International Review of Law, Computers & Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cirl20

Governing fiduciary relationships or building up a governance model for trust in AI? Review of healthcare as a socio-technical system

Mehmet B. Unver

To cite this article: Mehmet B. Unver (2023) Governing fiduciary relationships or building up a governance model for trust in AI? Review of healthcare as a socio-technical system, International Review of Law, Computers & Technology, 37:2, 198-226, DOI: 10.1080/13600869.2023.2192569

To link to this article: https://doi.org/10.1080/13600869.2023.2192569

Published online: 29 Mar 2023.

INTERNATIONAL REVIEW OF LAW, COMPUTERS & TECHNOLOGY
2023, VOL. 37, NO. 2, 198–226
https://doi.org/10.1080/13600869.2023.2192569

Routledge
Taylor & Francis Group

Governing fiduciary relationships or building up a governance model for trust in AI? Review of healthcare as a socio-technical system

Mehmet B. Unver

Hertfordshire Law School, University of Hertfordshire, Hatfield, UK

ABSTRACT

': Fiduciary law aims to mitigate the inherent risk of 'trust', which helps restore interpersonal trust. It remains to be answered how trust should be governed in an AI-driven socio-technical system where technical and social factors are involved including interpersonal relationships and AI-human interactions. Taking interpersonal trust as the backdrop of analysis, this article seeks answers to this question focusing on healthcare. It firstly draws a conceptual framework regarding 'trust' and investigates its interplay with AI as well as examines how it is governed under the fiduciary law. Subsequently, it upholds a socio-technical system perspective, examining how to enable and sustain trust in an AI-driven socio-technical system. A governance model is then developed to elicit 'intrinsic', 'dynamic' and 'ethical' values of trust attributed to various elements under a tri-partite framework. It is recognised that findings of the literature as to trust, its trajectory and implications can be implemented within the proposed framework. Furthermore, it brings novelty by re-conceptualising the elements of 'trust' and associated values, marking distinction to its interpersonal roots and fiduciary relationships. It is considered this governance model, by upholding a holistic viewpoint, provides a generalisable framework that can construct, maintain and restore trust in AI-driven socio-technical systems.

KEYWORDS
Trust; fiduciary law; artificial intelligence

Introduction

'Trust' is 'a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another' (Rousseau et al. 1998, 395). There are widely acknowledged elements that constitute 'trust' such as 'perceived competence', 'perceived benevolence' and 'perceived integrity' (Rousseau et al. 1998; Meijer and Grimmelikhuijsen 2020). For instance, if person A trusts person B to act for or on behalf of himself/herself, B is expected to put A's interests first based on

CONTACT Mehmet B. UNVER m.unver@herts.ac.uk Hertfordshire Law School, University of Hertfordshire, De Havilland Campus, Mosquito Way, Hatfield, AL10 9EU, UK

the trust and the presumed trust relationship between them. This also reflects the approach taken by the fiduciary law in its conception of obligations and remedies towards the unexpected behaviours of the 'trustee'. For instance, in a scenario where the B (trustee) betrays A (trustor) failing to meet his/her expectations, common law imposes equity remedies for the violation of fiduciary obligations, e.g. duty of loyalty, no profit rule, no conflict rule.

Fiduciary law can be taken as a baseline as to how trust is governed with respect to interpersonal relationships, e.g. via restorative remedies for the unexpected consequences of a trust relationship. This part of English common law consist of not only legal but also moral and ethical obligations on the fiduciaries to whom trust is reposed by others. While trust is inherent in law and ethics, this relationship needs to be checked out against the impact of AI, in particular concerning socio-technical systems, where there are a number of actors engaging with AI. Against this background, one can ask not only how the legal concept of interpersonal trust interacts with the real-world scenarios of AI but also how to understand the trust in an AI-driven socio-technical system, given the need to enable trust against the human and non-human factors involved in it.

Seeking answers to the above questions, this article firstly sets out a definitional framework regarding 'trust' and examines different approaches to define 'trust' conceptually, mainly from the perspective of moral psychology, and legally, based on the English fiduciary law. It then examines trust in (relation to) AI and from the socio-technical perspective after drawing a framework of socio-technical system theory and its application AI-driven context. After all, a model is built up to demonstrate how trust can be governed from a socio-technical perspective with a review of healthcare sector.

Given the restraints of fiduciary law, e.g. focused on interpersonal trust from a restorative viewpoint, and the fact that cutting-edge AI technologies that can erode trust, a broader perspective is upheld. This is mainly to investigate how trust should be governed in an AI-driven socio-technical system where technical and social factors are involved including interpersonal relationships and AI-human interactions. In this regard, expanding the idea of mitigating the risks inherent within the 'trust' relationships towards enabling and sustaining trust during the AI life cycle, the article focuses on the main drivers to elicit the full value of trust given the role(s) taken by AI agents and their capabilities in a socio-technical system.

Against this background, this article makes the discussion based on the domain of healthcare as a socio-technical system, considering that in this sector AI is employed by multiple actors, e.g. clinicians, for various tasks, e.g. diagnosis of diseases, personalised medicine, patient care monitoring, disease management using robotics, based on interaction with the stakeholders, e.g. manufacturers, insurers, towards the provision of healthcare to the patients. In this regard, the article delves into the interaction of human and non-human factors and derivation of full value from overall trust in healthcare as a socio-technical system.

After all, a governance model is developed finding that overall trust can be distilled from the interaction of all the (healthcare) actors via pertinent tools, safeguards and remedies to ensure reliability, transparency, adaptiveness, competence, responsibility and accountability. These elements of trust are categorised under a tri-partite framework including (i) elements of intrinsic value, (ii) elements of dynamic value and (iii) elements

of ethical value. As demonstrated in the healthcare domain, the concept of 'trust' is primarily predicated on 'reliability' and 'transparency' which provide the *intrinsic* value of it. On top of these antecedents, 'trust' can develop with a *dynamic* value to be derived from 'adaptiveness' and 'competence'. Finally, 'responsibility' and 'accountability' of the actors can bring about *ethical* value.

On this basis, it is concluded the proposed model can enable and sustain 'trust', eliciting its full value in an AI-driven socio-technical system. It is recognised that findings of the literature as to 'trust', its trajectory and implications can be implemented within this framework. Given all this, it is considered this governance model, by upholding a holistic viewpoint, provides a generalisable framework that can construct, maintain and restore trust in AI-driven socio-technical systems.

Trust: conceptual and legal framework

Overview of 'Trust': focus on interpersonal trust

According to Rousseau et al. (1998, 395) 'trust' is 'a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another'. In case trustee, namely the person who is trusted, fails to act the way s/he is expected, the trustor would then become betrayed. This level of risk or vulnerability accepted by the trustor is a pre-requisite for trust, although having a degree based on the trustworthiness of the trustee. Trust is an attitude that we have towards people whom we hope will be trustworthy, where trustworthiness is a property, not an attitude (McLeod 2021). Which characteristics a trustor takes to be important for assessing a trustee's trustworthiness may depend on the characteristics of the trustor (sometimes called trustfulness), the situation in which the interaction takes place, the nature of the task, the type of agent the trustor is engaging with, and the propensity of the trustor to engage in trust relationships (Starke et al. 2022). Given this fact, trust is situation-specific, and can be described with the following parameters: 'A trustor A that trusts (judges the trustworthiness of) a trustee B with regard to some behaviour X in context Y at time t' (Sharan and Romano 2020, 2).

Interpersonal trust establishes the ground upon which the definitional framework is set up for trust from the beginning, although it is widely acknowledged institutional trust, trust in government, trust in AI and self-trust exist (McLeod 2021). All the definitions of trust assume the presence of some form of positive expectation regarding the intentions and behaviour of the object of trust (Rousseau et al. 1998). The three most cited elements of trust are 'perceived competence', 'perceived benevolence' and 'perceived integrity' (also sometimes called honesty) (Meijer and Grimmelikhuijsen 2020). All these perceived elements together demonstrate that trust is a subjective and elusive concept also being dependent on a variety of factors, mainly based on past experiences and future optimism, along with the degree of vulnerability and dependency on the part of each party.

According to McLeod, trust is 'warranted' when it is justified or well-grounded, where well-grounded trust successfully targets a trustworthy person (McLeod 2021). In her view, trust happens to be justified (i) sometimes when the trustee is not in fact trustworthy, which suggests that the epistemology of trust is relevant, or (ii) often because some value will emerge from the trust or because it is valuable in and of itself (McLeod

2021). Last but not least, trust is considered to be 'plausible' when people feel optimism toward one another, hoping by such trust to elicit, in the fullness of time, more responsible and responsive trustworthy behaviour (McLeod 2021). A counter claim is that such trust involves the normative attitude that the trustee ought to do what one trusts him or her to do, rather than optimism that s/he will do it (McLeod 2021).

For instance, when P needs private medical advice and for this purpose consults D, this means P has a trust in D with regard to the specific matter(s) of consultation. P might have in mind that D specialises in a particular area of medicine being a qualified doctor and would have an *epistemic* trust based on this belief (Faulkner 2011; Origgi 2020). When D's advice works out with a satisfactory result, there arises a clear justification for P to believe D is a trustworthy doctor. Before this, P's philosophical state might be closely connected to others' belief and testimony, e.g. in the community where s/he lives. If this belief is sourced from a social norm or expectation impacting P's decision, this means *normative* trust exists (Faulkner 2011; Origgi 2020; Carter 2020; Ryan 2020; Nickel 2022). Regardless of any social or epistemological ground, P's previous appointments might have gone well making her/him to have a strong belief that D will continue to behave in the same professional and trustworthy manner. Then we can assume there is a *predictive* trust in this attitude of P asking for D's advice each time s/he needs (Faulkner 2011). It is also noteworthy that a P's decisions in building trust to D might be affected by several reasons including (positive) bias, which can result in *affective* trust possibly having some epistemological features in it (Faulkner 2011; Origgi 2020; Ryan 2020; Nickel 2022).

Against this background, some distinguish 'emotional' and 'cognitive' constructs of trust while discussing trust in AI (Glikson and Wooley 2020). From this point of view, trust would sustain either on the cognitive construct that involves rational evaluation of the trustee and situational features or on the emotional construct that is mostly built upon more human emotions or both (Glikson and Wooley 2020). Another distinction can be made as to the 'normative' and 'discretionary' aspects of the trust (Nickel 2022). According to Nickel, one accords another entity discretionary authority because one has relevant normative and predictive expectations toward it; following on this, a user trusts an AI application when s/he is disposed to give it discretionary authority over practically important questions on the basis of normative and predictive expectations about its performance in context (Nickel 2022). Albeit with such different accounts and underlying factors to build trust, the 'value of trust' arguably surfaces most when the trustworthiness (of AI) is visible and proven usually based on both cognitive decisions and behavioural interactions.

