Ethics and Information Technology (2022) 24:7
https://doi.org/10.1007/s10676-022-09630-5

ORIGINAL PAPER

Trust in medical artificial intelligence: a discretionary account

Philip J. Nickel[1]

Accepted: 5 January 2022 / Published online: 24 January 2022

Abstract
This paper sets out an account of trust in AI as a relationship between clinicians, AI applications, and AI practitioners in which AI is given discretionary authority over medical questions by clinicians. Compared to other accounts in recent literature, this account more adequately explains the normative commitments created by practitioners when inviting clinicians' trust in AI. To avoid committing to an account of trust in AI applications themselves, I sketch a reductive view on which discretionary authority is exercised by AI practitioners through the vehicle of an AI application. I conclude with four critical questions based on the discretionary account to determine if trust in particular AI applications is sound, and a brief discussion of the possibility that the main roles of the physician could be replaced by AI.

Keywords Artificial intelligence · Trust in AI · Discretion · Normative expectations · Future of medicine

Introduction

Medical artificial intelligence ("AI") applications are more than just an algorithm and a flow of data. They incorporate an interface with professional users and a sociotechnical embedding. The functioning of the AI application depends on this embedding (van de Poel, 2020, p. 391). As Taddy argues, "the success or failure of an AI system is defined in a specific *context*, and you need to use the structure of that context to guide the architecture of your AI" (2019, p. 65). In the field of medicine, clinical validation of AI applications must be shown in a real-world environment (Nagendran et al., 2020). As Durán and Jongsma observe, "Algorithms in medicine should be understood to provide *input* for clinical decision-making but they cannot decide by themselves how [to act] on the results" (2021, p. 333). The AI application, encompassing algorithms and an interface, is accorded authority by the clinician. It might not make decisions by itself, but it can strongly influence decision-making. When used in this way as clinical decision support tools, AI applications are embedded in clinicians' work in structured ways. For example, they might be designed to flag patients as high risk based on what is feasible in the organizational context and be expected in a certain percentage of cases to lead to additional clinical measures such as discussing an intervention with patients (Gallagher et al., 2020). In this paper, I argue that the discretionary authority accorded to AI applications based on reasonable expectations about their functioning is a significant kind of trust.

Recently a number of philosophers have criticized the idea of trust in medical AI. The critiques not only deny that we *can* meaningfully trust AI (Ryan, 2020; Wolfensberger & Wrigley, 2019), but also maintain that we *should not* do so (Hatherley, 2020; Ryan, 2020; Tallant, 2019). According to these critiques, genuine trust is a human interpersonal concept, depending on rich affective and normative attitudes. It cannot meaningfully be applied to medical AI because the underlying technology does not have the characteristics necessary to underwrite these attitudes. Moreover, it should not be applied to medical AI because doing so leads to a corrupted form of trust in a domain where rightful trust is of paramount importance. Properly speaking, one can only *rely* on medical AI, but not trust it.

Previous defenses of trust in AI use a non-normative concept of trust. Ferrario et al. (2020a) claim that there is a "simple" notion of trust involving the *reduction of monitoring and control* of the outputs of AI. I go a different direction, arguing that there is a stronger notion of trust in AI applications involving *giving it discretionary authority*, a kind of normative authority. To a greater degree than the

✉ Philip J. Nickel
p.j.nickel@tue.nl

[1] Department of Philosophy and Ethics, School of Innovation Sciences, Eindhoven University of Technology, PO Box 513, 5600MB Eindhoven, The Netherlands

simple notion, this stronger, discretionary notion explains why inviting user trust entails moral commitments. When the designers and deployers of AI invite user trust relating to a certain understood medical objective served by the AI application, for example deciding on a treatment, they thereby assume the discretionary authority to answer medical questions in context. We cannot tell this story without a robust notion of trust. I will leave it open whether this form of trust is reducible to trust in the engineers and deployers of AI applications. In order to maintain this as a live option, I sketch a philosophically viable reductive account of trust in AI in "Trust in AI practitioners" section.

The theoretical background for this account is a pragmatic view that sees how we define a concept like trust as partly determined by the concept's explanatory value (McLeod, 2002; Nickel, 2017). Interpersonal trust is used to explain distinctive patterns of cooperation, as well as the moral commitments that come along with these patterns. Trust marks out a distinctive way of adopting and relating to technologies such as automation, robots, and AI, and signals the moral commitments of those who design and deploy them.

