this assessment of challenges in translating responsible AI into practice is intimately linked to the epistemic and moral requirements of trustworthiness: (i) validating the understanding of an ethical principle in general and within a specific context (such as medicine or even some medical specialisation such as oncology) is paramount for an ecosystem entity to assess motivation attributions and exhibit trust-responsiveness, (ii) if AI ethics tools lack support for an appropriate problem-solving competence, they do not help in meeting the epistemic requirement and (iii) while compliance tests are an essential ingredient to communicating trustworthiness, only a true cultural shift would provide justifiable grounds for others to make themselves vulnerable by exhibiting trust.

An ecosystem's responsible innovation platform aspect could try to mitigate all three of these issues:

- **Research:** A platform aspect could aggregate and conduct research on domain- and actor-specific understandings of ethical principles, priorities and (changing) attitudes. Such research can be conducted both theoretically and empirically, as well as locally as well as more globally. Accordingly, similar platform aspects and actors on the local and global level can benefit from each other's research, such that, e.g., insights on other cultures can facilitate a more global market reach. However, surveys involving, e.g., local physicians, practices and patients would also more directly speak to developing within the local ecosystem and vice versa.
- **Operationalisation:** A platform aspect could also provide personnel to aggregate and conduct research on practical solutions for ethical issues, as well as to support during the actual and specific operationalisation attempts. This could—at times—amount to embedding an ethicist within a development team (e.g., McLennan et al. 2020) or further include an external ethics consultant (Blank et al. 2024) to aid in identifying and mitigating a particular set of ethical issues, conduct highly contextualised qualitative research or surveys on stakeholder preferences as well as triangulations with normative ethical requirements. In other instances, a mere networking function of the platform aspect could suffice, connecting developers between companies willing to share expertise on how to translate ethics into practice.
- **Strategic Planning:** A platform aspect's further function could consist of supporting the strategic development of action plans for addressing ethical challenges. This implies that the strategic importance must also be advertised on the platform level, ideally supported by evidence and "user stories" on the return of investments, success stories, or stories on potential disasters averted. For instance, the responsibility-by-design standard (CEN CWA 17796:2021, 2021) provides an appealing process by which companies and development teams can formulate their ethical vision, identify potential drivers and challenges, risks and barriers, as well as draw up a roadmap with concrete action plans and responsibilities. In linking responsible innovation conduct with a contextualised principle-based vision, responsible and trustworthy AI development is proceduralised in a quality assurance-like framework.

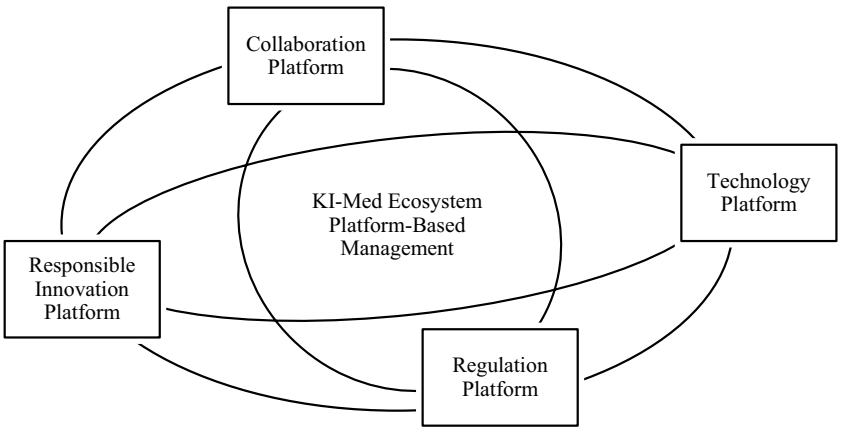

Fig. 1 Illustration of the platform-based governance components of the KI-Med ecosystem

4.2.3 Culture of self-reflection

A brief comment on institutionalizing or—at least—contributing to the ecosystem's self-reflective abilities is in order. Perhaps as part of a responsible innovation platform aspect, but possibly also as a general board of actors committed to critical revision, a trustworthy innovation ecosystem should engage in continuously questioning its conduct. For instance, in ethically-salient areas, where research is still advancing, actions should be promoted that add to possibilities to revise, monitor and provide long-term evaluations of implementations and (semi-)standardized conduct. For instance, fairness metrics may be deployed—even under consultation with diverse stakeholders—that would appear as insufficient after continued scrutiny. Data-driven methods to identify biases may turn out to be prone to statistical artifacts of fairness criteria. Current questions of whether or not to include demographics in data to avoid unwanted biases may find improved answers over time (Petersen et al. 2023). Self-reflection and -revision of actors, platform processes and governance policies can add to critical evaluations and corresponding counteractions to maintain integrity.

We have now set out theoretical ideas on what factors could support a responsible and trustworthy innovation ecosystem in AI for health. We will commence by sketching the case of the KI-Med ecosystem.

5 Future work: the case of the Northern German KI-Med ecosystem

Finally, we illustrate our conceptual argument by describing the initiation of one particular AI ecosystem. We first describe its current structure and will later move to critically discuss its current implementation as well as future avenues for more closely attaining all possible merits of a trustworthy AI ecosystem. [1]

The KI-Med ecosystem has emerged from a large third-party-funded consortium project for translating AI research into medical applications, originally titled KI-SIGS, which is a German acronym that translates into "AI Spaces for Intelligent Health Systems". As Fig. 1 illustrates, besides ecosystem management, the project implemented four core platform aspects: the collaboration platform providing internal and external means of communication and exchange; the technology platform providing reusable AI solutions; the regulation platform providing support for the eventual regulatory approval of products and the responsible innovation platform. We will focus on the latter and comment on platform interaction when appropriate.

Figure 2 depicts an UML-inspired diagram of the ecosystem structure including the platform-based management, the platform aspects, external stakeholders and translational research projects—all complete with a coarse description of properties and functions.

5.1 The responsible innovation platform

Part of the ongoing practice of the so-called responsible innovation platform (henceforth abbreviated as RI-P) aspect at the KI-Med ecosystem, in general, can be represented as strategic, operative and subsidiary support for innovators active in translatory efforts from AI-based research into medical practice as illustrated in Fig. 3.

On the operative level, the RI-P supports research teams explicitly addressing ethical, legal and societal aspects (ELSA) within the development context, such as AI

[1] KI-MED is the German name for a Northern German ecosystem, composed of the German acronym for artificial intelligence (AI), Künstliche Intelligenz (KI), and its abbreviated domain of application, medicine (Med).