On the opposite side, Ryan (2020) rejects the idea of developing trust in AI, arguing AI cannot be something that has the capacity to be trusted for it does not possess emotive states or cannot be held responsible for their actions. In his view, AI should not be viewed as trustworthy because it undermines the value of interpersonal trust, anthropomorphises AI (the affective account of trust), and diverts responsibility from those developing and using AI (the normative account of trust) (Ryan 2020). As a contrary argument, responsibility would be regarded as a factor to influence development and maintenance of trust, as predicated in the proposed governance model within this article. Likewise, Lewis and Marsh (2022, 35) posit that 'trust is about much more than just attribution of responsibility or blame'.

Against the above background one can grapple to estimate which trust is the one we develop across to a device or service driven by AI. AI's peculiar features such as working with high computation and data but no consciousnesses make the trust in relation to AI more elusive and hardly recognisable. There are many aspects that can indicate the nature of such a trust, including the nature of AI service or function expected by the user, the context AI is embedded, whether the AI is hidden or visible as well as the disposition and needs of the trustor. Not delving into such factors and their investigation, this article aims at examining how to enable trust in a socio-technical system driven by AI comparing to interpersonal trust and its reflections under fiduciary law. To that end, this study looks into fiduciary relationships and related duties and remedies under English common law for they take 'interpersonal trust' as granted and prescribe a legal framework to govern it.

'Trust' in fiduciary law: how interpersonal trust is governed under law?

Trust means a relationship between 'trustor' and 'trustee' embodying some risks associated with the former's dependence to the latter. If this concludes with betrayal of the former, this would have some legal consequences which are manifested well in fiduciary law. Given this, this part of the article is focused on fiduciary duties and responsibilities and the consequences of failure of them under common law, mainly to shed light on the interplay between trust and law. This is also considered to benefit upholding a wider perspective in relation to how to govern trust in a socio-technical system where many agents co-work towards a common task for which trust is developed by the users or beneficiaries. While the beneficiaries also partake in fiduciary relationships, the 'trust' they develop in interpersonal interactions and in socio-technical systems might differ from each other. Despite this fact, considering the latter embodies the former, fiduciary law would offer guidance before upholding a socio-technical perspective.

Fiduciary law refers to common law principles and rules governing relationship between a fiduciary and beneficiary. 'Fiduciary' is a person who is entrusted with a responsibility to pursue the interests of a 'beneficiary' for the interpersonal trust developed between them. Fiduciary rules under common law systems mitigate the risks and vulnerabilities inherent in such trust relationships. Inspired by fiduciary rules, law and policy makers often adopt comparable rules to govern relationships based on interpersonal trust, e.g. between fiduciaries such as doctors, lawyers, financial advisors and beneficiaries such as patients, clients and businesses. Given this, fiduciary law is taken as the reference in this study to find out how trust is governed under English common law.

Historically, under fiduciary law, when a person (A) acts for, or on behalf of, another person (B) in circumstances which give rise to a relationship of trust and confidence,[1] A is B's fiduciary and owes B fiduciary duties.[2] A fiduciary is not said to be subject to the rules of the duties because he occupies the position of a fiduciary in a relationship, but rather he is said to be a fiduciary because he is subject to the rules of the duties for the purposes of that relationship (Atkins 2017). A fiduciary relationship may arise on an ad hoc basis if the relationship in question has the requisite attributes, whereas there is a presumption that fiduciary duties are owed within certain categories of relationships, i.e. between trustee and beneficiary, agent and principal, partner and co-partner, solicitor and client, director and company, promoter and company, although rebuttable.[3] Courts

have discretion to decide whether such a relationship has arisen along with the fiduciary duties and obligations considering each case within its entirety of circumstances. It has thus far been possible for professional advisers, bank managers, mortgagees, doctors and employees to be subject to fiduciary obligations even though they do not fit into the recognised relationships which are fiduciary *per se* (Panesar 2005).

There are a number of fiduciary duties arising out of such relationships based on trust and confidence. The core fiduciary duty is the obligation of loyalty, which distinguishes fiduciary law duties from other common law duties.[4] This core fiduciary obligation, having a prescriptive nature, requires that the fiduciary pursue the best interests of his principal or beneficiary, that he will be loyal to the interests of another (Practical Law Corporate 2022). The essential idea is that fiduciaries are not permitted to use their positions for their own private advantage but are required to act unselfishly in what they perceive to be the best interests of the other person (Practical Law Corporate 2022). Fiduciaries must act only in the interests of the other person in the fiduciary relationship and not adversely to them, and in particular must subordinate any personal interests they have.[5]

From an early authority of equity law, two accompanying fiduciary duties arise as follows:

> It is an inflexible rule of the Court of Equity that a person in a fiduciary position […] is not unless otherwise expressly provided for, entitled to make a profit; he is not allowed to put himself in a position where his duty and interest conflict. (Lord Herschell, *Bray v Ford* [1896] AC 44 para. 57)

Two key fiduciary duties can be derived out of this landmark case: 'duty to avoid conflicts of interest' ('no conflict of interest' rule) and 'duty to avoid unauthorised or secret profits' ('no profit' rule), which are proscriptive in nature. The former requires fiduciaries not place themselves in a position where their duty and interest conflict or where there is a real possibility they may conflict.[6] The latter means fiduciary must not make a profit or allow his own interests to run contrary to those of his beneficiary in the absence of express and valid authorisation (Atkins 2017).

There are various obligations in the field of company law that are originated from these two particular fiduciary duties being imposed on company directors, as set out in Company Act 2006. Not being limited to company law or corporate law, there are various areas of law heavily affected or governed by fiduciary law such as trust law, agency law and professional practice (Velasco 2021). For instance, delivery of financial services is regulated under the under Financial Services Act 2021 incorporating fiduciary type obligations when provided by financial advisors and firms.[7] In such cases of application, fiduciary duties are often extended with duty of good faith, duty of care, duty of confidentiality and full disclosure (Practical Law Corporate 2022; Velasco 2021; Hall 2019) or with some modifications over the legacy obligations.

Among such obligations, notably, it is the obligation of loyalty which is unique to a relationship in which is found a fiduciary standard of trust and confidence (Atkins 2017). This trust and confidence are what is protected by the specifically designed duties, imposing ethics not found in other commercial relationships (Atkins 2017). In fact, fiduciary law employs strong moral rhetoric that is uncharacteristic outside of criminal law (Velasco 2021). This very ground of fiduciary law creates a standard of trust which, though adjustable to each context, embeds some ethical rules to law to maintain that

standard. This ethical thrust and moral rhetoric is reflected also in the comprehensiveness of the equity remedies in fiduciary law which include monetary awards, injunctive and declaratory reliefs. Not only equitable compensation but also other restitutionary remedies such as accounting for profits, disorgement and constructive trusts, unwinding remedies such as equitable rescission and supervisory remedies such as declaratory judgements are embodied within the body of fiduciary remedies (Velasco 2021; Bray 2019).

Artificial intelligence (AI) and trust

AI in general

Artificial Intelligence (AI) is an umbrella term for a range of computer systems that demonstrate intelligence within the meaning of performing tasks commonly associated with human mind such as reasoning, generalising, problem solving or learning. While AI is used to describe (typically digital) artefacts that extend any of the capacities related to natural intelligence (Bryson 2018), this approach might not be sufficient to explain all the attributes of AI. Not only do existing applications of AI already show super-human performance with regard to specific tasks such as playing and winning at chess, but also, they do not have to function like the human mind, nor do they need to exhibit self-awareness and consciousness to perform tasks that would otherwise require intelligence when done by humans (Konig et al. 2022). While conceptually AI is difficult to define, the EU's Proposed AI Act makes a broad definition including a number of approaches and techniques.[8]

Among the techniques and approaches referred to in the EU's Proposed AI Act, the most prominent one is Machine Learning (ML) which mainly build on the artificial neural networks (ANNs). ANNs are computational models that use a simplified understanding of how the human brain learns through the use of essentially statistical models, involving several layers each consisting of multiple neurons, with neurons in each layer linked to the neurons in the previous and subsequent layers by synapses (Markou and Deakin 2021). Deep Learning (DL), a subset of ML, although vary depending on applications, generally involve large ANNs where 'depth' is determined by the number of hidden layers and neurons within them (Markou and Deakin 2021).

AI employ cognitive processes to solve the problems humans face, if not the same way, usually by detecting, observing and interacting with the environment and creating new correlations across myriad data. These computational correlations and solutions are realised by AI agents. At the heart of modern of AI lies the concept of *agents* which is a piece of software within a larger computer system performing a function on behalf of a user or another software agent and situated in an environment and interact with that environment while showing a certain degree of autonomy (Pasquale 2022).

The concept of *agent*, entailing all the AI techniques and approaches, is different from the legal concept of agent in the sense that the former does not necessarily act on behalf of a person ('beneficiary') and provide certain outputs based on the perceptions from and interactions with the environment, fulfilling a task of information gathering, filtering and/or correlating which result in automated decision making (ADM) so many times. Regardless of legal debate around agency,[9] it is important to underpin the distinction between *autonomy* and *control* to fully comprehend the reach of AI capabilities. Broadly speaking,

the more autonomy is given to the agents embedded in AI system, the less control is meant to be on the part of the human decision makers in such systems,[10] also being echoed with the artificial moral agents.

Against morally problematic situations which cannot be fully controlled by humans, whether or not artificial moral agents should be acknowledged in philosophy is an ongoing debate. Explicit moral agents, e.g. fully autonomous AI agents, are arguably situated somewhere in between moral subjects in the Kantian sense, who act from duty, and Kant's example of the prudent merchant whose self-interest only accidentally coincides with moral duty (Misselhorn 2022). It is also noteworthy that even if artificial moral agents do not fulfil the conditions for trustworthiness, trust may play a role with respect to their design and development, e.g. via developing codes of conduct, standards and certifications (Misselhorn 2022).

The environment AI agents interact with might be physical e.g. in object recognition systems such as autonomous vehicles (AVs) or virtual, e.g. AI-driven computer games, or alternatively they are software agents that are designed to interact with the physical world in a specific way, as do the chatbots that operate as text-based or voice-controlled conversational agents (Pasquale 2022). AVs manifest such capabilities of AI including systems of sensors and processing capacity that generate new complexities in the extract, transform and load process of their data systems (OECD 2019). AVs have light detection and ranging systems that can map out the environment, computer vision technologies that can track the eyes and focus of drivers and determine when they are distracted, and increasingly with a new capability of split-second operational decisions (OECD 2019), which all are governed by AI agents posing a significant degree of autonomy.[11]

While performance indicators in the AV industry would entail speed control, high-performing sensors, fuel efficiency, etc. they would differ when we mention socio-technical AI systems, which depend on not only technical hardware but also human behaviours and social institutions (Benk 2022). Given this, AI's computational capabilities need to respond to both the social fabric embedded in a socio-technical system as well as related moral and ethical concerns alongside the economic needs for a proper functioning.