The critique of trust in AI

It is useful to begin by summarizing recent philosophical literature on trust in medical AI. According to several recent critiques, trust toward AI is inappropriate because it attributes impossible characteristics to AI. A typical critique is that of Ryan (2020), who holds that genuine trust must be *normative* or *affective*, not merely rational or calculating. On normative views, trust is partly constituted by a person's normative expectation that the object of trust ought to live up to certain standards or fulfill certain commitments. The attitude of trust includes ascribing normative agency and responsibility to the object of trust. On affective views, trust ascribes a motive of goodwill to its object, such that it is assumed to be affectively and motivationally responsive to the fact that it is being relied upon. On both normative and affective views of trust, AI does not possess the characteristics to support the ascription of normative agency or goodwill; an indication of this is that we would be reluctant to say that AI can *betray* those who rely on it (ibid., 2757). Ryan emphasizes the wrongful nature of placing trust in AI: both the incorrectness of anthropomorphizing AI in itself, as well as the negative consequences for attributions of responsibility. Placing trust in AI harmfully confuses questions of responsibility: "Normative accounts of trust require moral agents to be held responsible for their actions" (ibid., 2762). The specific wrong involved is to impute responsibility to AI because this "inappropriately elevate[s] AI, while disavowing the responsibility of those developing and implementing it" (ibid)..

Hatherley (2020) presents an argument with a similar premise, claiming that AI can merely be reliable but not trustworthy, and that trustworthy AI is therefore a conceptual misunderstanding. By contrast to Ryan's argument, however, Hatherley emphasizes that trust in medical AI threatens to "displace the epistemic authority of human clinicians" (p. 478) by replacing the key tasks that physicians perform, thereby putting the trust relationship between patient and physician at risk. For Hatherley, what is most important is the downstream effects of relying on AI for these key tasks, rather than the inaptness of the attitude of trust toward AI. In fact, it hardly matters to Hatherley's overall argument that one cannot genuinely trust AI, since the point is simply that if AI develops superior reliability in diagnosis, prognosis, and treatment, it will threaten to supplant trust in human clinicians (480).[1] I will come back to this point about the displacement of epistemic authority in "Four critical questions about trust in medical AI" and "Conclusion" sections. For now let us focus on the claim that one cannot trust AI.

Ferrario et al. (2020a) respond explicitly to Hatherley, arguing that there is a "simple" notion of trust that might guide the reliance of a physician on an AI application after an initial period of experience: "After a sufficient number of trials, the physician would eventually entertain beliefs on the performance and error patterns of the medical AI. Therefore, at the next interaction with the medical AI, the physician could trust the AI by relying on it without updating these beliefs. This is expressed by a disposition of the physician to exert little [effort] and time in further activities instrumental to belief updating" (ibid).. For Ferrario et al., this new disposition in the physician's behavior, in which s/he exhibits little or no monitoring of the AI's outputs, counts as trust. It seems to follow that the disposition not to monitor or check the AI application can be understood as an attitude of the clinician toward herself. She takes no normative attitude toward the functioning of the AI. In this respect, simple trust is the clinician's own responsibility, and she has only herself to blame if it proves unwise.

Ferrario et al. remark that their account is not meant to restrict the grounds that may be involved in the formation

[1] Not every critique includes this claim. For example, Bryson (2018) gives two arguments why we should not trust AI. The first is that humans are not on the same level as AI. For Bryson, it is only appropriate to trust somebody or something that is a peer, and humans are not peers of AI. The second argument is that we should not trust AI, because we should strive for something better. We should make those who design and deploy AI accountable, in such a way that we do not need trust. One might question this latter argument by observing that accountability can support trust, rather than replacing it. Explaining why a person gets into to a driverless car calmly (or fearfully) to get from point A to point B seems to be furthered by making reference to their trust (or distrust). Accountability of manufacturers may tilt the balance toward trust, rather than making trust irrelevant.

of trust in AI (2020a, see footnote 4). Hence the motive of the person who trusts could be affective or normative, but it could equally be completely "rational," i.e., strategic. More to the point, the central necessary and sufficient element of their account of simple trust—abstaining from monitoring or controlling the relied upon entity or having a belief that one does not need to—is neither normative nor affective in itself (2020b, p. 537). Their account is anti-normative and anti-affective in the sense that normative and affective components of trust are superfluous to trust in AI. Below I argue that a more normative account of trust based on discretionary authority is useful to explain the interconnections between the expectations of users, the invitation of trust within user interfaces, and the commitments of AI practitioners. To explain these interconnections, we need an attitude of trust that is *normative* and *directed toward AI and AI practitioners*. The normativity is both functional, in the sense that one expects the AI application to perform certain functions, and moral, in the sense that if the application has been advertised by practitioners as being reliable for those functions in context and this is not the case, moral blame may be apt.

Normative trust in AI

My positive account of trust in medical AI applications is based on the discretionary authority given to them by clinicians. In my view of trust, one entity is disposed to give a second entity *discretion* over some matter of value on the basis of normative and predictive expectations about that second entity. This conception is inspired by the philosophical literature on trust, which has emphasized both the *discretion* and *vulnerability* entailed by trust (Baier, 1986), as well as the (often implicit or ascribed) moral *commitments* involved in both trust and distrust (Hawley, 2014). The philosophical literature differs from much of the literature in economics and game theory in its insistence that there is an essential normative, moral, or affective component to trust (Cohen, 2020). Although I do not maintain that this normative component is the same for all kinds of trust, I do think it applies to the case of trust in AI.