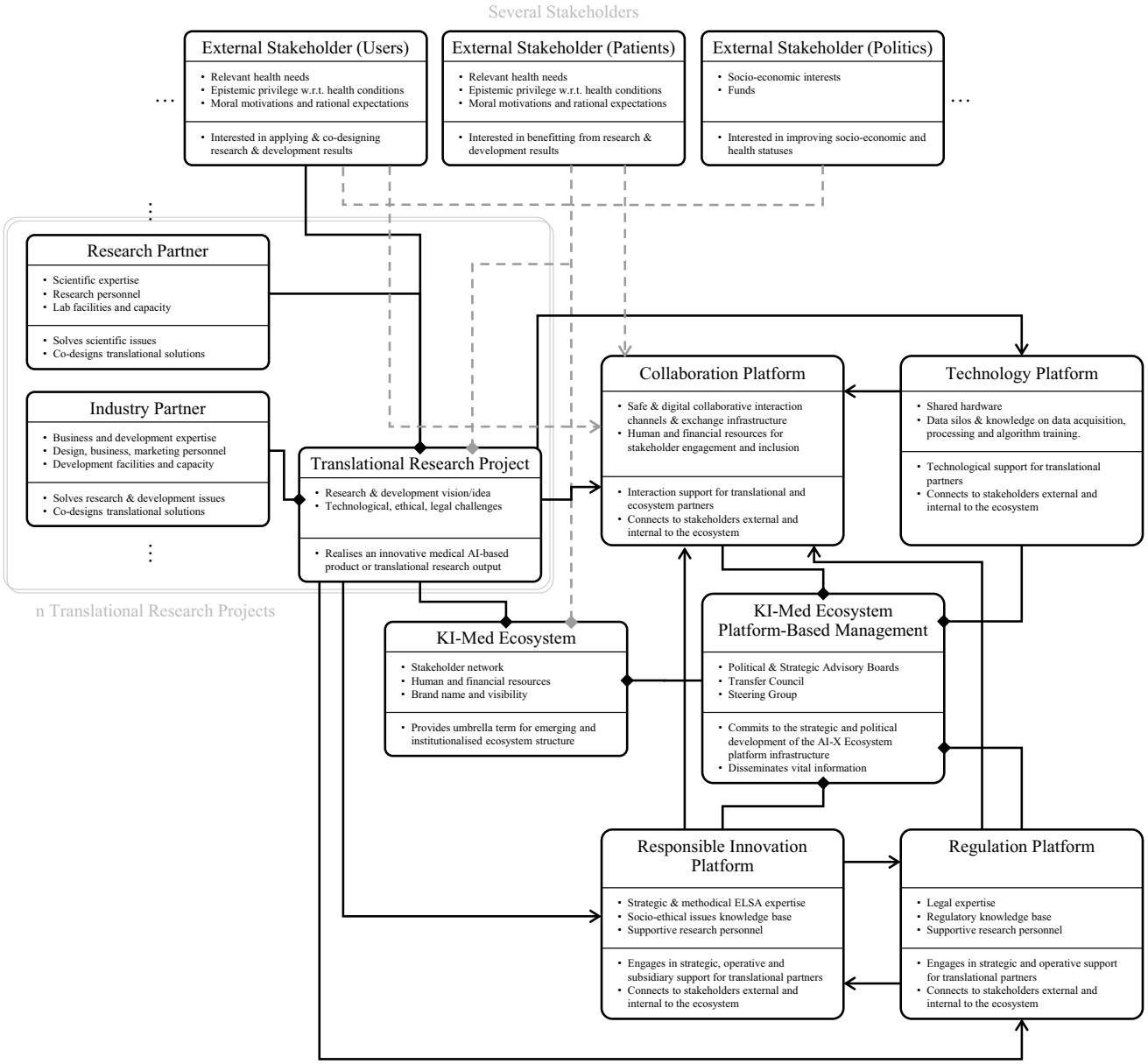

Fig. 2 UML-inspired Diagram of the Ecosystem Structure. Fields are separated into three parts (title, properties, functions). Interconnections are arrows (denotes 'uses support of and contributes to'), or diamonds (denotes 'is part of'). The initial third-party funded ecosystem consisted of nine translational research projects (not all are shown) typically composed of at least one research and an industry partner. Further external stakeholders (physicians, health institutions like hospitals, etc.) were typically involved. Platform aspects were offerings, but—apart from the collaboration platform—it was not mandatory to make use of them. Grey dashed lines indicate connections that proposals in this contribution are addressing to strengthen

explainability or algorithmic biases. Support during problem formulation and evaluation, task definitions and even algorithm selection may be offered as an ethicist is embedded into the development team (Blank et al. 2024). The RI-P supervised work on surveying potential technological solutions to both ethical and regulatory challenges and compiled as a significant collaborative endeavour within the AI ecosystem in collaboration with the regulation platform (Petersen et al. 2022).

On a strategic level, the RI-P ventures to support research teams in identifying and reflecting upon the major ethical challenges. According to the responsibility-by-design approach (CEN CWA 17796:2021, 2021), RI-P members worked with one of the nine translational research project teams to plan and map actions onto ethical issues for mitigation within a so-called RRI roadmap (Blank et al. 2024). The process offered both a sensitising and participatory way of encouraging ethical reflection within the development team

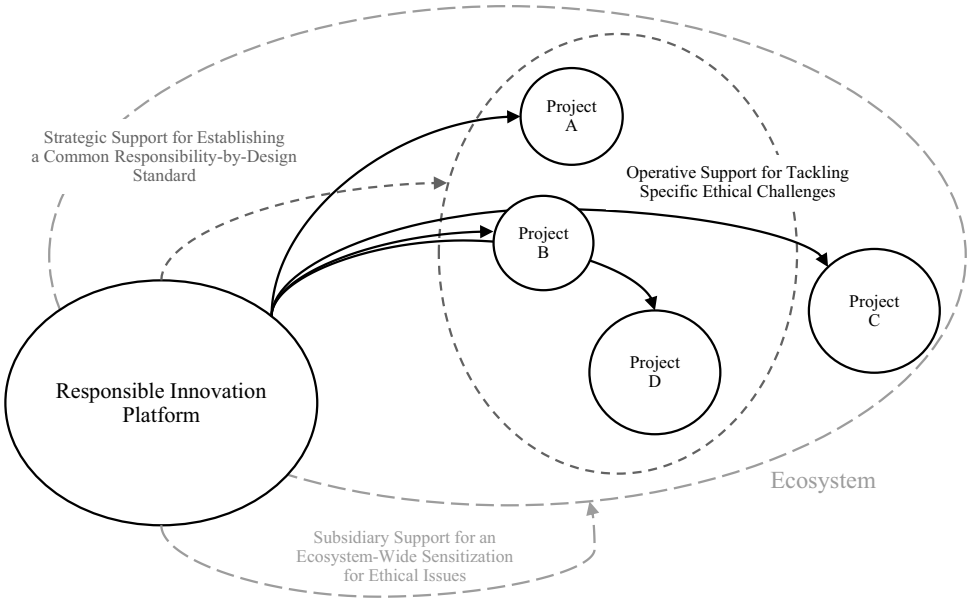

Fig. 3 Strategic, operative and subsidiary support functions of the responsible innovation platform