Trust in (relation to) AI: from interpersonal to socio-technical perspective

Threading the needle from fiduciary law to socio-technical systems

Fiduciary law is steeped in moral issues such as trust, vulnerability and abuse of power (Velasco 2021). As implicated above, the governance of fiduciary relationships is built on interpersonal trust and how to remedy unexpected consequences of breaching this trust. The rationale behind fiduciary duties lies at the fact that the beneficiary (trustor) is vulnerable and needs protection from any abuse at the hands of fiduciary (trustee). Since the latter has the power to betray the former, the fiduciary obligations are imposed and equitable remedies are invoked to protect the latter. Crucially, the common law does not satisfy with contractual and tort obligations and remedies and creates a space for equity law to protect the vulnerable, mainly for the ethical and moral reasons which go to the roots of the trust as placed at the core of fiduciary law.

By making a connection between another area of law and fiduciary law, legal scholars may be able to borrow and incorporate its strong moral rhetoric without being encumbered by its strict technical duties (Velasco 2021). However, even this approach would

not be sufficient for the areas where trust is of paramount importance being affected by technologies such as AI. In the case of autonomous AI agents, how to govern trust becomes trickier, given the multiple actors, e.g. software developer, manufacturer, user; wide range of factors, e.g. the data on which the algorithm is trained; the techniques for modelling; design and architecture; as well as the impact the use of technology on people's decisions and lives. Not only the multiple agencies but also the dynamic relationships among the parties in the context of an AI-driven socio-technical system such as healthcare would require a broader perspective to govern trust in such a context.

For instance, if a fiduciary misdirects their principal's property, it is important for the purposes of deciding which remedy to impose that the Court determines the nature of the obligations that the fiduciary owed, for example, whether the property was the subject of legal or equitable obligations, or both (Ryan 2021). This standpoint takes trust on board as a per se relationship, e.g. fiduciary and property owner, and elaborates on certain obligations and potential remedies to be applicable in the case of betrayal of the latter. This rather 'restorative' approach would however fall insufficient for the transformative nature of trust in socio-technical systems which involve multiple actors including AI taking part as a sub-system. In such a context, the actors are often interconnected with each other carrying out the defined task that is influenced by technical and social elements and their interaction.

Given all this, trust in a socio-technical system needs to be entangled by considering the interactions and trust relationships among different actors including AI agents. A typical socio-technical system consists of subsystems that interact with each other being oriented to the same goal, e.g. delivery of healthcare services to the patients. By definition, in a socio-technical system, multiple stakeholders, e.g. engineers, business owners, and customers interact with the AI to achieve their goals according to their roles, through the use of system interfaces, over time, and following the rules of the system social institution (Benk 2022). Taking the healthcare as an example, the structural core of such a system comprises of people, e.g. providers, patients, patient family; performing various tasks, e.g. diagnosis, treatment; within a physical environment, e.g. cancer setting, home care; using tools and technologies, e.g. AI-enabled technologies, consumer health informatics tools; within an organisational context, e.g. guidelines to integrate AI results into decision-making (Choudhury and Asan 2020).

Socio-technical system theory and its application to AI-driven contexts

While the socio-technical system theory originally examines 'both the technical system and the social system and their interrelations on the work group level', it aims to explore learning and behaviours regarding the resources and interactions of the system that involves current technology, interpersonal interactions, language, and external environment (Yu, Xu, and Ashton 2023). The sociotechnical perspective acknowledges that a system's outcomes depend on mutual influences between technical and social structures, as well as between instrumental and humanist values (Dolata, Feuerriegel, and Schwabe 2022). From this point of view, socio-technical perspective can be built on interpersonal interactions and other interactions e.g. human-AI that are focused on functionality and technical efficiency of these systems.

The socio-technical approach implies that the technical and social subsystems of work cannot be decoupled and are inter-related; the compatibility and interaction between the

two sub-systems determine the effectiveness of a work system (Holdsworth and Zaghloul 2022). Having said that, interconnected elements of a socio-technical system, in which an AI agent subsists as a sub-system, cannot be considered in isolation from each other. Inter-connection of subsystems is a key aspect of an AI-driven socio-technical system where multiple actors interact and co-work enabling functionality of the whole system. Human-AI interactions in the system may comprise the periodic validation of batches of images, AI outcomes and the performance of the AI (data engineer, data scientist, clinicians, and the AI), the discussion of a given image and its AI outcome to determine diagnosis (patient, clinician, and the AI), or the treatment for the fracture (clinician, surgeon, and the AI) (Benk 2022).

The placement of trust in socio-technical system often requires a belief about its trustworthiness including AI agents. This notion seems to portray the EU's Ethics Guidelines for Trustworthy AI, which underlines that trust 'concerns not only the technology's inherent properties, but also the qualities of the socio-technical systems involving AI applications' (IHLEG 2019, 5). This standpoint brings us to a rather broad perspective to incorporate not only the AI agents but also the overall socio-technical system they are embedded, while constructing and maintaining trust. The EU Commission attributes a number of properties, i.e. lawfulness, ethics and robustness, to trustworthy AI, and regards 'trustworthiness' as a pre-requisite 'to ensure that we can trust the sociotechnical environments in which [AI systems] are embedded' (IHLEG 2019, 4). This approach prioritising 'trustworthiness' under governing mechanisms and principles can also be seen within the UK approach as well. For instance, the recently issued UK guidance refers to a list of 12 requirements in its governance framework by which to develop public trust as well as respond to other concerns (UK Department of Health and Social Care 2021). This suggests a rather rational approach, arguably along with a re-conceptualisation of trust focused on trustworthiness and the risk of harm on trusting activity (Lee 2022).

However, marking a distinction to cognitive trust, there are some emotional aspects that might have a role in building trust. In this regard, human-likeness is a noteworthy property of AI being echoed with the term 'anthropomorphism' which is found to increase trust as far as virtual or robotic AI are concerned (Lockey et al. 2021; Glikson and Woolley 2020). Empirical research so far broadly support the proposition that anthropomorphism increases in trust in AI, although some studies demonstrate it can also cause discomfort in the case of highly human like robots (Lockey et al. 2021; Glikson and Woolley 2020). This implies that trust in (relation to) AI also encompasses some properties of interpersonal trust. For instance, the factors of interdependency and vulnerability are also discernible in an AI-driven socio-technical system where human factors rely on AI. Yet, the purposes of efficiency and cost-saving usually prevail in multi-actor contexts such as healthcare (Holdsworth and Zaghloul 2022) which would make trust less affective and more predictive. Furthermore, trust towards a socio-technical system would have a normative and/or epistemological nature interrupting or intersecting with interpersonal trust. Overall, it is unrealistic and unreasonable to delve into each interpersonal or two-sided relationships including human-AI interactions in such a context.

In multi-agent systems, AI users e.g. clinicians place trust in the institution and working there to meet their normative commitments and to act in the best interests of trustors, e.g. patients. Despite the complexity and multi-dimensionality of decision-making, the user does not infer that one can, or does, trust to technologies that the institution is

using, no matter how advanced, autonomous, or intertwined they are within the business practices of the institution (Ryan 2020). By the same token, if users are sceptical towards institutionalised healthcare in general, they might also be sceptical towards robots used for that purpose in the same context (Aroyo et al. 2021).

Arguably, there exists a hierarchical relationship between the human and the machine, in which the final control and responsibility always remain with the human, suggesting machine agency is the capacity attributed to machines to evoke changes within a socio-technical system by autonomously carrying out a task to reach a certain goal (Zafari and Koeszegi 2018). This implicates that trust in such contexts needs to be considered holistically taking account of the AI autonomy alongside the original attributes of trust that can be developed within a socio-technical system in which AI is embedded.

This approach highlights the dynamic, hierarchical, and interactive subsystems and the need to enhance 'trust' in an AI-driven socio-technical system which should be revisited in view of the roles of the actors involved. While finding out the legal implications come to the fore in interpersonal trust, as manifested in fiduciary law, how to enable and sustain trust from a socio-technical perspective appears to be the leading concern from a holistic viewpoint. Having said that, trust in socio-technical systems need to be considered as a broader concept than trusting AI on an individual basis. Trust should thus be the central tenet of such a socio-technical system, encompassing the stages of AI life cycle, more explicitly during design, development and deployment of AI (Leslie 2021).

In such a socio-technical system, integrity and inter-connectedness of the sub-systems requires an overall approach incorporating designation of responsibilities of the medical staff, e.g. clinicians, doctors, determination of the extent to which the inner processes, life cycles of AI are to be kept transparent, and setting out the ethical and legal rules that govern the workflows, e.g. diagnosis, prognosis or treatment involving AI-assisted tools such as in detection of cancer. This arises as a necessity both to provide healthcare following the ethical principles, i.e. beneficence, nonmaleficence, autonomy and justice (Attfield 2023), and to successfully run the whole system including AI agents to enable trust and elicit its full value.

From this point of view, this study regards trust as worthy of value to be constructed, maintained and restored within and across the sub-systems of socio-technical system. In the field of healthcare, any institutional approach oriented to trust should thus enable 'trust' from the very beginning, e.g. training of AI, to the further steps of AI trajectory, e.g. day-to-day usage, along with ethical safeguards, e.g. concerning accountability for any damage resulting from AI reliance. Hence, cognitive processes embedded in an AI-driven healthcare system need to be run and monitored entailing the management of the risks related to transparency, reliability and fairness, considering all the potential dark holes of trust.

Building up a governance model to enable trust in an AI-driven socio-technical system

While AI is mostly used for the sake of technical efficiency and cost saving, the socio-technical character of AI-driven systems compels any consideration of trust to be made from a holistic viewpoint. Multi-agent relationships are a more combination of trust (interpersonal and institutional) and reliance (with AI and other technologies being used) (Ryan

2020). In this approach, AI functions as a sub-system which should serve to the overall trust across to a socio-technical system such as healthcare. This suggests 'trust' needs to be revitalised by building up a 'governance model' to facilitate identifying the key processes, tools, safeguards and remedies. In other words, trust can be constructed and sustained rather than being regarded as granted. Based on this approach, this article proposes a framework by taking trust as the main thread of a 'governance model' in which all the AI-inclusive processes, ethical and legal challenges can be worked out.

Construction, maintenance and restoration of trust would enable effective solutions to tackle opacity as well as bias and discrimination in AI-enabled processes. In fact, AI processes and systems can raise various concerns, which can relate to the opacity of underlying software, potential challenges over human agency and control, and inequalities being leveraged and/or duplicated (Sartori and Theodorou 2022). Not limiting to these, many of the ethical and legal challenges that can arise from reliance on AI can be preempted and mitigated by building a trust-centric structure. From this point of view, this study proposes a governance model built upon 6 elements, i.e. reliability, transparency, adaptiveness, competence, responsibility and accountability, which can cumulatively create the full value of trust, as shown in the below figure (Figure 1).

In this model 'trust' is revitalised with a view to demonstrate how it can be governed by leveraging the tools, safeguards and remedies to elicit the full value of trust from a socio-technical perspective. All of the elements that underlie trust in an AI-driven socio-technical system are related to each other, yet their value relates to their respective role and functionality within the overall structure. For instance, reliability and transparency of an AI-driven socio-technical system are key to the 'intrinsic' value to be derived from it. In close relation to this, competence and adaptiveness surface over time, e.g. when the AI users find out solutions via this system on the face of new problems, delivering a 'dynamic' value. Last but not least, 'ethical' value arises from the implementation of safeguards and remedies to restore trust in view of the ethical and legal challenges that might otherwise erode it, e.g. when an AI-based diagnosis has turned out to be biased for the training data.