I aim to show that there is a plausible notion of trust in medical AI that is grounded in reasonable, realistic attitudes of clinicians and explains the moral commitments of AI practitioners. It not my goal to show that AI can meet all of the moral and affective conditions of traditional philosophical accounts of interpersonal trust. I give two main arguments for the positive view. The first is that it best explains the way that AI practitioners describe what they are doing, and the moral commitments they take on by inviting clinician trust in AI applications. The second is that the key elements of the view are fitting and reasonable as an interpretation of clinician attitudes toward AI applications. Accepting the view means taking a broader view of what trust is. However, a range of conceptions of trust, including self-trust, trust in animals and institutions, and infant-parent trust, is already widely accepted (McLeod, 2002). Clearly, not all of these conceptions satisfy the same conditions. People make use of them because they have practical and explanatory value.

AI practitioners often use the language of trust in relation to AI and exploit social and affective features of the user interface to invite trust.[2] In a characteristic summary article written for an audience of programmers and information technology managers, Siau and Wang argue that trust is important for "acceptance and continuing progress and development of artificial intelligence" (2018, p. 47). They avoid adopting a definition of trust, preferring a general statement that trust is both cognitive and behavioral, where the cognitive component includes "a set of specific beliefs dealing with benevolence, competence, integrity, and predictability" (ibid).. They view trust in AI as unproblematic but do not sharply distinguish it from strategic reliance. Let us call this the *AI practitioner's view*.

The AI practitioner's view is oriented toward fostering trust in AI, robots, and automation for organizational and instrumental ends. This often involves the presentation of AI in the guise of affective and social characteristics. For example, Siau and Wang state, "Humans are social animals. Continuous trust can be enhanced with social activities. A robot dog that can recognize its owner and show affection may be treated like a pet dog, establishing emotional connection and trust" (Siau & Wang, 2018, p. 51). Linking this with AI, they advise other practitioners that "the representation of an AI as a humanoid or a loyal pet will facilitate initial trust formation" (ibid., 52). Along similar lines, a widely referenced synthesis of the literature on trust in automation refers to many means of fostering trust through the exploitation of affective and social cues (Hoff & Bashir, 2015). The use of cues to encourage social, affective, and normative responses disposing a person toward reliance is what I mean by *inviting* trust.

Incidentally, this aspect of the AI practitioner's view is the target of Ryan's critique from the previous section, in which anthropomorphism, via trust, leads to wrongful attitudes and confusion about the apportionment of responsibility for AI's effects. Recall that Ryan's critique is based on two main premises: first, that trust in AI implies ascribing it moral and affective characteristics that it does not possess; and second, that doing so leads to incorrect and confused

[2] I use the term *AI practitioners* to refer to professionals in the fields of AI, robotics, and automation whose job it is to apply AI in a practical context.

beliefs about responsibility for AI. I aim to undermine the first of these premises by developing a discretionary account of trust in AI and showing how this can be reduced to trust in AI practitioners, and the second by pointing out the genuine ethical implications of inviting trust in AI, suggested by the discretionary account.[3] Ryan is right that our trust attitudes toward AI *can* be wrongly anthropomorphic. But they do not have to be.

The central notion on which I pin the positive account of trust in AI is *giving discretionary authority*. Discretion refers to a circumscribed authority accorded to another entity; it is a frequently mentioned hallmark of trust. The legal scholar H.L.A. Hart writes that discretion's "distinguishing feature" is that the answer to a question "is not determined by principles which may be formulated beforehand, although the factors which we must take into account and conscientiously weigh may themselves be identifiable" (Hart, 2013, p. 661). On Hart's view, discretion involves being entrusted with the authority to determine the answer to such a question (ibid., 665). We may conclude that, like trust itself, giving discretion can usefully be understood as a three-place relation of one entity to another entity, such that the first comes to place this kind of *discretionary authority* in the hands of a second within a domain of interaction (Baier, 1986).

Under some conditions it can be reasonable for a clinician to assume that an AI application is less subject to relevant errors and "noise" than his or her own judgment in a given domain, and therefore reasonable to grant it discretionary authority in answering certain questions (Durán & Jongsma, 2021). In practice, it may be difficult to convince physicians to accept such authority, but it is the stated aim of AI practitioners in the medical field (Polonski, 2018). There are many pragmatic grounds why discretionary authority is given to answer a question: because the answer is *arbitrary*, because it is a matter of *convention*, or because it involves *coordination problems* such as a Prisoners Dilemma (Raz, 1986, pp. 48–50). These grounds can apply in the domain of medicine, where there are instances of arbitrariness (e.g., where two treatments are evidentially non-inferior to one another), convention (e.g., where it is useful for multiple clinicians to adopt a standard approach), and even coordination problems involving factual questions (Nyland et al., 2017; Reay & Hinings, 2009). There is always a danger of injustice in the arbitrary exercise of discretion (Pratt & Sossin, 2009). This is connected to the essential vulnerability of trust (Baier, 1986).