and increasing the team's ability to strategically allocate and plan for resources to address potential issues. Simultaneously, an ethical vision is developed, potentially refined and reflected upon during the inclusive process. In addition, the positive ethical potential of the innovation is highlighted and more clearly communicated. While exemplified as a process on a single project, the responsibility-by-design standard is supposed to be adopted on a wider scale throughout the ecosystem.

On a subsidiary level, the RI-P contributes to, collects, or initiates the production of tools for the independent analysis of ethical challenges by the research and development teams. Tools, such as the "ethics canvas" (Reijers et al. 2016), are proposed for integration on a procedural level. In addition, the RI-P has also worked on a primer meant to sensitise researchers to the significance, implications, and potential remedies of a range of ethical issues and principles, which can be used as a reference by developers. The primer is tailored to the needs of the local AI-based medical innovation ecosystem, which have been assessed in workshops beforehand.

The KI-Med ecosystem—as the follow-up to the BMWK[2]-funded KI-SIGS consortium project—is still in its infancy. Consequently, its corresponding commitment towards implementing responsible AI is still growing. While the RI-P is supported by theoretical research, such as on the epistemological and ethical utility of explicable medical AI (Herzog 2022a, b), it still lacks the resources to produce empirical insights, particularly into the ethical perception of the groups of patients and physicians. In the following discussion, we will delve deeper into how the current implementation of the KI-Med ecosystem compares with our theoretical demands on responsible and trustworthy innovation ecosystems in AI for health.

6 Discussion

We conclude by discussing future avenues to develop the KI-Med ecosystem into a responsible and trustworthy innovation ecosystem in AI for health according to the above-mentioned tenets. We will also comment on the practicality and challenges for this to happen. We do not claim that the above or the following activities and governance perspectives guarantee that the ecosystem architecture will indeed lead to increased trust. After all, even if the ecosystem proves itself to be facilitating trustworthy innovation, it may well not lead to actual perceived trust, even though it may be warranted. On the other hand, untrustworthy socio-technological solutions should not be ennobled by ecosystem activities geared towards increasing trust. Rather, ecosystem governance should encourage, maintain, perhaps control, and definitely make.

6.1 Linking local and global ecosystem levels

The consortium project has established a clear platform structure with properly defined responsibilities. However, whether each platform aspect can continue to assume the responsibilities remains a question of appropriate funding. As the KI-Med ecosystem is emerging from the consortium through decentralised funding and various funding agencies,

[2] Bundesministerium für Wirtschaft und Klimaschutz — Federal Ministry for Economic Affairs and Climate Action.

a coherent maintenance of the ecosystem governance structure is not guaranteed. Reminiscent of an assemblage (Buchanan 2021; Nail 2017), actors sustaining the former platform aspects are heterogeneous, consisting of particular individuals, technological artefacts, core ideas as well as companies or research groups close to the subject matter. Such a fragile, even volatile, transition constitutes risks for the establishment of trustworthiness but also opportunities for new actors subscribing to responsible innovation.

6.2 Managing and facilitating stakeholder inclusion and interaction

Within the transitional stage between third-party-funded projects and emerging ecosystem, the platform aspects largely depend on academia. However, the inclusion of further societal actors is desirable to support trustworthiness. For instance, an ecosystem-wide ethics board should include stakeholders from industry, academia, politics, medical personnel, and society at large. On the other hand, while close academic ties to the ecosystem's platforms may be advantageous, their conduct need not follow the incentives and logic of the academic domain. An ecosystem can benefit from technological, collaboration and responsible innovation platform aspects whose main aim is to support the societal and economic players of the ecosystem. Clearly, societal actors may even be opposed to particular technological interventions or even general technological streams, such as AI. It is important to engage with stakeholder groups rejecting essential aspects that the ecosystem nurtures, understand the origins and reasons for rejection and take these seriously.

6.3 Balancing cooperation with competition

Within the current KI-Med ecosystem structure, internal incentive structures are gravely underdeveloped. Setting out the right incentives for engaging with the cooperative modes offered by, e.g., the RI-P is paramount to adopting trustworthy innovation processes. Part of the burden also lies with the platform aspect itself, which should offer a comprehensive package of guidance and ethical reflection processes, clearly advertised from success stories as part of the shared knowledge base. When connecting with global movements, such as responsible innovation indices (cf. Nazarko 2020) or the responsibility-by-design standard, competition can be spurred within the ecosystem while providing grounds for a competitive advantage as a trustworthy ecosystem.

6.4 Promoting a transparent reporting culture

Regular conferences and project reports, particularly on responsible innovation conduct, have been held within the KI-SIGS consortium project duration, and it appears that this kind of practice will be sustained by a group of individuals, academic research groups and industrial sponsors within the KI-Med ecosystem. Though the RI-P has always been prominently featured in the yearly conferences, reporting standards that address a wide range of stakeholders and levels of expertise have not been established beyond the requirements of the funding program. Such a reporting culture could be featured within an ecosystem-wide code of ethics to which individual ecosystem players can adhere.

6.5 Distributing ethical responsibility

During its project duration, the consortium has not produced a clear value proposition developed from within the ecosystem in a participatory way that all relevant actors have subscribed to. Even though they may not be a far cry from a proposition that could result from such a process, the current values advertised by the project leaders have been derived in a top-down fashion. Accordingly, the emerging KI-Med ecosystem needs an inclusive process of determining and regularly updating a common ethical vision and an ethics board that is representative and diverse to support its trustworthiness.

6.6 Platform aspect supporting responsible innovation conduct

While the RI-P has succeeded in test-driving the three aspects of subsidiary, operational and strategic support for responsible innovation, its resources were insufficient to conduct some—or all—of the additional tasks we have outlined above. For instance, research on particular patient and physician preferences relevant to specific innovation projects has been planned, but could not be supported, let alone carried out, by the RI-P on a platform level. However, this kind of empirical work—and supporting it—has been shown to be a highly relevant but sometimes neglected part of the innovation work.