Having said that, this model pays attention to construction, maintenance and restoration of trust taking it broader than the fiduciary law does, considering the latter provides remedies applicable under certain circumstances, e.g. a solicitor's giving advice to a client, company director's acting on behalf of a company, a trust's managing assets for settlors. In this approach, trust would be achieved from distillation of the human-human and AI-human interactions in a socio-technical system. Once this happens, in other words when the multiple actors interactively develop and maintain the overall trust, all of them including patients as well as clinicians would benefit from this along with the full value demonstrated above.

With respect to the value of the trust, a framework is drawn within the model by categorising each of the 6 elements under 3 parts: (i) elements of intrinsic value, (ii) elements of dynamic value and (iii) elements of ethical value. Under this tripartite framework, no structural link or sequence exists between these elements, since all the elements drive each other in carrying out the socio-technical tasks. These three inter-related chains of trust include the gears making the whole system work up and running eliciting the full value of trust out of that. From the systems viewpoint, the overall trust becomes more than the sum of all the constituent elements for the dynamic and distilled nature of

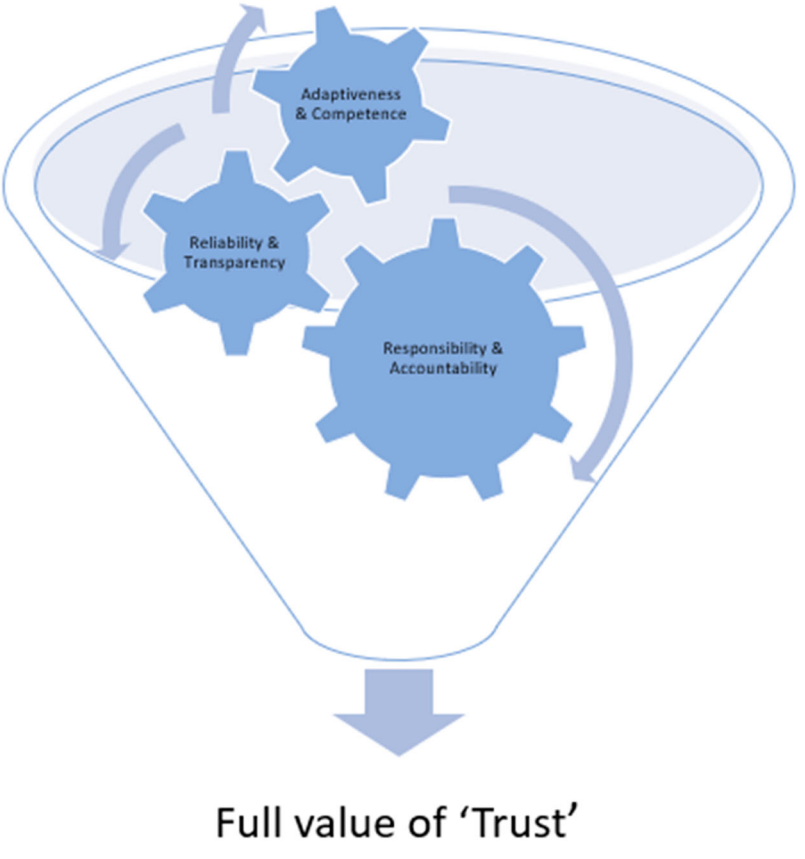

Figure 1. Full value of 'Trust'.

trust from a socio-technical perspective. On the other hand, the full value can decrease in case an element does not perform well towards the functioning of the whole system, e.g. because of a broken or damaged gear.

Framing governance model with a focus on healthcare

Elements of intrinsic value

In AI-driven sociotechnical systems such as healthcare, AI users e.g. clinicians need to be provided with information about the properties of AI, since this is often necessary to develop trust, from the beginning. In this category is included 'reliability' and 'transparency' of AI, which are considered to be the antecedents of 'trust' creating its intrinsic value. This value would then become augmented along with other features, e.g. 'competence' being added on to 'reliability' and 'transparency'. While 'competence' overlaps with as well as complements 'reliability', they are distinguished for the purpose of this study.

For instance, before using navigation app or a GPS map, we would prefer to find out some useful information and/or ask any experienced user or friend ('testimony') with regard to some key aspects such as what the fault ratios are or whether alternative routes are given for cyclists, pedestrians, drivers, etc. The former aspect is usually

echoed with 'reliability' meaning the same or similar outcome being reached on the part of all the users, whereas the latter ('competence') denotes capability of the AI tool or agent to offer diverse options and capabilities on top of the reliability.

Evaluating 'reliability' by comparing to higher performance or quality on the basis of 'competence' is usually distinctive because the former addresses the robustness of AI, i.e. concerning whether deviations from normal functioning occur as usually labelled as 'breakdowns' or 'malfunctioning' (Starke et al. 2022). Technical requirements and standards are entailed within the meaning of technical robustness to ensure the AI agent be not open to malicious use and not cause any unintentional harm (IHLEG 2019).

Along with 'reliability', 'transparency' needs to be noted as a key aspect and antecedent of trust in relation to AI. Transparency indicates the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment along with the provenance and dynamics of the data that is used and created by the system (Virginia Dignum 2020).

Reliability

In the case of AI, reliability is often difficult to assess, especially in the context of high machine intelligence, as learning from data can lead technology to exhibit different behaviours, even if the underlying objective function remains the same (Glikson and Wooley 2020). In research conducted with robots, working with inconsistent (medium level) reliability is found to be more confusing to the participants with their trust being lower than in the consistent low level reliability condition (Glikson and Wooley 2020). It is found low reliability significantly decreases trust, not being limited to the robots and entailing other types of AI, e.g. virtual and embedded (Glikson and Wooley 2020).

Reliability means safety and robustness also need to be found in AI. Unreliable performance early in one's experience with an AI system may cause more significant trust breakdown than failure later in an interaction (Lockey et al. 2021) and safe and technically robust AI systems would have a key role in building trust from the outset. There are various ways and procedures to ensure this such as testing and approval processes run by the US Food and Drug Administration (FDA). Likewise, under the EU's Proposed AI Act, an assessment for high-risk AI systems would require a full, effective and properly documented ex ante compliance with all the regulatory requirements, compliance with quality and risk management systems and post-market monitoring (Council of the EU 2022, Art. 17–19). However, non-deterministic AI systems e.g. ML would require more than just the application of quality assurance protocols designed for conventional software systems, incorporating built-in test techniques (Middleton 2022). All these processes and techniques facilitate robustness and cybersecurity of the systems, leveraging reliable AI and enabling the intrinsic value of trust.

Reliability of AI is a key aspect to build trust in (relation to) AI through which inaccurate outcomes including bias and discrimination can be mitigated. Reliability would co-relate with accuracy or validity, and for reliable AI these latter aspects might be needed (Grimm, Grossman, and Cormack 2021). Unfair outcomes including bias and discrimination, whether directly or indirectly, might stem from the quality of the training data or algorithmic design or programming (Grimm, Grossman, and Cormack 2021; Criado and Such 2019; Yeung 2019). For instance, in AI models, configuration of parameters includes classification and correlation of specific attributes, and this process might result in

certain features, e.g. being white, male, with no disability, being weighted more than others, e.g. black, female, disabled (Scantamburlo, Charlesworth, and Cristianini 2019; Criado and Such 2019; Mittelstadt 2016). This can cause other problems, for example, when a treatment actually works better for one gender or race than another, but the beneficial effect is masked by an overall i.e. combined accuracy rate that is low, or because the protected data is either not collected or not considered by the algorithm (Grimm, Grossman, and Cormack 2021). This situation would adversely affect reliability of AI. Such examples demonstrate that accuracy would become at stake yet would not be sufficient alone for building reliable and trustworthy AI. To deal with the problems of bias and discrimination, AI systems should therefore be transparent as well as reliable within the meaning of functioning within a reasonable margin of accuracy.

Transparency

ML-driven systems are often opaque in the sense that the patients affected by them can hardly ever comprehend how or why a certain input of data has been categorised and produced a certain output. This is echoed with the 'opacity' or 'black box' problem. For instance, medical research indicate that ML and DP techniques could be used to diagnose Alzheimer disease years before symptoms appear, by complementing the work of radiologists with other biochemical and imaging tests (Diogo, Ferreira, and Prata 2022; Brierley 2022; Fabrizio et al. 2021; Radiological Society of North America 2018). However, clinical usefulness, interpretability, and generalisability of the classifiers across datasets and MRI protocols remain limited, mainly for the black-box problem (Diogo, Ferreira, and Prata 2022). Likewise, in detection of breast cancer (Salvi and Kadam 2021), pancreatic cancer (Liu 2020), head and neck cancer (Ma et al. 2019), DP techniques e.g. ANNs can pose challenges to transparency, since how conclusions are drawn is mostly not visible to the clinicians, even to their programmers, let alone the patients. Ironically, black-box medicine might promise substantial benefits regarding diagnostics, personalised treatment, image analysis (Ford, Price, and Nicholson 2016) which usually has a cost in terms of explainability.

In the US, FDA has approved several black-box medical devices with 'locked' algorithms that generate the same result each time for the same input (FDA 2019).[12] This means exchanging transparency with predictability and promising accuracy rates. Until certain satisfactory rates are achieved, 'human oversight' would still be needed as the EU's Proposed AI Act suggests. This Regulation Proposal requires high-risk AI systems to be designed and developed in a manner that 'they can be effectively overseen by natural persons' (Council of the EU 2022, Art 14) incorporating appropriate human-machine interfaces (Schwemer, Tomada, and Pasini 2021). In case there is no or minimum risk, human oversight would arguably be reduced or diminished to embody a 'no-human-in-the-loop' approach.[13] This approach looms on the horizon along with the calls for the 'post-market surveillance' to mitigate the adverse consequences with monitoring (Samir 2022). Effectively, the barricade for entry of novel advanced algorithms has been lowered recently (Asan, Bayrak, and Choudhury 2020) and this trend seems to continue with differentiated policies of monitoring.

Overall, it may be more important to explain how a system has been validated and whether a particular use falls within the parameters with which the system can be expected to produce reliable results rather than explaining how an AI model arrives at

a particular judgement (WHO 2021). Clinicians require other types of information, even if they do not understand exactly how an algorithm functions, including the data on which it was trained, how and who built the AI model and the variables underlying the AI model (WHO 2021). Insofar as this is assured, the need for full transparency would be reduced for the complexity embedded in AI-driven healthcare system. From this point of view, the FDA's recently published 'Patient-Centred Approach' (FDA 2021) signifies an optimum way of dealing with the need for transparency vis-à-vis accuracy and predictability, given the explicit requirement that manufacturers ensure transparency towards users e.g. clinicians about the functioning of SaMD devices to the effect that they understand the benefits, risks, and limitations of these devices (Unver and Asan 2022).