In trust, one accords another entity discretionary authority because one has relevant normative and predictive expectations toward it.[4] One often reasonably expects it to fulfill the *role* or *function* designated for it within its context. For example, when an AI application has the function to support decisions about discharging patients from a given hospital ward, this makes it reasonable, other things equal, for a physician to expect it to do a decent job at making relevant predictions of readmission or death. Because this is a normative expectation, it is not about what the clinician predicts, but about what the technology is *for*. This normative expectation is the typical basis for evaluating whether the technology is well designed and works properly. Users and engineers apply values and norms to technologies deriving from these technologies' functions: "The notion of function plays a normative role … it tells us what an artefact ought to do" (Vaesen, 2013, p. 119). When such function-based expectations are relevant to the needs and goals of clinicians, they provide the basis for giving some of the clinician's own discretionary authority to the AI application, allowing it to (help) answer questions that previously went unanswered, or that were previously answered using other means.

Transferring discretionary authority to another entity carries distinctive moral weight. The entity exercising that discretion enjoins an obligation to deliver what its function or role requires in the context. A clinician may have formerly answered a question such as whether a given fragile patient is prescribed a given intervention using a small range of readily available data together with their own clinical observations. She made such judgments using her own professional discretion, developed through experience. Now, suppose she answers that question largely on the basis of a color-coded risk score based on algorithmic analysis of the patient's chart and biomarkers, adopted as part of an AI solution for the hospital in which she works. By taking over this discretionary authority from the clinician, the AI application becomes the *object*, but not the bearer, of a moral obligation. Its function is to answer a question within the context in such a way that values such as fairness, well-being, efficiency, and transparency are not undercut. AI practitioners are morally obligated, other things equal, to ensure that it carries out this function within these constraints. This obligation is owed *to*

[3] Regarding the second premise, it might be doubted whether the trust one develops in the anthropomorphized guise of AI really contributes to a responsibility gap. Typical organizational factors, such as the many actors involved in designing, deploying, and maintaining the AI, seem to offer a better explanation.

[4] It is conceivable that one accords another entity discretionary authority not because one thinks it can and should perform a given function or role, but merely because one has no other good option. This is not a case of trust. Conversely, one can have normative and predictive expectations of an entity without giving it any authority (perhaps because it is not relevant to one's needs, or because there are comparably good alternatives). This is also not trust. This is why both elements are required: discretionary authority must be based on normative and predictive expectations to count as trust.

the clinician, as the one whose discretionary authority has been transferred in an act of trust. This helps to bring out the specific normative force of trust in AI, as compared with the generic moral obligations surrounding any new care technology put into service for the first time.

AI whose function is to represent things as being a certain way, especially using natural language or programmed visual representation, entails normative expectations that can underwrite discretionary authority over factual questions. By issuing a factual statement such as a risk prediction for a patient or a classification of the parts of a scan, an AI application presents something as true. This is similar to assertion or written testimony. There is a widely discussed class of "norms of assertion" such that by making an assertion, a speaker is normatively committed to the truth, justification, or warrant of the asserted claim (Pagin, 2016). It is plausible to suppose that such norms of assertion underlie or even partly constitute testimonial trust (Simion, 2020), and to explain users' expectations of artificially generated speech and representations in the medical field using these same norms of assertion, considering that this is a context where truth and justification are important. Humans' evaluations of artificial speech for competence and dishonesty are parallel to their evaluations of human speech (Kneer, 2020). We can make sense of the idea of a robot or complex technological system being deceptive, and therefore we can apply a norm of non-deceptiveness to it (Nickel, 2013). This is not to say that people conflate algorithmic utterances with human ones. On the contrary, research shows that humans use different heuristics to evaluate an algorithm's utterance, compared with a human utterance (Efendic et al., 2020). Parallel points apply even if we think of AI applications as a kind of instrumentation rather than as an author of assertions (Freiman & Miller, 2020). The point is that the norms of assertion and representation provide a basis for giving authority over certain factual questions.

On this account, a user trusts an AI application when she is disposed to give it discretionary authority over practically important questions on the basis of normative and predictive expectations about its performance in context. It is realistic and reasonable for professionals to give such authority to the outputs of AI applications. Hence, the AI practitioners view is not wrong to describe what AI offers to clinicians in the language of trust, nor to offer suggestions for inviting trust through the user interface. Describing it this way does not lead ineluctably to malignant anthropomorphism or a responsibility gap. On the contrary, by inviting clinicians' trust, AI practitioners create normative commitments, inviting the user to give discretionary authority to the application on the basis of normative expectations about its performance. This transfer of authority implies ethical obligations on the part of the AI practitioner toward the clinician.

Trust in AI practitioners

A complete account of trust in medical AI must include trust in AI practitioners themselves. Indeed, it has sometimes been suggested that trust in AI is *nothing more* than trust in the designers, deployers, and overseers of the AI (Sutrop, 2019, p. 512). Let us call this the *reductive view*. In this section I sketch a version of the reductive view that supports the discretionary account. Let it be clear that this is not the view that trust in AI does not exist or is not explanatory, but rather the view that we can translate the moral content of statements about trust in AI into statements about human and institutional elements.