The RI-P has, however, been able to carry out research on expectations, preferences and expertise (concerning ethics) of the AI developers of the ecosystem to be able to adjust for the most promising tools to be devised for the subsidiary support in the ethical reflection of AI-based innovation in health.

By way of supporting the early adoption of the responsibility-by-design standard in one of the translational research projects within the consortium, it became clear that (i) a consensual RRI vision and (ii) acknowledging the epistemic deficits in tackling RRI challenges and allocating the proper resources for remedies would contribute to increased trustworthiness within the frame of a single innovation project. Setting this as a responsible innovation standard to be followed by many innovators within the ecosystem

would contribute to the trustworthiness of the ecosystem as a whole. However, additional ecosystem-level incentive structures are lacking, and additional resources for guiding through such processes still need to be accounted for. It also remains unclear how the ethical implementation of an algorithm can be continuously monitored, whether the ecosystem's governance can ensure this, and whether a company could be incentivised to spend the resources to do so.

6.7 Balancing institutional efforts with decentralized activities

In light of efforts to bring unity into activities that contribute to trustworthiness overall, care must be taken to avoid interference with and disruption of existing and effective trust dynamics. Actions of the larger ecosystem might be misunderstood or single failures, e.g., to provide transparency, may impair the ecosystem's overall reputation. It may turn out difficult to only engage in ecosystem-wide activities that supplement, rather than substitute working and entirely local trust relationships. The ecosystem should not create the illusion of unity out of motivations for better visibility or even marketing. Consequently, ecosystem governance should distribute responsibilities to the point that local caregivers own much of the direct interpersonal responsibilities while being backed up just enough such that they can rely on the technological instruments they use and be able to trust the technology providers, maintenance personnel, etc. This makes it clear that the trust relationships should not simply always target the patients as the final trustee, but rather that the relationships are hierarchical and aligned with responsibilities and the capacities to take over responsibility.

One cannot expect and want all stakeholders to conceive and recognize the ecosystem as a consistent whole. The whole point of reflecting upon institutional and non-institutional activities that actors within the ecosystem can do to contribute to its overall trustworthiness is to guide practitioners to choose from a wide range of possible actions that acknowledges and caters to the diverse stakeholders.

7 Limitations

Since we have mainly adopted the innovation ecosystem perspective offered by Stahl (2022), this paper suffers from a similar limitation as many of our propositions of thinking about innovation ecosystems and supporting their being declared trustworthy is not based on empirical evidence. Consequently, our exposition cannot provide proof that our concepts, or the concepts that we adopt, will prove productive in the sense of any kind of formally verifiable guarantee that trust relationships ensue. In fact, our work may even be only an initial step towards truly trustworthy ecosystems, because our arguments are thus that—under the constraints of the conceptualization of trustworthy technology as per Rieder et al. (2020) and Nickel et al. (2010)—trustworthiness of entities operating within the ecosystem may be plausibly increased (in the sense of being more likely). To speak of the entire ecosystem as trustworthy may, hence, be wrong. It may also not be possible to prove that the entire ecosystem is trustworthy. This, however, is also not the point of our contribution, as we seek to add to the discussions on ecosystem governance strategies that promote trustworthy conduct by building on Rieder's and Nickel's notion of trustworthy technology. We believe that the notion of trustworthiness that we have adopted here prompts new ideas in that regard. This does not foreclose that even more refined conceptualisations of trustworthiness may yield yet additional insights, or even that competing philosophies of trust (e.g., Ferrario et al. 2021) may demand entirely different approaches or can do without some of our proposed ones. It is thus, that we have not engaged in a full defense of the rational choice and motivation-attributing accounts as the only possible framework for conceptualizing trust. Rather, we have aimed to show its merits in medicine and in deriving plausible ecosystem governance strategies from it.

Another limitation of our contribution is that we do not engage with potential existential risks posed by (increasingly) autonomous AI as visibly discussed by notable figures like Yoshua Bengio, Geoffrey Hinton, Stuart Russell, Daniela Kahneman and others (Bengio et al. 2024). We believe that seriously engaging with governance structures of ecosystems that can provide guarantees that these kinds of harms are prevented is highly laudable and warrants a dedicated article (possibly more than one).

8 Conclusion

In this article, we have explored tenets that would constitute a responsible and trustworthy innovation ecosystem in AI for health. We have drawn on philosophical accounts of trust and the epistemic and moral requirements for trustworthiness first to discuss their significance in the medical domain. We have then applied these ideas to the ecosystem idea, for which we propose a platform-based governance structure that implements subsidiary, operational and strategic support functions, establishes a shared epistemic basis and connects to higher-order ecosystem structures on the national or global level. Finally, we have provided a closer look at one particular emerging medical AI ecosystem in Northern Germany and commented on its achievements and future potential to constitute a responsible and trustworthy innovation ecosystem.

We wanted to answer the research question: What constitutes a responsible AI ecosystem that can promote justifiable

trustworthiness in a healthcare environment? The answer to this question is not straightforward. We have shown that it calls for ecosystem governance structures and reporting processes that reflect the complexity of the ecosystem. This implies that different levels of ecosystems, from the local to the global level, are appropriately linked. Stakeholders and members of ecosystems need to be included and supported in productive interaction. There needs to be a balance between cooperation and competition within individual subsystems and between such subsystems. A core requirement is to establish transparency through an appropriate reporting culture. In addition, the various ethical concerns need to be recognised and addressed, which calls for distributing ethical responsibilities, which can be strengthened by introducing platform aspects into the ecosystem.

While these responses to our research question are a synthesis of ideas drawn from the literature and interpreted as helpful contributions to the responsible and trustworthy ecosystem concept, we have shown their practical relevance using the example of a real-life AI health ecosystem in Northern Germany. The example gives credibility to our claims but also shows that this work is only just beginning. We believe that the conceptual foundations we have provided here are a sound basis for more in-depth studies of our example case but also for other AI ecosystems.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflicts of interest None applicable.