Elements of dynamic value

We follow the suggestions of the AI agents which we trust based on their key features, e.g. reliability, transparency. While they exhibit a certain level of autonomy when computing results, we believe they provide a service in our best interest based on their capabilities. This trust belief reinforces our reliance on such agents, even sometimes we do not recognise this, such as while relying on a navigator that guides us while driving on the road. Another example includes in-car collision avoidance system which can sense danger and react faster than humans. We tend to feel more vulnerable in traffic than before and increasingly rely on these technologies with everyday enhancing capabilities also in view of their decreasing and affordable prices. Our trust in AI systems, among other factors, often depends on our experience as well as the third parties' belief or testimony (Jacovi 2021). Given the fact that throughout such learning process AI would respond better as our needs (re)surface and the capabilities of ML and DL enhance over time,[14] the value to be derived from this can be best described with the term 'dynamic'.

This situation makes 'competence' quite important in the sense that this experience itself is key to arriving to a trust decision, usually with new information and achievements, echoed with 'adaptiveness' in this study. We can either 'establish' or 'develop' a trust relationship based on our reliance. This would be a new experimental process enabling to assess the so-called competence and adaptiveness of AI, e.g. by testing its hidden or extra features. In case AI is not configured robust and reliable enough, such new features would not come up. Nor do they emerge unless data quality is accomplished. In the case of socio-technical system, reliance can equally make the trajectory of AI-related trust dynamic, although meeting the antecedents of trust (reliability and transparency) would lessen the potential risks.

In the healthcare sector, regardless of the degree of autonomy of the AI agents, clinicians are responsible for the diagnosis and treatment processes and are held liable for their erroneous decisions causing any damage for relying on AI (WHO 2021; Gerke, Minssen, and Cohen 2020). Even in situations when AI substitutes clinicians, e.g. such as in SEDASYS machines monitoring patients' breathing and heart rates and (re)setting does of anaesthesia accordingly (Pasquale 2022) the ultimate decision makers are the physicians in charge. Clinicians are usually trained to use AI for various purposes such as for gathering X-ray images, recommendations, etc. where necessary by obtaining informed consent from the patients.[15] All the key responsibilities being on the clinicians makes healthcare rather distinctive when compared to other sectors, e.g. AV industry

where manufacturers rather than operators are at the forefront.[16] In this regard, autonomy of the AI devices and software, embracing their 'competence' and 'adaptiveness', stands as a key factor in building of trust in socio-technical systems.

Overall, antecedents of trust, e.g. reliability and transparency, are complemented by AI having multi-purpose character and capabilities (competence) and being responsive to changing environmental conditions (adaptiveness). That is not to say, without these additional properties, trust will fall behind. However, in a dynamic and steadily changing environment, AI's having such complementary features would reinforce a trust decision and pre-empt any likelihood of deterioration in terms of trust relationship.

Adaptiveness

Adaptiveness is meant to cover AI algorithms that update themselves over time with new data. 'Adaptiveness' means self-updating with and learning through a continuous stream of incoming data. Adaptive algorithms enable adapted strategies and decisions according to the circumstantial changes based on the data gathered and pooled from great many sources. Vehicles can take advantage of the experience of other vehicles on the road, without human involvement, and the entire corpus of their achieved 'experience' is immediately and fully transferable to other similarly configured vehicles (West and Allen 2018). Adaptive AI solutions are getting widespread in various industries, e.g. finance, telecommunications, insurance, aiming to adjust the service delivery to the changes in the market.

Faster adjustment with wealthier data would result in more responsive and effective AI solutions and tools. Hardware or software level innovations can enable this, sometimes by means of the Internet of Things devices and chips integrated into the environment. For instance, in healthcare, Body Sensor Networks (BSNs) are becoming more prominent given their miniature size and capability to enable wireless data transfer. Major technical hurdles are related to continuous sensing and monitoring, requiring long-term stability of the sensors and low-power operation, bio-inspired design requirements and battery lifetime (Imperial College London 2022). 'Adaptiveness' of BSNs is key to cope with these hurdles. Along with real-time datasets, adaptive transmission data rate mechanism and self-adaptive routing algorithm would enable pervasive and cost-effective healthcare through wearable devices (Zahid 2022). Not being limited to BSNs, an adaptive ML algorithm changes its behaviour using a definitive learning process without requiring any manual input and might generate different outputs each time a given set of inputs is received due to learning and updating (Asan, Bayrak, and Choudhury 2020) with increasingly higher rates of success (Abdel-Jaber et al. 2022; Davenport and Kalakota 2019). Under this light, adaptiveness would increase not only efficiency but also improve the level of trust in AI-driven socio-technical systems.

From a technical point of view, adaptiveness is a key driver of competence for which computational power, processing latency and wireless data transfer need to be managed effectively. In this regard, success of AI needs to be considered in combination with other elements of IT and up-to-date datasets to derive enhanced competence. Hence, the quality of data used for AI algorithms is key for increased success rates to build trust. Not limited to this, all the related factors relevant to adaptiveness, e.g. IT resources, enhance 'competence' of AI and ultimately serve to building of trust in AI-driven socio-technical systems.

Competence

'Competence' is of a distinct nature although closely related to 'adaptiveness'. Competence is evaluated with the further capabilities than the baseline features of steady functionality towards a particular task. Normally, AI performs a task detecting its environment and working out the given problem via the algorithms trained based on datasets. If trained on a static pre-existing dataset, that situation signifies a model running on fixed parameters. To update the model based on new data or changing circumstances, we have to retrain it offline with the updated dataset (generally a computationally- and time-intensive process) and then redeploy it (Toews 2022). Competent AI, incorporating new variables and/or modifications over the model, would better cope with the real-world environments providing solutions against diverse circumstances.

Language learning models (LLMs) is an important AI/ML tool based on Bayesian programme learning which increasingly enables a stream-based learning process to be implemented across new data and languages (Toews 2022). Enhanced LLMs can dynamically incorporate far more inputs basis from their environment and decipher wide-ranging language structures with more accurate translations over time (Toews 2022; Zewe 2022). In their quest to develop an AI system that could automatically learn a model from multiple related datasets, the researchers chose to explore the interaction of phonology (the study of sound patterns) and morphology (the study of word structure) and achieved successful results of high-level language patterns for 58 languages (Zewe 2022). Notably, the 'competence' of AI here is related to not only multiple datasets being employed but also cross-learning across different languages of the AI system. This probabilistic method of ML has similar manifestations of success in the field of healthcare such as in prediction of diseases. Indicating signs of diabetes, oncology, liver and kidney diseases, such ML tools have so far been used in a variety of classification tasks, rather than a discreet function.

Competence in this regard embodies enhanced capabilities of AI to deal with different (iated) cases via models that update themselves based on new variables as well as wide-ranging datasets. This poses new affordances creating dynamic value, e.g. accomplishment of similar performance across different fields, on top of reliability features. This can be seen in case AI is proven to have reliability features in a field, e.g. breast cancer, and its capabilities is tested in another field, e.g. lung cancer. For instance, Watson for Oncology (WFO), an AI assistant decision system, was developed by IBM with the help of top oncologists from Memorial Sloan Kettering Cancer Center (Jie, Zhiying, and Li 2021). It took more than 4 years of training, based on national comprehensive cancer network cancer treatment guidelines and more than 100 years of clinical cancer treatment experience in the US, and can recommend appropriate chemotherapy regimens for specific cancer patients (Jie, Zhiying, and Li 2021). When compared to multidisciplinary teams of clinicians, WFO is recorded to have displayed a high rate of concordance overall (over 80%) across different regions and types of cancers, including breast cancer, rectal cancer, colon cancer, gastric cancer, lung cancer, ovarian cancer and cervical cancer (Jie, Zhiying, and Li 2021). While this clearly demonstrates 'competence' considering the concordance rates across a huge variety of regions and cancer types, disparate rates of concordance among the cancer types (breast cancer was the highest with 81.76% and gastric cancer was the lowest with 29.90%) and stages (stage I–III versus stage IV, the latter being less concordant) (Jie, Zhiying, and Li 2021) can lead to an

interpretation that the reliability of WFO (AI) is different across distinct cancer types and stages. In other words, one can argue WFO is more reliable in terms of giving recommendations as to the detection of breast cancer when compared to gastric cancer.

Elements of ethical value

Enablement of trust in an AI-driven socio-technical system, as lying at the centre of this proposed model, means not only construction and maintenance but also restoration of it. Safeguards and remedies to restore trust therefore stand as key factors that would impact the overall trust decision towards such systems. This can be compared to a situation when we would like to purchase an AV based on our knowledge and past experiences as to their reliability, transparency, performance, etc. In such a scenario, we usually require extra information as to the consequences in case of an accident, including to what extent the insurer or the manufacturer will meet the costs for damages caused by an erroneous AI system. Following on this, several questions remain to be answered as follows:

(ii) Who is responsible for taking the necessary safeguards to prevent any damage or harm that would result from erroneous AI system and to ensure the AV is steadily working in compliance with the standards and technical requirements?
(iii) Whether or to what extent explanation will be given for any AI-related damage or harm, e.g. for the faulty auto-pilot system, including the justifiable measures to pre-empt it with the causes of the problem?
(iiii) Who will be liable for the monetary and non-monetary damages or losses in case a bodily injury or death takes place?

While the first of the above questions is related to 'responsibility' of the related actors, the second question is related to 'accountability' and the third 'liability'. Although all these concepts are related to each other, 'liability' is not necessarily related to ethics but to the legal obligations to pay for damages or losses caused by one's actions, e.g. under civil or criminal law. Therefore, the below categorisation encompasses 'responsibility' and 'accountability' as the elements of trust having 'ethical' value that needs to be preserved via certain safeguards and remedies, as explained below.

Responsibility

Responsibility is an issue of governance by which to set out the roles and duties of each actor involved in a regulated setting. Regulation in this context needs to be considered as broad as to entail all types of regulation, including self and co-regulation, and by all actors and means. It is up to governments and citizens, namely the society to determine the extent to which AI and its actions are to be regulated including the degree to which responsibilities are detailed for each participant or agent. A responsible approach to AI is needed to ensure that systems are not only developed in a good way but also developed for a good cause (Dignum 2020). Responsible AI concerns not only the software system itself, but also, and foremost, the people, institutions and organisations that compose the socio-technical system (Dignum 2020).

In a socio-technical environment where AI can potentially affect other participants with respect to their rights or benefits, it should be ascertained which agent(s) or stakeholder (s) are to be held responsible and to what extent. For instance, an AV owner should know who is responsible for prevention of any collusion as well as overseeing and maintenance of the car including for all hardware and software incorporating AI agents that function autonomously. As the chain of responsibility grows, means are needed to link the AI systems' decisions to their input data and to the actions of stakeholders involved in the systems' decisions (Dignum 2020).