The reductive view must be squared with the impersonal relationship between professional users and AI practitioners. The user has often never met the AI practitioners and cannot identify them. In addition, individual practitioners usually do not personally communicate a commitment to users and have no personal knowledge of the users' specific reliance on the technology or the exact situations in which it occurs. Literature predicts that in the future, "reinforcement learning algorithms will become companion physician aids, unobtrusively assisting physicians and streamlining clinical care," and that "[machine learning] will become increasingly easy and commoditized" (Johnson et al., 2018, p. 2678). If it is unobtrusive and commoditized, it is likely to be impersonal to a high degree, ruling out all but the most abstract trust in AI practitioners. Therefore, to make sense of the reductive view, one must suppose that trust does not require acquaintance or concrete commitments between the parties. This is considered a problem by some theorists of trust, because it precludes the possibility that the trusted party takes the specific reliance of the trusting party into account in her actions, and that the trustee bases her trust on specific knowledge of the motivations of the trusted (Hardin, 2006; McLeod, 2000).

Despite this concern, there are many cases where we speak of trust toward people with whom we are not acquainted. Suppose during a period stuck at home in quarantine, one might order many consumer goods online and receive multiple postal packages, but never actually see a person delivering the packages, and not be able to distinguish the voices of the delivery people heard over the intercom. Suppose on the basis of beliefs about and experiences with package delivery services, the client comes to trust *whoever delivers packages to the apartment building* without even knowing if there is just one person or multiple people who do so. This seems like a real instance of interpersonal trust. It is in this sense that we should understand a reductive view on which trust in an AI application is ultimately vested in *whoever designed and deployed it*.

Parallel points can be made about the AI practitioner's conceptualization of the use contexts and users.

Another challenge for the reductive view is that it needs to provide a means by which discretion is exercised by the AI practitioner. Bluntly put, the clinician relies on the outputs of the AI application, not on the outputs of the practitioner. When the clinician accepts a factual claim on the basis of the application, s/he is not accepting it on the authority of an AI practitioner. When s/he judges that *it* is doing a good job, this is not the same as judging that *they* are doing a good job. (Good practitioners could produce a bad AI application, and bad ones could produce a good application). In addition, no practitioner invites the user's trust directly, at least not typically. For these reasons, trust toward the practitioners does not obviously explain the normative dimensions of trust in AI, for it is not closely related to the authority accorded to the outputs of the AI application.

A version of the reductive view on which both the AI application and the practitioners are objects of user attitudes and are linked to one another can solve this problem. As explained previously, the AI application itself is a proximal, concrete object of reasonable normative expectations and is given discretion to answer certain questions. The AI practitioners are an indirect object with whom the clinician is not acquainted, bearing the moral obligation to ensure those expectations are fulfilled. The clinician trusts the practitioners *through* the application. When the practitioners invite and sustain user trust, it is also through the application. An (admittedly imperfect) analogy might be drawn to how a spectator relates to a composer through the experience of a work of music, never seeing or hearing the composer (and perhaps not even knowing their name) but coming to form a judgment about them through the experience of the work. Conversely, the composer forms expectations about the audiences and performances of the work. By analogy, on this reductive view, practitioners play an essential role in inviting and supporting trust in the technology, one layer removed from the experience of the user. They are the ultimate indirect object of user trust in the application. When they knowingly place an AI application that answers questions for clinicians into a practical context, this carries moral weight and invites moral expectations because of the human agency behind it.

Four critical questions about trust in medical AI

The concept of discretionary authority, introduced in "Normative trust in AI" section, defines the way that certain functions of the physician are transferred to AI applications. Discretion has a finite scope: it is always given relative to a question or a domain of activity. Dworkin (1977) characterizes legal discretion as an empty "donut hole" in which the rules plus the evidence do not determine a definitive answer to a question.[5] The judgment of the professional is normally active in this empty space, where there is no easy answer to a question. This is also the space where AI can be deployed to discern evidence more finely as well as settle pragmatic issues such as consistency across clinical departments, marshaling of data in ways that are useful for research, and efficient use of scarce resources. Even when professionals are given the power to override the judgment or recommendation given by the application, it can have a strong influence both psychologically and institutionally on what professional judgments are found reasonable (see Fagan & Levmore, 2019 for judicial examples).[6]

Using the account developed here, we can ask four critical normative questions, a proper answer to which should satisfy us that trust is well-placed. This helps respond to concerns about the replacement of the core functions of physicians by AI and displacement of patient-physician trust (Hatherley, 2020, op cit).:

(1) Over what question does the clinician give discretionary authority to an AI application?
(2) Are there good epistemic and pragmatic reasons for giving this discretion?
(3) Are these same reasons available to the clinician?
(4) Is the discretionary authority properly limited in scope?