References

AAMC Principles of Trustworthiness (2021) Association of American Medical Colleges. https://www.aamchealthjustice.org/our-work/trustworthiness/trustworthiness-toolkit

Alpay L, Verhoef J, Toussaint P (2006) What makes an "informed patient"? The impact of contextualization on the search for health information on the Internet. Stud Health Technol Inform 124:913–919

Andras P, Esterle L, Guckert M, Han A, Lewis PR, Milanovic K, Payne T, Perret C, Pitt J, Powers ST, Urquhart N, Wells S (2018) Trusting intelligent machines deepening trust within socio-technical systems. IEEE Technol Soc Mag 37(12):76–83. https://doi.org/10.1109/MTS.2018.2876107

Anoop VS, Asharaf S (2022) Integrating Artificial Intelligence and Blockchain for Enabling a Trusted Ecosystem for Healthcare Sector. In: Chakraborty C, Khosravi MR (Eds.), Intelligent Healthcare (pp. 281–295). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-8150-9_13

Ayling J, Chapman A (2022) Putting AI ethics to work: are the tools fit for purpose? AI and Ethics 2(3):405–429. https://doi.org/10.1007/s43681-021-00084-x

Baier A (1986) Trust and antitrust. Ethics 96(2):231–260

Bartneck C, Lütge C, Wagner A, Welsh S (2021) An introduction to ethics in robotics and AI. Springer Int Publish. https://doi.org/10.1007/978-3-030-51110-4

Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Darrell T, Harari YN, Zhang YQ, Xue L, Shalev-Shwartz S, Hadfield G, Clune J, Maharaj T, Hutter F, Baydin AG, McIlraith S, Gao Q, Acharya A, Krueger D, Mindermann S (2024) Managing extreme AI risks amid rapid progress. Science 384(6698):842–845. https://doi.org/10.1126/science.adn0117

Berg M, Goorman E (1999) The contextual nature of medical information. Int J Med Inform 56(1–3):51–60. https://doi.org/10.1016/S1386-5056(99)00041-6

Bertelsmann Foundation (2023) Trusted Health Ecosystems – Development of a national platform strategy for the healthcare system [Project Webpage]. Bertelsmann Foundation. https://www.bertelsmann-stiftung.de/en/our-projects/trusted-health-ecosystems/project-description

Binkley C (2021) The Physician's Conundrum: assigning moral responsibility for medical artificial intelligence and machine learning. Verdict — Legal Analysis and Commentary From Justia. https://verdict.justia.com/2021/02/08/the-physicians-conundrum

Birch J, Creel K, Jha A, Plutynski A (2022) Clinical decisions using AI must consider patient values. Nat Med 28(2):226–235

Bjerring JC, Busch J (2021) Artificial intelligence and patient-centered decision-making. Philos Technol 34(2):349–371. https://doi.org/10.1007/s13347-019-00391-6

Blank S, Mason C, Steinicke F, Herzog C (2024) Tailoring responsible research and innovation to the translational context: The case of AI-supported exergaming. Ethics Inform Technol. https://doi.org/10.1007/s10676-024-09753-x

Bockelmann N, Schetelig D, Kesslau D, Buschschlüter S, Ernst F, Bonsanto MM (2022) Toward intraoperative tissue classification: exploiting signal feedback from an ultrasonic aspirator for brain tissue differentiation. Int J Comput Assist Radiol Surg 17(9):1591–1599. https://doi.org/10.1007/s11548-022-02713-0

Bolte L, Van Wynsberghe A (2024) Sustainable AI and the third wave of AI ethics: a structural turn. AI Ethics. https://doi.org/10.1007/s43681-024-00522-6

Borenstein J, Grodzinsky FS, Howard A, Miller KW, Wolf MJ (2021) AI ethics: a long history and a recent burst of attention. Computer 54(01):96–102. https://doi.org/10.1109/MC.2020.3034950

Buchanan I (2021) Assemblage theory and method. In Assemblage Theory and Method. https://doi.org/10.5040/9781350015579

Branford J (2023) 'Experiencing AI and the Relational "Turn" in AI ethics' International Conference on Computer Ethics: Philosophical Enquiry (CEPE) Chicago IL

Bynum TW (2006) Flourishing Ethics. Ethics Inform Technol 8(4):157–173

Cabitza F, Rasoini R, Gensini GF (2017) Unintended consequences of machine learing in medicine. JAMA J Am Med Assoc 318(6):517–518. https://doi.org/10.1001/jama.2017.7797

Cahai A hoc C on AI (2022) Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law