Responsibility responds to the need for mechanisms that enable AI systems to function ethically and legally. And at this point, responsibility serves to building 'trust', enabling humans to participate in and figure out the best ways to govern an AI-driven automated process. It is widely acknowledged that there should be human oversight for the involvement of AI in the regulated processes (IHLEG 2019; Middleton 2022). This is sometimes echoed by the term 'human in the loop' implicating the role of the human actors to lead, design and/or validate the processes where AI is embedded.[17] Although the degree to which human should be in the loop is debated (Grimm, Grossman, and Cormack 2021),[18] human oversight appears to be key to build trust not only during the design and validation processes but also during use and reliance on AI.

For instance, Clinical Decision Support Systems (CDSSs),[19] which are designed to assist clinicians with their decision-making based on prior successful diagnoses, treatment, and prognostication (Lysaght, Lim, and Xafis 2019) embody the mechanisms of human oversight – technically signifying human-in-command approach in which humans have the control and command over the AI recommendations. In this example, based on the published data on its performance characteristics including assigned weights to the data, a clinician can determine whether s/he should override the recommendations given by the CDSS or not (Lysaght, Lim, and Xafis 2019). The CDSS can predict outcomes based on various social determinants of health, which can arguably lead to further discrimination of already-marginalised groups and communities (Lysaght, Lim, and Xafis 2019) and this fact requires the clinicians be knowledgeable about the capabilities and limits of such systems. On this note, CDSS-made decisions would complicate the responsibilities regarding how far information acquired by these AI systems may be used to build patient profiles by private health insurers as well as public health systems. From this point of view, socio-technical systems in which AI is deployed and used needs to be regulated particularly for the bias and discrimination risks, based on well-balanced and formulated responsibilities.

Accountability

Accountability refers to the requirement for the system to be able to explain and justify its decisions to users and other relevant actors (Dignum 2020). Rationale and default outcome based on AI can diverge from the well-established values, principles as well as rationale that used to drive decisions or assessments made by humans, resulting in accountability gap (Widlak, Van Eck, and Peeters 2020). Accountability gap can also emerge when the physicians follow AI recommendations in view of their success rate for another community or country that has different cultural or social norms. Besides, such a gap would be sourced from the 'many hands problem' or the 'traceability of

harm' which bedevils health-care decision-making systems even in the absence of AI (WHO 2021, 43). To ensure accountability, decisions should be derivable from, and explained by, the decision-making mechanisms used (Dignum 2020). It also requires that the moral values and societal norms that inform the purpose of the system as well as their operational interpretations have been elicited in an open way involving the stakeholders (Dignum 2020).

Accountability enables identification of the causes for decisions and outcomes, indicating responsible actors to provide further explanation regarding the life cycles of AI, e.g. the diagnosis and treatment processes for the healthcare sector. From a broader point of view, it would mean risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties in an AI-driven healthcare (Rodrigues 2020). In managing the related risks and potential harms, a distinction can be made between accountability for content and for operation (Academy of Royal Medical Colleges 2019). According to such distinction, a clinician might be accountable for not using an algorithm or device correctly, but in the event of harm being caused by incorrect content rather than improper use, the accountability must lie with those who designed and then quality assured it (WHO 2021).

Clinicians should not, however, be fully exempt from accountability for errors in content, not checking out whether an automated technology meets their needs or those of the patient (WHO 2021) given the risk of automation bias (Grimm, Grossman, and Cormack 2021). At this point, accountability and liability intersect with each other, particularly where the responsible actors need to follow standard of care.[20] While physicians must be able to trust an AI algorithm, they should not ignore their own expertise and judgement and simply rubber-stamp the recommendation of a machine for the possibility of automation bias (WHO 2021; Blasimme and Vayena 2020).

This broader approach means an expansive vision through which responsibilities of the agents can be reinforced. Given the fact AI is ever fast increasing along with likelihood of unpredictable consequences, such a broader approach seems compelling also from the perspective of trust. Hence, rather than relying on a narrowly formulated right to explanation such as under the GDPR,[21] broadly formulated accountability rules are advisable in an era where AI increasingly inhabits in modern life and society. Ethical value of trust would then be compounded with the intrinsic and dynamic values that emerge from the initial decision and relationship of trust.

Conclusion

We increasingly rely on the automated services and processes run by AI. Trust has always been a debated topic not only given the transformation led by AI, but also for its autonomous nature and ensuing legal and ethical challenges. Given such challenges, 'trustworthiness' has come to the fore, being connoted with 'trustworthy AI' (IHLEG 2019), yet governance of trust should not be boiled down to the ethics and standards in relation to AI. Its subjective nature would similarly mislead any debate in the sense that interpersonal interactions cannot be indicative of 'trust' without elaboration of the socio-technical systems AI is embedded. Considering that socio-technical systems involve many human and non-human factors that engage with AI, assessment of 'trust' in such a context requires a holistic viewpoint to look into this multifaceted and elusive concept.

Notwithstanding, interpersonal relationships that constitutes the core of fiduciary law offers a useful vision through which trust can be analysed with an expansive outlook in relation to AI. Following this approach, this study has reviewed various aspects and accounts of trust with a view to first flesh out the common law perspective, i.e. based on the fiduciary duties and remedies, and then build up a governance model with a focus on AI-driven socio-technical systems, based on 6 components, i.e. reliability, transparency, adaptiveness, competence, responsibility and accountability. Although inspired by the fiduciary law, this governance model upholds an expansive vision emphasising the construction, maintenance and restoration of 'trust' along the life cycle of AI, to underlie a trust-centric structure in an AI-driven socio-technical system.

According to the proposed governance model, 'trust' is taken as the main thread of a socio-technical system for which various tools, safeguards and remedies need to be in place. While the fiduciary law acknowledges the trust as granted and invokes remedies to restore it, e.g. in the case of betrayal of the trustor, this study does not consider 'trust' as *per se* or a granted feature of two or multi-sided relationship. Nor does it attempt to make an unequivocal definition or boundaries of 'trust' while acknowledging it is situational and context-specific. Rather, boundaries are set for the governance of trust in view of the interaction of human and non-human factors that can unlock the full value of trust to be derived from 'reliability and transparency', 'adaptiveness and competence' and 'responsibility and accountability'. Interactive subsystems should aim to distil intrinsic, dynamic and ethical values from these trust chains in an AI-driven socio-technical system.

After all, it is concluded that the proposed tri-partite governance model can bring about the so-called full value of trust in an AI-driven socio-technical system in the presence of the constitutive elements working in coherence. It is recognised that the findings of the literature as to trust, its trajectory and implications can be implemented within this framework. As demonstrated in the healthcare domain, the concept of 'trust' is primarily predicated on the components of 'reliability' and 'transparency' which provide its *intrinsic* value. On top of these antecedents, 'trust' can develop with a *dynamic* value to be derived from 'adaptiveness' and 'competence'. Finally, 'responsibility' and 'accountability' of the actors can bring about *ethical* value within an AI-driven socio-technical system.

Against this background, it is plausible that without one of these components, trust would not be implicated fully in an AI-driven socio-technical environment. As can be seen from the review of healthcare, a *holistic* approach incorporating all the elements of trust is crucial for a sustainable model and its implementation. Overall, it is considered that, this governance model, by upholding a holistic viewpoint, provides a generalisable framework that can enable trust in AI-driven socio-technical systems.

Notes

1. Although trust and confidence are key to fiduciary relationship, the relationship must also require the exercise of judgment and the making of discretionary decisions by A on behalf of B; alternatively, the giving of advice by A to B where A has a substantial degree of power over B's decision-making, and constitute trust and confidence in A's loyalty such

that A will put aside any personal interest and act solely in the interests of B (*Al Nehayan v Kent, Leggatt LJ, paras. 159 and 165)*.

2. According to the landmark decision of *Bristol & West Building Society v Mothew*, 'a fiduciary is someone who has undertaken to act for or on behalf of another in a particular matter in circumstances which give rise to a relationship of trust and confidence. The distinguishing obligation of a fiduciary is the obligation of loyalty. The principal is entitled to the single-minded loyalty of his fiduciary' (*Bristol & West Building Society v Mothew (t/a Stapley & Co) [1998] Ch 1)*.
3. Practical Law Corporate (2022) 'Fiduciary duties' (Practical Law UK Practice Note 8-107-4883).
4. *Bristol & West Building Society v Mothew* judgement sets out as follows:

   A fiduciary is someone who has undertaken to act for or on behalf of another in a particular matter in circumstances which give rise to a relationship of trust and confidence. The distinguishing obligation of a fiduciary is the obligation of loyalty. The principal is entitled to the single-minded loyalty of his fiduciary. (*Bristol & West Building Society v Mothew (t/a Stapley & Co) [1998] Ch 1)*.
5. Regarding the fiduciary duty to subordinate any personal interests to those of the other person, see *Children's Investment Fund Foundation (UK) v Attorney General [2020] UKSC 33*; *Tulip Trading Ltd v Bitcoin Association for BSV [2022] EWHC 667 (Ch); Kyla Shipping Co Ltd v Freight Trading Ltd [2022] EWHC 1625 (Comm)* (Practical Law Corporate 2022).
6. Regarding the 'no conflict of interest' rule, see *Aberdeen Railway v Blaikie Bros (1854) 1 Macq 461; Boardman v Phipps [1967] 2 A.C. 46; Bristol & West Building Society v Mothew* (Practical Law Corporate 2022).
7. The recently adopted Financial Services Act (FCA) 2021 sets out a framework under which banks, wealth managers, financial advisors or any other person or body that manages money for others are subject to prudential regulatory standards that are reminiscent of fiduciary duties and obligations. The new rules under FCA require firms to provide consumers with information they can understand, offer products and service that are fit for purpose and provide helpful customer service (Financial Conduct Authority 2021).
8. According to the EU's Proposed Regulation (AI Act), 'AI system' means 'a system that is designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of objectives using machine learning and/or logic and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts' (Council of the EU 2022, Art 3(1)).
9. For a comprehensive discussion around the concept of 'agency', see Aksoy (2022).
10. Regarding a more detailed explanation on the concepts of 'autonomy' and 'control' in the context of AI, see Wheeler (2020) and Powers and Ganascia (2020). Regardless of this discussion, an autonomous AI agent does not have itself have a concept of 'its own behalf', meaning they take the orders from the AI users and fulfil them, while still having and operating some sense of autonomy.
11. While fully autonomous AI agents can hardly be mentioned for AVs, the more agents involved in the decision making process, the more control assumed by them, as can be seen through the six-stage standard AV types developed by the Society of Automotive Engineers (SAE): (i) Level 0 (no driving automation), (ii) Level 1 (driver assistance), (iii) Level 2 (partial driving automation), (iv) Level 3 (conditional driving automation), (v) Level 4 (high driving automation) and (vi) Level 5 (full driving automation) (SAE International 2021).
12. The US Food and Drug Administration (FDA) categorises Software into three classes: (a) Software as a Medical Device (SaMD), (b) software in a medical device, and (c) software used in the manufacture or maintenance of a medical device. FDA defines SaMD as ' … AI/ML-based Software, when intended to treat, diagnose, cure, mitigate, or prevent disease or other conditions, are medical devices under the FD&C Act and called Software as a Medical Device' (FDA 2019). SaMD ranges from smartphone applications to view radiologic images for diagnostic purposes to Computer-Aided Detection software to post-processing of images to detect breast cancer (FDA 2017).