[5] In fact, Dworkin was caricaturing the legal positivist view of discretion using the donut metaphor (1977, 31ff).. On his own view, the empty space is actually filled with reasons of a different kind—not rules, but principles and policies requiring legal judgment.

[6] There are at least two other possible ways in which AI can affect professional medical discretion. First, AI applications can be used to block professionals from stepping outside of their proper area of discretion (by e.g., blocking discrimination or medical error). Second, they can be used to widen the scope of discretion, e.g., by generating risk scores that empower professionals to treat additional patients or widen the scope of an intervention program (see Brayne 2017 for comparable examples from law enforcement). However, I do not examine such cases here. I focus on the situation in which AI answers questions that reduces the discretion of the professional by replacing their judgment.

Let us discuss each of these questions in turn, using the example of AI applications that predict readmission to hospital. I then conclude with some general reflections.

(1) *Over what question does the clinician give discretionary authority to an AI application?* The questions over which the clinician gives discretion to an AI application are dependent on the application's specific function. For example, an application might be designed to answer the question, "What is likely to be the most impactful intervention for the prevention of hospital readmission for this patient" on the basis of electronic health record data (Jamei et al., 2017). If the clinician judges that it is worthwhile to rely on the application, and that the application should be able to answer this question adequately, we can speak of a trust relation that disposes her to give discretionary authority to the application. By fostering these expectations, the practitioner takes on commitments of due care toward clinician and patient.

(2) *Are there good epistemic and pragmatic reasons for giving this discretion?* There are different kinds of valid reasons for giving discretionary epistemic authority to an AI application. On the one hand, the paradigmatic epistemic reason for doing so is the application's reliability: it makes fewer incorrect judgments than the clinician (e.g., about the desirability or undesirability of an intervention). This reason related to reliability can obtain even if the internal workings of the AI application are opaque to the user, as Durán and Jongsma (2021) argue. On the other hand, some reasons for giving discretionary authority to AI are pragmatic, such as when allowing the AI application to answer a given question would reduce costs to the hospital, or when doing so leads to more consistent decisions among different physicians on a ward. Then again, there are reasons with both an epistemic and a pragmatic aspect, such as when using the AI application makes it easier to conduct research comparing interventions and thereby acquire new knowledge.

(3) *Are these same reasons available to the clinician?* There are many channels for clinicians to learn about valid reasons to give discretion to an AI application, such as direct involvement in design and implementation, communication with a hospital technology coordinator, trainings (Durán & Jongsma, 2021, p. 334), or simply interaction with the application interface. Scholars of trust classify these channels of information into those prior to direct experience with the system (which could influence a decision to adopt or trust in the system) and those derived from direct experience with the system (Hoff & Bashir, 2015). Explainable AI has received a great deal of attention as a way of bolstering the accountability of professionals and creating transparency. Such explanations can offer a way to establish understanding of an AI application during one's experience with it, and this can create or sustain trust (Hoffman et al., 2018). However, it would be a mistake to regard explanations offered by AI for a given output as being the same reasons that justify giving it discretionary authority. Discretion can be granted before one even has experiences with actual reliance or been offered explanations for particular decisions.

The user interface plays an important role in revealing the reasons for relying on the AI application to the clinician. For example, the risk of hospital readmission and intervention options for particular patients are visualized or presented in a certain way in the user interface of a decision support system. This gives the user cues about what kinds of reasons are being used to draw inferences by the AI application. When the information presented to the clinician consists of an estimated percentage likelihood for each patient that s/he will need to be readmitted within 30 days, this suggests that the relevant outputs are drawn from a statistical relationship between the data in the patient's electronic health record and other similar patients' average readmission rate. Color coding and classification of risks as "high" or "moderate" suggest pragmatic factors. A picture of an interface containing these factors is presented in Schreiner et al. (2020), where the authors study how clinicians combine automated risk scores with human expert judgment to make their own risk estimates.

An AI application can have multiple functions and users with different interests. The potential lack of awareness of and transparency about the multiple functions of an AI application is a major ethical issue leading to a negative answer to question (3). AI applications can have secondary functions or uses unrelated to the aims for which the clinician is meant to use the application. An example would be the use by management of an AI-based clinical support application to assess and compare clinician performance, or cost-outcome ratios. When the clinician is not aware of and would not endorse secondary uses such as this one, the clinician's trust can be said to be *unsound* (Voerman & Nickel, 2017).

(4) *Is the discretionary authority properly limited in scope?* Some of the most important ethical risks of AI in medicine are related to lax guardianship over the boundaries of discretionary authority. In such cases, trust in the AI application is too broad or too strong. Examples include function creep, automation bias, and deskilling. *Function creep* is defined as the use of data and analytical tools originally designed for one purpose (e.g., quality control) for another purpose (e.g., surveillance) (cf.