(CM(2021)173-add). https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a4e8a5
Cardella V (2020) Rationality in mental disorders: Too little or too much? Eur J Anal Philos 16(2):13–36. https://doi.org/10.31820/ejap.16.2.1
CEN CWA 17796:2021 (2021) Responsibility-by-design—Guidelines to develop long-term strategies (roadmaps) to innovate responsibly (Standard CEN CWA 17796:2021). https://standards.iteh.ai/catalog/standards/cen/8e3cfe68-8449-49f8-b87c-d3efe20da158/cwa-17796-2021
Clarke R (2019) Principles and business processes for responsible AI. Comput Law Secur Rev 35(4):410–422
Danaher J, Nyholm S (2024) The ethics of personalised digital duplicates: a minimally viable permissibility principle. AI Ethics. https://doi.org/10.1007/s43681-024-00513-7
Depraetere I, Caët S, Debulpaep S, Ezzahid S, Janke V (2023) Building a child's trust before a medical procedure: a linguistic case study. Appl Linguist. https://doi.org/10.1093/applin/amad080
Digital Catapult (2020) Lessons in practical AI ethics: Taking the UK's AI ecosystem from 'what' to 'how.' Digital Catapult. https://assets.ctfassets.net/nubxhjiwc091/xTEqMcYudwQ7GHZWNoBfM/c2a2d55a0ee1694e77634e240eafdfdf/20200430_DC_143_EthicsPaper__1_.pdf
Dignum V (2019) Responsible artificial intelligence (O'Sullivan B, Woolridge M), (Eds.). Springer. https://doi.org/10.1007/978-3-030-30371-6
Dubber MD, Pasquale F, Das S (2020) The Oxford handbook of ethics of AI. Oxford University Press
Duenser A, Douglas DM (2023) Whom to trust, How and Why: untangling artificial intelligence ethics principles, trustworthiness, and trust. IEEE Intell Syst 38(6):19–26. https://doi.org/10.1109/MIS.2023.3322586
Durán JM, Formanek N (2018) Grounds for trust: essential epistemic opacity and computational reliabilism. Mind Mach 28(4):645–666. https://doi.org/10.1007/s11023-018-9481-6
Durán JM, Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical. AI J Med Ethics. https://doi.org/10.1136/medethics-2020-106820
Edelman GmbH (2023) The 2023 Edelman Trust Barometer—Special Report: Trust and Health. https://www.edelman.com/trust/2023/trust-barometer/special-report-health
European Commission (2020) On Artificial Intelligence—A European approach to excellence and trust (COM(2020) 65 final). European Commission
European Commission (2021a) Proposal for a Regulation laying down harmonised rules on artificial intelligence (COM(2021) 206 final). European Commission. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence
European Commission (2021b) Proposal for a Regulation on a European approach for Artificial Intelligence (COM(2021) 206 final). European Commission. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence
Ferrario A, Loi M, Viganò E (2021) Trust does not need to be human: It is possible to trust medical AI. J Med Ethics 47(6):437–438. https://doi.org/10.1136/medethics-2020-106922
Findlay M, Seah J (2020) An ecosystem approach to ethical AI and data use: experimental reflections. 2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G) 192–197. https://doi.org/10.1109/AI4G50087.2020.9311069
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. https://dash.harvard.edu/handle/1/42160420
Flick C, Zamani ED, Stahl BC, Brem A (2020) The future of ICT for health and ageing: unveiling ethical and social issues through horizon scanning foresight. Technol Forecast Soc Chang 155:119995. https://doi.org/10.1016/j.techfore.2020.119995
Floridi L (2019a) Establishing the rules for building trustworthy AI. Nat Mach Intellig 1(6):261–262. https://doi.org/10.1038/s42256-019-0055-y
Floridi L (2019b) Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol 32(2):185–193. https://doi.org/10.1007/s13347-019-00354-x
Gardiner R (2008) The transition from "informed patient" care to "patient informed" care. Stud Health Technol Inform 137:241–256
Gawer A, Cusumano MA (2014) Industry platforms and ecosystem innovation: platforms and innovation. J Prod Innov Manag 31(3):417–433. https://doi.org/10.1111/jpim.12105
Gillespie N, Lockey S, Curtis C, Pool J, Ali Akbari (2023) Trust in artificial intelligence: A global study. The University of Queensland; KPMG Australia. https://doi.org/10.14264/00d3c94
Gillespie T (2019) Systems engineering for ethical autonomous systems. Instit Eng Technol. https://doi.org/10.1049/SBRA517E
Guterres A (2020) The Highest Aspiration—A Call to Action for Human Rights. Unitted Nations. https://www.un.org/sg/sites/www.un.org.sg/files/atoms/files/The_Highest_Asperation_A_Call_To_Action_For_Human_Right_English.pdf
Hadwick D, Lan S (2021) Lessons to be learned from the Dutch childcare allowance scandal: a comparative review of algorithmic governance by tax administrations in The Netherlands, France and Germany. World Tax J https://doi.org/10.59403/27410pa
Hall W, Pesenti J (2017) Growing the artificial intelligence industry in the UK. Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf
Hallensleben S, Hustedt C, Fetic L, Fleischer T, Grünke P, Hagendorff T, Hauer M, Hauschke A, Heesen J, Herrmann M, Hillerbrand R, Hubig C, Kaminski A, Krafft T, Loh W, Otto P, Puntschuh M (2020) From principles to practice—an interdisciplinary framework to operationalise AI ethics. Bertelsmann Stiftung; VDE
Hansen L, Siebert M, Diesel J, Heinrich MP (2019) Fusing information from multiple 2D depth cameras for 3D human pose estimation in the operating room. Int J Comput Assist Radiol Surg 14(11):1871–1879. https://doi.org/10.1007/s11548-019-02044-7
Haque A, Milstein A, Fei-Fei L (2020) Illuminating the dark spaces of healthcare with ambient intelligence. Nature. https://doi.org/10.1038/s41586-020-2669-y
Hatherley JJ (2020) Limits of trust in medical AI. J Med Ethics 46(7):478–481. https://doi.org/10.1136/medethics-2019-105935
Hengstler M, Enkel E, Duelli S (2016) Applied artificial intelligence and trust-The case of autonomous vehicles and medical assistance devices. Technol Forecast Soc Chang 105:105–120. https://doi.org/10.1016/j.techfore.2015.12.014
Heilinger J-C (2022) The ethics of AI ethics. A constructive critique. Philos Technol 35(3):61. https://doi.org/10.1007/s13347-022-00557-9
Herzog C (2019) Technological opacity of machine learning in healthcare. 2nd Weizenbaum Conference: Challenges of Digital Inequality - Digital Education, Digital Work, Digital Life. https://doi.org/10.34669/wi.cp/2.7
Herzog C (2022a) On the ethical and epistemological utility of explicable AI in medicine. Philos Technol 35(50):31. https://doi.org/10.1007/s13347-022-00546-y
Herzog C (2022b) Inexplicable AI in medicine as a form of epistemic oppression. IEEE International Symposium on Technology and