13. No human-in-the-loop approach contrasts with human-in-the-loop approach which means a control mechanism by human is incorporated into ADM process run by AI such as in humans marking false positives in the email spam filters. See also infra notes 17 and 18.
14. For instance, when an AI user (human) sets the destination for the AI agent (driverless car), the latter determines the way to reach that destination in an autonomous way, while still obeying the former's order. In doing so, the autonomy the latter has entails the algorithms pre-programmed that usually develop over time based on new data, ending up new ways of ADM (Wheeler 2020).
15. For processing of patents' medical data, 'informed consent' needs to be obtained from them in line with the data protection laws, e.g. Articles 6 and 9 of the EU General Data Protection Regulation (GDPR). Notwithstanding, from the perspective of liability regime there is presumption that there would not arise liability for failing to inform patients about the use of medical AI to help formulate treatment recommendations (Cohen 2020).
16. Notwithstanding, Pasquale argues liability should lie at the vendors because AI and robotics increasingly replace a skilled medical professional shifting the burden (Pasquale 2022).
17. Human-in-the-loop refers to the capability for human intervention in every decision cycle of the system (IHLEG 2019).
18. As the White Paper of the European Commission (2020, 21) establishes, the appropriate type and degree of human oversight may vary from one case to another. According to White Paper, 'it shall depend in particular on the intended use of the systems and the effects that the use could have for affected citizens and legal entities'.
19. CDSS, being programmed with rule-based systems, fuzzy logic, artificial neural networks, Bayesian networks, as well as ML techniques, works by continuously monitoring information that clinicians enter into the EHR (Lysaght, Lim, and Xafis 2019). As information is recorded, the CDSS can analyse the entries in real time along with other clinically relevant data including test results from pathology laboratories, radiological departments, genetics departments, and ambulatory settings, as well as research results stored in biobanks, clinical trials, and databanks of genome sequences (Lysaght, Lim, and Xafis 2019).
20. For example, in the UK law, a doctor will not be liable in negligence if s/he adopts a treatment accepted at the time as proper by a responsible body of medical opinion, even if other medical professionals would disagree (Turner 2020). Otherwise, if a doctor follows an ML model that results in erroneous prediction or treatment plan, application of it would create liability for the doctor who decides to do so. Only if the balancing condition (between the respective costs and benefits of accuracy and explainability) is met, the use of the model should be deemed generally legitimate (but not yet obligatory) (Hacker et al. 2020). Under German law, for example, new medical methods meet the standard of care if the marginal advantages vis-à-vis conventional methods outweigh the disadvantages for an individual patient (Hacker et al. 2020).
21. Article 5(2) of the GDPR imposes an accountability obligation on the data controllers yet does not detail this obligation except stating the controller be 'responsible for, and be able to demonstrate compliance with' the governing principles under Article 5(1). On the other hand, there exist information and transparency obligations under Articles 12–14 and 22 of the GDPR serving to keep the controllers accountable to some extent, e.g. requiring 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject' in the case of automated decision making process. Watcher, Mittelstadt and Floridi (2017) explores the potential accountability gaps within the boundaries of the GDPR with a focus on the 'right to explanation' and refers to the lack of precise language as well as explicit and well-defined rights and safeguards against ADM under the GDPR. Likewise, Edwards and Veale (2017) explain the shortcomings of this right to ensure transparency and accountability and argue this right would even lay a ground for a new kind of transparency fallacy.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Abdel-Jaber, H., D. Devassy, A. Al Salam, L. Hidaytallah, and M. El-Amir. 2022. "A Review of Deep Learning Algorithms and Their Applications in Healthcare." *Algorithms* 15 (71). doi:10.3390/a15020071.

Academy of Royal Medical Colleges. 2019. *Artificial Intelligence in Healthcare* https://www.aomrc.org.uk/reports-guidance/artificial-intelligence-in-healthcare/.

Aksoy, P. C. 2022. "AI as Agents." In *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*, edited by L. A. DiMatteo, C. Poncibò, and M. Cannarsa, 146–160. Cambridge: CUP.

Aroyo, A. M., J. De Bruyne, O. Dheu, E. Fosch-Villaronga, A. Gudkov, H. Hoch, S. Jones, et al. 2021. "Overtrusting Robots: Setting a Research Agenda to Mitigate Overtrust in Automation." *Paladyn, Journal of Behavioral Robotics*. doi:10.1515/pjbr-2021-0029.

Asan, O., A. E. Bayrak, and A. Choudhury. 2020. "Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians." *Journal of Medical Internet Research* 22 (6), doi:10.2196/15154.

Atkins, M. 2017. "What is the Purpose of the Ongoing use of Fiduciary Duties in English Business law, with Particular Reference to Breaches of Duty in Relation to Bribery, Secret Profits, Conflicts of Interest and Unconscionability?" PhD Diss. Lancaster University.

Attfield, R. 2023. *Applied Ethics: An Introduction*. Cambridge: Polity.

Benk M., S. Tolmeijer, F. Von Wangenheim, and A. Ferrario. 2022. "The Value of Measuring Trust in AI – A Socio-Technical System Perspective" Workshop on Trust and Reliance in AI-Human Teams (TRAIT), New Orleans, LA.

Blasimme, A., and A. E. Vayena. 2020. "The Ethics of AI in Biomedical Research, Patient Care, and Public Health." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 703–718. New York: OUP.

Bray, S. L. 2019. "Fiduciary Remedies." In *The Oxford Handbook of Fiduciary Law*, edited by Evan J. Criddle, 449–467. New York: OUP.

Brierley, C. 2022. "AI Could Detect Dementia Years Before Symptoms Appear" University of Cambridge. Accessed August 12, 2021. https://www.cam.ac.uk/stories/AIdementia.

Bristol & West Building Society. v Mothew (t/a Stapley & Co) [1998] Ch 1, [1996] EWCA Civ 533.

Bryson, J. J. 2018. "The Past Decade and Future of AI's Impact on Society." In *In Towards a New Enlightenment? A Transcendent Decade*. Madrid: BBVA.

Carter, J. A. 2020. "Trust and its Significance in Social Epistemology." In *Oxford Handbook of Social Epistemology*, edited by J. Lackey and A. McGlynn. Oxford: OUP. https://www.jadamcarter.com/research.

Choudhury, A., and O. Asan. 2020. "Role of Artificial Intelligence in Patient Safety Outcomes: Systemic Literature Review." *JMIR Medical Informatics* 8 (7): 1–24. doi:10.2196/18599.

Cohen, I. G. 2020. "Informed Consent and Medical Artificial Intelligence: What to Tell the Patient?" *SSRN Electronic Journal* 108: 1425–1469. doi:10.2139/ssrn.3529576.

Council of the European Union. 2022. Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (AI Act) and amending certain Union legislative acts, 25 November 2022.

Criado, N., and J. M. Such. 2019. "Digital Discrimination." In *Algorithmic Regulation*, edited by K. Yeung and M. Lodge, 82–97. New York: OUP.

Davenport, T., and R. Kalakota. 2019. "The Potential for Artificial Intelligence in Healthcare." *Future Healthcare Journal* 6 (2): 94–98. doi:10.7861/futurehosp.6-2-94.

Dignum, V. 2020. "Responsibility and Artificial Intelligence." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 703–718. New York: OUP.

Diogo, V. S., H. A. Ferreira, and D. Prata. 2022. "Early Diagnosis of Alzheimer's Disease Using Machine Learning: A Multi-Diagnostic, Generalizable Approach." *Alzheimer's Research & Therapy* 14 (107), doi:10.1186/s13195-022-01047-y.

Dolata, M., S. Feuerriegel, and G. Schwabe. 2022. "A Sociotechnical View of Algorithmic Fairness." *Information Systems Journal* 32 (4): 754–818. doi:10.1111/isj.12370.

Edwards, L., and M. Veale. 2017. "Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy you are Looking for." *Duke Law & Technology Review* 16: 18–84. https://scholarship.law.duke.edu/dltr/vol16/iss1/2.

European Commission. 2020. "White Paper on Artificial Intelligence - A European Approach to Excellence and Trust, COM (2020) 65 final (19 February 2019)." https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Fabrizio, C., A. Termine, C. Caltagirone, and G. Sancesario. 2021. "Artificial Intelligence for Alzheimer's Disease: Promise or Challenge?" *Diagnostics* 11 (8), doi:10.3390/diagnostics11081473.

Faulkner, P. 2011. *Knowledge on Trust*. Oxford: OUP.

FDA. 2017. "What are Examples of Software as a Medical Device?" https://www.fda.gov/medical-devices/software-medical-device-samd/what-are-examples-software-medical-device.

FDA. 2019. "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) – Discussion Paper". https://www.regulations.gov/document?D = FDA-2019-N-1185-0001.

FDA. 2021. "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan." https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.

Financial Conduct Authority. 2021. "FCA to Introduce new Consumer Duty to Drive a Fundamental Shift in Industry Mindset." *Press Release.* Accessed December 7, 2021. https://www.fca.org.uk/news/press-releases/fca-introduce-new-consumer-duty-drive-fundamental-shift-industry-mindset.

Ford, R. A., W. Price, and I. Nicholson. 2016. "Privacy and Accountability in Black-box Medicine." *Michigan Telecomm & Technology of Law Review* 23 (1): 1–43. https://repository.law.umich.edu/mttlr/.

Gerke, S., T. Minssen, and G. Cohen. 2020. "Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare." In *Artificial Intelligence in Healthcare*, edited by Adam Bohr and Kaveh Memarzadeh, 295–396. Academic Press. doi:10.1016/B978-0-12-818438-7.00012-5

Glikson, E., and A. W. Woolley. 2020. "Human Trust in Artificial Intelligence: Review of Empirical Research." *Academy of Management Annals* 14 (2): 627–660. doi:10.5465/annals.2018.0057.

Grimm, P. W., M. R. Grossman, and G. V. Cormack. 2021. "Artificial Intelligence as Evidence." *Northwestern Journal of Technology and Intellectual Property* 19 (1): 9–106. https://jtip.law.northwestern.edu/issues/artificial-intelligence-as-evidence/.

Hacker, P., R. Krestel, S. Grundmann, and F. Naumann. 2020. "Explainable AI Under Contract and Tort law: Legal Incentives and Technical Challenges." *Artificial Intelligence and Law* 28: 415–439. doi:10.1007/s10506-020-09260-6.

Hall, M. A. 2019. "Fiduciary Principles in Health Care." In *The Oxford Handbook of Fiduciary Law*, edited by Evan C. Cridle, Paul B. Miller, and Robert H. Sitkoff, 286–302. New York: OUP.

Holdsworth, M., and M. Zaghloul. 2022. "The Impact of AI in the UK Healthcare Industry: A Socio-Technical System Theory Perspective" *Proceedings of the 8th International Workshop on Socio-Technical Perspective in Information Systems Development* (STPIS 2022) Reykjavik, Iceland, August 19–20, 2022. 52–63. https://ceur-ws.org/Vol-3239/paper6.pdf.