Koops, 2021). In a clinical context, an example is using an application designed to promote patient outcomes at the department level to evaluate individual employee performance. Aaen et al.'s (2021) study of the Danish DAMD project suggests that reuse of medical data (analytics) for multiple purposes is a significant risk when multiple and changing stakeholders are involved in the data "ecosystem". *Automation bias* is the widely documented tendency to over-trust automation, even when one's own judgment should raise a red flag about doing so (Goddard et al., 2012). For example, clinicians sometimes treat the output of an automated diagnosis as a baseline for their own judgment, using the widely studied psychological "anchoring" heuristic to make a judgment in a way that does not sufficiently correct for errors (Bond et al., 2018). *Deskilling* refers to a process by which trust in automation or other forms of technology causes one to lose important skills that require use and practice to be maintained (Vallor, 2015). Learning from failure, one of the core practices of modern medicine, can be blocked in cases where there is little insight into the reasons why a judgment or treatment was recommended by AI. The problem of opacity in some AI systems can therefore contribute to deskilling (Macrae, 2019).

A theoretical benefit of the account of trust in AI developed here is that it allows us to think of these issues in terms of the moral obligations AI practitioners come to have toward clinicians when AI applications that invite trust are deployed in context. Yet clinicians need to think critically about these constraints on the proper delegation of discretionary authority to AI, exhibiting what Manson and O'Neill call "intelligent trust" (2007). The critical questions above are meant to refine this idea.

Conclusion

In the future, the domain over which the clinician transfers discretionary authority to AI applications could conceivably increase to the point where it would encompass many of the questions that patients now turn to clinicians to answer. This is what Hatherley has in mind when worrying that most of the roles essential to the physician–patient trust relationship might be transferred to AI. It would be a gradual and piecemeal process at first. Shaw et al. argue, quoting Agrawal et al. (2018, p. 125), that 'the actual implementation of AI is through the development of tools'. The unit of AI tool design is not 'the job' or 'the occupation' or the 'the strategy', but rather 'the task'". Therefore, for a health care provider to be entirely replaced, every single task performed by that provider would need to be automated by [a machine learning] tool or handed off to a different human' (Shaw et al., 2019). Although this could be justified if the four critical questions from "Four critical questions about trust in medical AI" section are answered to our satisfaction for each of the clinician's tasks, we might still worry that something of value, essential to the patient-physician relationship, has been lost.

A possible response to Hatherley's concerns is that clinicians will take on new roles no less essential for the future practice of medicine. Two that are often mentioned in the discussion of AI by medical professionals are the roles of researcher and care manager. The researcher role cannot easily be replaced by AI because the past is a moving target: as the world and technological opportunities change, old datasets become outdated and predictions less accurate. In order for medicine to learn and improve, systematic research needs to be designed and carried out by humans. In order to make AI work, innovation and research will become an even greater part of the clinician's role than they already are (Institute of Medicine, 2007). Patients must trust clinician-researchers when consenting to serve as research participants, a fact "especially relevant to health care Al systems, which require diverse and large amounts of data (from people) in order to optimize outcomes" (Feldman et al., 2019, p. 405).[7]

The role of care manager is also mentioned frequently as one that cannot easily be replaced by AI (Briganti & Moine, 2020). For example, one radiologist pondering the impact of AI writes that "there is little likelihood of patients in the near future accepting to be cared for and treated entirely by a machine without human intervention when they can be made to feel secure by help provided to doctors by technology," and that "radiologists will go from the status of 'ball counters' with coarse tools to that of data controllers processing increasingly sophisticated quantified results" (El Hajjam, 2020). At least in the medium term, patients' trust in AI will be strongly mediated by their trust in physicians. Patients will place their individual health interests in the hands of physicians working with AI, and this act of trust will require that physicians are not themselves robotic or algorithmic in their decision making.

This leaves us with a picture of the future of trust in medical AI where the desired situation is one in which four elements—patients, clinicians, AI applications, and AI practitioners—are aligned in a relationship of properly bounded trust and corresponding ethical commitment even as the practice of medicine itself changes. By accounting for this complexity, the normative account of trust presented in this paper provides insight and horizons of research for

[7] There are potential conflicts between the roles of researcher and care professional that can complicate trust in data-intensive health care (Faden et al., 2013).

clinicians, AI practitioners, and ethicists in this rapidly developing field.

Funding Funding was provided within the NWO-MVI project "Mobile Support Systems for Behavior Change" (grant number 100-23-616) and the project "Ethics of Socially Disruptive Technologies" funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (grant number 024.004.031).

References

Aaen, J., Nielsen, J. A., & Carugati, A. (2021). The dark side of data ecosystems: A longitudinal study of the DAMD project. *European Journal of Information Systems*. https://doi.org/10.1080/0960085X.2021.1947753

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.

Baier, A. (1986). Trust and antitrust. *Ethics, 96*, 231–260.

Bond, R. R., et al. (2018). Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *Journal of Electrocardiology, 51*, S6–S11.

Brayne, S. (2017). 2017 Big data surveillance: The case of policing. *American Sociological Review, 82*(5), 977–1008.