Society, IEEE International Symposium on Technology and Society, Hong Kong, Hong Kong
Ibáñez JC, Olmeda MV (2022) Operationalising AI ethics: How are companies bridging the gap between practice and principles? Explor Study AI Soc 37(4):1663–1687. https://doi.org/10.1007/s00146-021-01267-0
IEEE Computer Society (2021) IEEE standard model process for addressing ethical concerns during system design—7000–2021 (7000–2021) [Standard]. https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html
Independent High-Level Expert Group on Artificial Intelligence Set Up By the European Commission (2019) Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation
Iqbal S, Altaf W, Aslam M, Mahmood W, Khan MUG (2016) Application of intelligent agents in health-care: review. Artif Intell Rev 46(1):83–112. https://doi.org/10.1007/s10462-016-9457-y
Jelinek T, Wallach W, Kerimi D (2020) Policy brief: The creation of a G20 coordinating committee for the governance of artificial intelligence. AI Ethics. https://doi.org/10.1007/s43681-020-00019-y
Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intellig 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2
Karnam S (2017) Hybrid doctors: the need risen from informed patients. J Clin Diagnost Res. https://doi.org/10.7860/JCDR/2017/23163.9200
Kazim E, Koshiyama AS (2021) A high-level overview of AI ethics. Patterns. https://doi.org/10.1016/j.patter.2021.100314
Kempt H, Freyer N, Nagel SK (2022) Justice and the normative standards of explainability in healthcare. Philos Technol 35(4):100. https://doi.org/10.1007/s13347-022-00598-0
Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. LIPIcs 67, 43:1–43:23. https://doi.org/10.4230/LIPICS.ITCS.2017.43
Lane M, Williams M, Broecke S (2023) The impact of AI on the workplace: main findings from the OECD AI surveys of employers and workers (OECD Social, Employment and Migration Working Papers No. 288; OECD Social, Employment and Migration Working Papers, Vol. 288). https://doi.org/10.1787/ea0a0fe1-en
Laugharne R, Priebe S, McCabe R, Garland N, Clifford D (2012) Trust, choice and power in mental health care: Experiences of patients with psychosis. Int J Soc Psychiatry 58(5):496–504. https://doi.org/10.1177/0020764011408658
Laux J, Wachter S, Mittelstadt B (2023) Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. Regul Governance. https://doi.org/10.1111/rego.12512
Lehoux P, Roncarolo F, Silva HP, Boivin A, Denis J-L, Hébert R (2019) What health system challenges should responsible innovation in health address? Insights From an International Scoping Review. Int J Health Policy Manag 8(2):63–75. https://doi.org/10.15171/ijhpm.2018.110
London AJ (2022) Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care? Cell Rep Med 3(5):100622. https://doi.org/10.1016/j.xcrm.2022.100622
Mantelero A, Esposito MS (2021) An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. Comput Law Secur Rev 41:105561. https://doi.org/10.1016/j.clsr.2021.105561
Manzeschke A (2015) MEESTAR: Ein Modell zur ethischen Evaluierung sozio-technischer Arrangements in der Pflege- und Gesundheitsversorgung. In: Weber K, Frommeld D, Manzeschke A, Fangerau H (Eds.), Technisierung des Alltags—Beitrag für ein gutes Leben? (pp. 263–283). https://elibrary.steiner-verlag.de/book/99.105010/9783515110099
Markus ML (2001) Toward a theory of knowledge reuse: types of knowledge reuse situations and factors in reuse success. J Manag Inf Syst 18(1):57–93. https://doi.org/10.1080/07421222.2001.11045671
Martinho A, Kroesen M, Chorus C (2021) A healthy debate: exploring the views of medical doctors on the ethics of artificial intelligence. Artif Intell Med 121:102190. https://doi.org/10.1016/j.artmed.2021.102190
McDougall RJ (2019) Computer knows best? The need for value-flexibility in medical AI. J Med Ethics 45(3):156–160. https://doi.org/10.1136/medethics-2018-105118
McLennan S, Fiske A, Celi LA, Müller R, Harder J, Ritt K, Haddadin S, Buyx A (2020) An embedded ethics approach for AI development. Nat Mach Intellig 2(9):488–490. https://doi.org/10.1038/s42256-020-0214-1
Minkkinen M, Zimmer MP, Mäntymäki M (2021) Towards ecosystems for responsible AI: expectations on sociotechnical systems, agendas, and networks in EU documents. In: Dennehy D, Griva A, Pouloudi N, Dwivedi YK, Pappas I, Mäntymäki M (Eds.), Responsible AI and analytics for an ethical and inclusive digitized society (Vol. 12896, pp. 220–232). Springer International Publishing. https://doi.org/10.1007/978-3-030-85447-8_20
Montemayor C, Halpern J, Fairweather A (2022) In principle obstacles for empathic AI: Why we can't replace human empathy in healthcare. AI Soc 37(4):1353–1359. https://doi.org/10.1007/s00146-021-01230-z
Moore JF (1993) Predators and prey: a new ecology of competition. Harvard Business Review 75–86
Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L (2021a) Ethics as a service: a pragmatic operationalisation of AI ethics. Mind Mach 31(2):239–256. https://doi.org/10.1007/s11023-021-09563-w
Morley J, Floridi L, Kinsey L, Elhalal A (2021b) From What to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In: Floridi L (Ed.), Ethics, governance, and policies in artificial intelligence (pp. 153–183). Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-1_10
Nail T (2017) What is an assemblage? SubStance 46(1):21–37. https://doi.org/10.3368/ss.46.1.21
Nazarko L (2020) Responsible research and innovation in enterprises: benefits, barriers and the problem of assessment. J Open Innova Technol Market Complex 6(1):12. https://doi.org/10.3390/joitmc6010012
Nickel PJ, Franssen M, Kroes P (2010) Can we make sense of the notion of trustworthy technology? Knowl Technol Policy 23(3–4):429–444. https://doi.org/10.1007/s12130-010-9124-6
Nishant R, Kennedy M, Corbett J (2020) Artificial intelligence for sustainability: challenges, opportunities, and a research agenda. Int J Inf Manage 53:102104. https://doi.org/10.1016/j.ijinfomgt.2020.102104
NIST (2022) AI risk management framework: second draft. https://www.nist.gov/document/ai-risk-management-framework-2nd-draft
OECD (2019) Recommendation of the Council on Artificial Intelligence [OECD Legal Instruments]. OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
Perry A (2023) AI will never convey the essence of human empathy. Nat Hum Behav 7(11):1808–1809. https://doi.org/10.1038/s41562-023-01675-w
Petersen E, Potdevin Y, Mohammadi E, Zidowitz S, Breyer S, Nowotka D, Henn S, Pechmann L, Leucker M, Rostalski P, Herzog C (2022) Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions. IEEE Access 10:58375–58418. https://doi.org/10.1109/ACCESS.2022.3178382

Petersen E, Ferrante E, Ganz M, Feragen A (2023) Are demographically invariant models and representations in medical imaging fair? (arXiv:2305.01397). arXiv. http://arxiv.org/abs/2305.01397

Petkovic D, Kobzik L, Ghanadan R (2020) AI ethics and values in biomedicine technical challenges and solutions. Pac Symp Biocomput 25(2020):731–735. https://doi.org/10.1142/9789811215636_0064

Platt J, Nong P (2023) An ecosystem approach to earning and sustaining trust in health care—too big to care. JAMA Health Forum 4(1):e224882. https://doi.org/10.1001/jamahealthforum.2022.4882

Porcari A, Pimponi D, Borsella E, Mantovani E (2019) PRISMA RRI-CSR Roadmap. 710059

Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. Nat Med. https://doi.org/10.1038/s41591-021-01614-0

Rampton V, Böhmer M, Winkler A (2022) Medical technologies past and present: how history helps to understand the digital era. J Med Human 43(2):343–364. https://doi.org/10.1007/s10912-021-09699-x

Read L, Korenda L, Nelson H (2021) Rebuilding trust in health care. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/health-care/trust-in-health-care-system.html

Reijers W, Calvo A, Lewis D, Levacher K (2016) The ethics canvas: a tool for practising ethics in responsible research and innovation. 21st International Conference on Applications of Natural Language to Information Systems, Salford, UK. www.adaptcentre.ie