Imperial College London. 2022. "Body Sensor Networks." https://www.imperial.ac.uk/hamlyn-centre/research/sensing/body-sensor-networks/.

Independent High-Level Expert Group (IHLEG) on Artificial Intelligence (set up by the European Commission). 2019. "Ethics Guidelines for Trustworthy Artificial Intelligence." https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Jacovi, A, A. Marasović, T. Miller, and Y. Goldberg. 2021. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. doi:10.1145/3442188.3445923.

Jie, Z., Z. Zhiying, and L. Li. 2021. "A Meta-Analysis of Watson for Oncology in Clinical Application." *Scientific Reports* 11 (5792), doi:10.1038/s41598-021-84973-5.

Konig, P. D., T. D. Krafft, W. Schulz, and K. A. Zweig. 2022. "Essence of AI: What is AI?" In *In The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*, edited by L. A. DiMatteo, C. Poncibò, and M. Cannarsa, 18–34. Cambridge: CUP.

Lee, S. S. 2022. "Philosophical Evaluation of the Conceptualisation of Trust in the NHS' Code of Conduct for Artificial Intelligence-Driven Technology." *Journal of Medical Ethics* 48: 272–277. doi:10.1136/medethics-2020-106905.

Leslie, D., J. Cowls, M. Katell, and M. Briggs. 2021. *Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A Primer Prepared for the Council of Europe*. The Alan Turing Institute. doi:10.5281/zenodo.4369743.

Lewis, P. R., and S. Marsh. 2022. "What is it Like to Trust a Rock? A Functionalist Perspective on Trust and Trustworthiness in Artificial Intelligence." *Cognitive Systems Research* 72: 33–49. doi:10.1016/j.cogsys.2021.11.001.

Liu, K. 2020. "Deep Learning to Distinguish Pancreatic Cancer Tissue from non-Cancerous Pancreatic Tissue: A Retrospective Study with Cross-Racial External Validation." *Digital Health* 2: 303–313. doi:10.1016/S2589-7500(20)30078-9.

Lockey, S., N. Gillespie, D. Holm, and I. A. Someh. 2021. "A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions" P*roceedings of the 54th Hawaii International Conference on System Sciences*. January 5–8. Kauai, HI, USA.

Lysaght, T., H. Y. Lim, V. Xafis, and K. Y. Ngiam. 2019. "AI-Assisted Decision-Making in Healthcare." *Asian Bioethics Review* 11: 299–314. doi:10.1007/s41649-019-00096-0.

Ma, L., G. Lu, D. Wang, X. Qui, Z. G. Chen, and B. Fei. 2019. "Adaptive Deep Learning for Head and Neck Cancer Detection Using Hyperspectral Imaging." *Visual Computing for Industry, Biomedicine, and Art* 2(18). doi:10.1186/s42492-019-0023-8.

Markou, C., and S. Deakin. 2021. "Ex Machina Lex: Exploring the Limits of Legal Computability." In *Is Law Computable: Critical Perspectives on Law and Artificial Intelligence*, edited by Simon Deakin and Christopher Markou, 31–66. Cambridge: CUP.

McLeod, C. 2021. "Trust." In *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2021/entries/trust/.

Meijer, A., and S. Grimmelikhuijsen. 2020. "Responsible and Accountable Algorithmization: How to Generate Citizen Trust in Governmental Usage of Algorithms." In *The Algorithmic Society: Technology, Power and Knowledge*, edited by M. Schuilenburg and R. Peeters, 52–66. New York: Routledge.

Middleton, S. E., E. Letouzé, A. Hossaini, and A. Chapman. 2022. "Trust, Regulation, and Human in-the-Loop AI Within the European Region." *Communications of the ACM* 65 (4): 64–68. doi:10.1145/3511597.

Misselhorn, C. 2022. "Artificial Moral Agents: Conceptual Issues and Ethical Controversy." In *The Cambridge Handbook of Responsible Artificial Intelligence*, edited by S. Voeneky, 31–49. Cambridge: CUP.

Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society July-December* 2016: 1–21. doi:10.1177/2053951716679679.

Nickel, P. J. 2022. "Trust in Medical Artificial Intelligence: A Discretionary Account." *Ethics and Information Technology* 24: 7. doi:10.1007/s10676-022-09630-5.

OECD. 2019. *Artificial Intelligence in Society*. Paris: OECD Publishing. doi:10.1787/eedfee77-en.

Origgi, G. 2020. "Trust and Reputation as Filtering Mechanisms of Knowledge." In *The Routledge Handbook of Social Epistemology*, edited by M. Fricker, P. J. Graham, D. Henderson, and N. J. L. L. Pedersen, 78–86. New York: Routledge.

Panesar, S. 2005. "Fiduciary Relationships and Constructive Trusts in a Commercial Context." *International Company and Commercial Law Review* 16 (12): 474–484.

Pasquale, F. 2022. "Liability Standards for Medical Robotics and AI." In *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*, edited by L. A. DiMatteo, C. Poncibò, and M. Cannarsa, 200–212. Cambridge: CUP.

Powers, T. M., and J. Ganascia. 2020. "The Ethics of the Ethics of AI." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 27–51. New York: OUP.

Practical Law Corporate. 2022. "Fiduciary Duties" Practical Law UK Practice (Note 8-107-4883).

Radiological Society of North America. 2018. "Artificial Intelligence Predicts Alzheimer's Years Before Diagnosis." *ScienceDaily*. Accessed November 6, 2018. www.sciencedaily.com/releases/2018/11/181106104249.htm.

Rodrigues, R. 2020. "Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities." *Journal of Responsible Technology*, doi:10.1016/j.jrt.2020.100005.

Rousseau, D. M., S. B. Sitkin, R. S. Burt, and C. Camerer. 1998. "Not so Different After all: A Cross-Discipline View of Trust." *Academy of Management Review* 23 (3): 393–404. doi:10.5465/amr.1998.926617.

Ryan, M. 2020. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." *Science and Engineering Ethics* 26: 2749–2767. doi:10.1007/s11948-020-00228-y.

Ryan, P. 2021. *Trust and Distrust in Digital Economies*. London: Routledge.

SAE International. 2021. "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104 (Revised version 2021-04-30) https://www.sae.org/standards/content/j3016_202104/.

Salvi, S., and A. Kadam. 2021. "Breast Cancer Detection Using Deep Learning and IoT Technologies." *Journal of Physics Conference Series*, 1831. https://iopscience.iop.org/article/10.10881742-6596/1831/1/012030/meta.

Samir, A. E., L. Brattain, and M. Baikpour. 2022. "Evolving Role of Artificial Intelligence in Radiological Imaging: No-Human-in-the-Loop AI-enabled Healthcare: Risk, Rewards, and Regulation." https://www.fda.gov/media/135732/download.

Sartori, L., and A. Theodorou. 2022. "A Sociotechnical Perspective for the Future of AI: Narratives Inequalities, and Human Control." *Ethics and Information Technology* 24: 4. doi:10.1007/s10676-022-09624-3.

Scantamburlo, T., A. Charlesworth, and N. Cristianini. 2019. "Machine Decisions and Human Consequences." In *Algorithmic Regulation*, edited by Karen Yeung and Martin Lodge, 49–81. New York: OUP.

Schwemer, S. F., L. Tomada, and T. Pasini. 2021. "Legal AI Systems in the EU's Proposed Artificial Intelligence Act." *CEUR Workshop Proceedings (LegalAIIA 2021)* 2888: 69–76. https://curis.ku.dk/ws/files/282950779/paper8_2.pdf.

Sharan, N. N., and D. M. Romano. 2020. "The Effects of Personality and Locus of Control on Trust in Humans Versus Artificial Intelligence." *Heliyon* 6 (8): e04572. doi:10.1016/j.heliyon.2020.e04572.

Starke, G., R. V. D. Brule, B. S. Elger, and P. Haselager. 2022. "Intentional Machines: A Defence of Trust in Medical Artificial Intelligence." *Bioethics* 36(2): 154–161. doi:10.1111/bioe.12891.

Toews, R. 2022. "What Artificial Intelligence Still Can't Do" Forbes. https://www.forbes.com/sites/robtoews/2021/06/01/what-artificial-intelligence-still-cant-do/?sh = 733cdfdd66f6.

Turner, J. 2020. *Robot Rules: Regulating Artificial Intelligence*. London: Palgrave Macmillan.

UK Department of Health and Social Care. 2021. "A Guide to Good Practice for Digital and Data-Driven Health Technologies." https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology.

Unver, M., and O. Asan. 2022. "Role of Trust in AI-Driven Healthcare Systems: Discussion from the Perspective of Patient Safety." *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* 11 (1): 129–134. doi:10.1177/2327857922111026.

Velasco, J. 2021. "Fiduciary Judgement Rules." *William and Mary law Review* 62 (4): 1397–1448. https://scholarship.law.wm.edu/wmlr/vol62/iss4/8.

Wachter, S., B. Mittelstadt, and L. Floridi. 2017. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7 (2): 76–99. doi:10.1093/idpl/ipx005.

West, D. M., and J. R. Allen. 2018. "How Artificial Intelligence is Transforming the World?" *Brookings*. https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/.

Wheeler, M. 2020. "Autonomy." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 343–357. New York: OUP.

Widlak, A., M. Van Eck, and R. Peeters. 2020. "Towards Principles of Good Digital Administration: Fairness, Accountability and Proportionality in Automated Decision-Making." In *The Algorithmic*

Society: Technology, Power and Knowledge, edited by M. Schuilenburg and R. Peeters, 67–84. New York: Routledge.

World Health Organization. 2021. *Ethics and Governance of Artificial Intelligence for Health (WHO Guidance)*. Geneva: WHO.

Yeung, K. 2019. "Why Worry About Decision-Making by Machine?" In *Algorithmic Regulation*, edited by K. Yeung, and M. Lodge, 21–48. New York: OUP.

Yu, X., S. Xu, and M. Ashton. 2023. "Antecedents and Outcomes of Artificial Intelligence Adoption and Application in the Workplace: The Socio-Technical System Theory Perspective." *Information Technology & People* 36 (1): 454–474. doi:10.1108/ITP-04-2021-0254.

Zafari, S., and S. T. Koeszegi. 2018. "Machine Agency in Socio-Technical Systems: A Typology of Autonomous Artificial Agents" IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO) Genoa, Italy, September 27–28.

Zahid, N., A Hassan, S. U. R. Kamboh, A. Alkhayyat, and L. Wang. 2022. "AI-Driven Adaptive Reliable and Sustainable Approach for Internet of Things Enabled Healthcare System." *Mathematical Biosciences and Engineering* 19 (4): 3953–3971. doi:10.3934/mbe.2022182.

Zewe, A. 2022. "AI That Can Learn the Patterns of Human Language" MIT News Office. https://news.mit.edu/2022/ai-learn-patterns-language-0830.