Ricci Lara MA, Echeveste R, Ferrante E (2022) Addressing fairness in artificial intelligence for medical imaging. Nat Commun. https://doi.org/10.1038/s41467-022-32186-3

Rieder G, Simon J, Wong P-H (2020) Mapping the stony road toward trustworthy AI: expectations, problems, conundrums. SSRN Electron J. https://doi.org/10.2139/ssrn.3717451

Ronicke S, Hirsch MC, Türk E, Larionov K, Tientcheu D, Wagner AD (2019) Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. Orphanet J Rare Dis 14(1):69. https://doi.org/10.1186/s13023-019-1040-6

Ruokonen F (2013) Trust, trustworthiness, and responsibility. In: Mäkelä P, Townley C (Eds.), Trust: analytic and applied perspectives (pp. 1–14). BRILL. https://doi.org/10.1163/9789401209410

Ruotsalainen P, Blobel B (2020) Health information systems in the digital health ecosystem—problems and Solutions for ethics, trust and privacy. Int J Environ Res Public Health 17(9):3006. https://doi.org/10.3390/ijerph17093006

Ruotsalainen P, Blobel B (2022) Transformed health ecosystems—challenges for security, privacy, and trust. Front Med 9:827253. https://doi.org/10.3389/fmed.2022.827253

Ryan M (2020) In AI We trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 26(5):2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Sand M, Durán JM, Jongsma KR (2022) Responsibility beyond design: physicians' requirements for ethical medical AI. Bioethics 36(2):162–169. https://doi.org/10.1111/bioe.12887

Sarasohn-Kahn J (2022) People have lost trust in healthcare systems because of COVID. How can the damage be healed? World Economic Forum. https://www.weforum.org/agenda/2022/03/trust-health-economy-pandemic-covid19

Schiff D, Rakova B, Ayesh A, Fanti A, Lennon M (2020) Principles to practices for responsible AI: closing the gap (arXiv:2006.04707). arXiv. http://arxiv.org/abs/2006.04707

Siau K, Wang W (2020) Artificial intelligence (AI) ethics: ethics of AI and ethical AI. J Database Manag (JDM) 31(2):74–87

Sisk B, Baker JN (2019) A model of interpersonal trust, credibility, and relationship maintenance. Pediatrics 144(6):e20191319. https://doi.org/10.1542/peds.2019-1319

Stahl BC (2021) Artificial intelligence for a better future: an ecosystem perspective on the Ethics of AI and emerging digital technologies. Springer International Publishing. https://doi.org/10.1007/978-3-030-69978-9

Stahl BC (2022) Responsible innovation ecosystems: Ethical implications of the application of the ecosystem concept to artificial intelligence. Int J Inf Manage 62:102441. https://doi.org/10.1016/j.ijinfomgt.2021.102441

Stahl BC (2023) Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems. Sci Rep 13(1):7586. https://doi.org/10.1038/s41598-023-34622-w

Stahl BC, Andreou A, Brey P, Hatzakis T, Kirichenko A, Macnish K, Laulhé Shaelou S, Patel A, Ryan M, Wright D (2021) Artificial intelligence for human flourishing – Beyond principles for machine learning. J Bus Res 124:374–388. https://doi.org/10.1016/j.jbusres.2020.11.030

Stahl BC, Antoniou J, Bhalla N, Brooks L, Jansen P, Lindqvist B, Kirichenko A, Marchal S, Rodrigues R, Santiago N, Warso Z, Wright D (2023) A systematic review of artificial intelligence impact assessments. Artif Intell Rev. https://doi.org/10.1007/s10462-023-10420-8

Stahl BC, Rodrigues R, Santiago N, Macnish K (2022) A European agency for artificial intelligence: protecting fundamental rights and ethical values. Comput Law Secur Rev 45:105661. https://doi.org/10.1016/j.clsr.2022.105661

Shteynberg G, Halpern J, Sadovnik A, Garthoff J, Perry A, Hay J, Montemayor C, Olson MA, Hulsey TL, Fairweather A (2024) Does it matter if empathic AI has no empathy? Nat Mach Intellig 6(5):496–497. https://doi.org/10.1038/s42256-024-00841-7

The European Parliament and the Council of the European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council—Artificial Intelligence Act. Off J Eur Union. https://doi.org/10.5040/9781782258674

The State Chancellery of Schleswig-Holstein, Germany (2021) Artificial Intelligence – Strategic objectives and areas of activity for Schleswig-Holstein, Version 2.0 (p. 40) [Political Strategic Agenda]. https://www.schleswig-holstein.de/DE/Landesregierung/Themen/Digitalisierung/Kuenstliche_Intelligenz/KI_Strategie/_documents/ki__ai_strategy_download.pdf?__blob=publicationFile&v=3

Topol EJ (2019) High-performance medicine: The convergence of human and artificial intelligence. Nat Med. https://doi.org/10.1038/s41591-018-0300-7

Tsujimoto M, Kajikawa Y, Tomita J, Matsumoto Y (2018) A review of the ecosystem concept—Towards coherent ecosystem design. Technol Forecast Soc Chang 136:49–58. https://doi.org/10.1016/j.techfore.2017.06.032

UK Government (2021) National AI Strategy. https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version

UK Government (2023) A pro-innovation approach to AI regulation (Command Pape CP 815). Department for Science, Innovation and Technology. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

Ulnicane I, Knight W, Leach T, Stahl BC, Wanjiku W-G (2021) Framing governance for a contested emerging technology:insights from AI policy. Policy Soc 40(2):158–177. https://doi.org/10.1080/14494035.2020.1855800

UNESCO (2020) First draft of the recommendation on the Ethics of Artificial Intelligence (SHS / BIO / AHEG-AI / 2020/4 REV.2). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000373434

van den Eede Y (2011) In Between Us: On the Transparency and opacity of technological mediation. Found Sci 16(2–3):139–159. https://doi.org/10.1007/s10699-010-9190-y

Vayena E, Blasimme A (2018) Health research with big data: time for systemic oversight. J Law Med Ethics 46(1):119–129. https://doi.org/10.1177/1073110518766026

Verbeek P (2006) Materializing morality: design ethics and technological mediation. Sci Technol Human Values 31(3):361–380

Walden D, Roedler G, Forsberg K, Hamelin RD, Shortell T (2015) INCOSE systems engineering handbook—A guide for system life cycle processes and activities (4th ed.). International Council on Systems engineering (INCOSE)

Winfield AFT, Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos Trans Royal Soc: Math Phys Eng Sci 376(2133):20180085. https://doi.org/10.1098/rsta.2018.0085