Briganti, G., & Le Moine, O. (2020). Artificial intelligence in medicine: Today and tomorrow. *Frontiers in Medicine*. https://doi.org/10.3389/fmed.2020.00027

Bryson, J. J. (2018). *AI and global governance: No one should trust AI*. United Nations Centre for Policy Research. Retrieved May 21, 2021 from https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html

Cohen, M. A. (2020). Trust in economy. In J. Simon (Ed)., *The Routledge handbook of trust and philosophy* (pp. 283–297). Routledge.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*, 329–335. https://doi.org/10.1136/medethics-2020-106820

Dworkin, R. (1977). *Taking rights seriously*. Harvard University Press.

Efendic, E., van de Calseyde, P. P. F. M., & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes, 157*, 103–114.

El Hajjam, M. (2020). Toward an augmented radiologist. In B. Nordlinger, C. Villani, & D. Rus (Eds)., *Healthcare and artificial intelligence*. Springer.

Faden, R. R., Kass, N. E., Goodman, S. N., Pronovost, P., Tunis, S., & Beauchamp, T. L. (2013). An ethics framework for a learning health care system: A departure from traditional research ethics and clinical ethics". *Ethical Oversight of Learning Health Care Systems, Hastings Center Report Special Report, 43*(1), S16–S27. https://doi.org/10.1002/hast.134

Fagan, F., & Levmore, S. (2019). The impact of artificial intelligence on rules, standards, and judicial discretion. *Southern California Law Review, 93*, 1.

Feldman, R., Aldana, E., & Stein, K. (2019). Artificial intelligence in the health care space: How we can trust what we cannot know. *Stanford Law and Policy Review, 30*, 399–419.

Ferrario, A., Loi, M., & Viganò, E. (2020a). Trust does not need to be human: It is possible to trust medical AI. *Journal of Medical Ethics*. https://doi.org/10.1136/medethics-2020-106922

Ferrario, A., Loi, M., & Viganò, E. (2020b). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy and Technology, 33*, 523–539. https://doi.org/10.1007/s13347-019-00378-3

Freiman, O., & Miller, B. (2020). Can artificial entities assert? In S. Goldberg (Ed)., *The Oxford handbook of assertion*. Oxford University Press.

Gallagher, D., Zhao, C., Brucker, A., Massengill, J., Kramer, P., Poon, E. G., & Goldstein, B. A. (2020). Implementation and continuous monitoring of an electronic health record embedded readmissions clinical decision support tool. *Journal of Personalized Medicine, 10*(3), 103. https://doi.org/10.3390/jpm10030103

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*, 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Hardin, R. (2006). *Trust*. Polity Press.

Hart, H. L. A. (2013). Discretion. *Harvard Law Review, 127*, 652–665.

Hatherley, J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*, 478–481. https://doi.org/10.1136/medethics-2019-105935

Hawley, K. (2014). Trust, distrust, and commitment. *Noûs, 48*, 1–20.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*, 407–434.

Hoffman, R.R., Mueller, S.T., Klein, G., & Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. http://arXiv.org/1812.04608v2

Institute of Medicine Roundtable on Evidence-Based Medicine. (2007). *The learning healthcare system: Workshop summary*. National Academies Press.

Jamei, M., Nisnevich, A., Wetchler, E., Sudat, S., Liu, E., & Upadhyaya, K. (2017). Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PLoS ONE, 12*, 7. https://doi.org/10.1371/journal.pone.0181173

Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology, 71*, 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

Kneer, M. (2020). *Can a robot lie?*. https://doi.org/10.13140/RG.2.2.11737.75366

Koops, B.-J. (2021). The concept of function creep. *Law, Innovation and Technology, 13*, 29–56.

Macrae, C. (2019). Governing the safety of artificial intelligence in healthcare. *BMJ Quality & Safety, 28*, 495–498.

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge University Press.

McLeod, C. (2000). Our attitude towards the motivation of those we trust. *The Southern Journal of Philosophy, 38*, 465–479.

McLeod, C. (2002). *Self-trust and reproductive autonomy*. MIT Press.

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., et al. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ, 368*, m689. https://doi.org/10.1136/bmj.m689

Nickel, P. J. (2013). Artificial speech and its authors. *Minds and Machines, 23*, 489–502.

Nickel, P. J. (2017). Being pragmatic about trust. In P. Faulkner & T. Simpson (Eds)., *The philosophy of trust* (pp. 195–213). Oxford University Press.

Nyland, K., Morling, C., & Burns, J. (2017). The interplay of managerial and non-managerial controls, institutional work, and the coordination of laterally dependent hospital activities. *Qualititative Research in Accounting and Management, 14*, 467–495.

Pagin, P. (2016). Assertion. In E. N. Zalta (Ed). *The Stanford encyclopedia of philosophy (Winter 2016 Edition)*.

Polonski, V. (2018). People don't trust AI: Here's how we can change that. *The Conversation*. Retrieved June 27, 2021 from https://theconversation.com/people-dont-trust-ai-heres-how-we-can-change-that-87129

Pratt, A., & Sossin, L. (2009). A brief introduction of the puzzle of discretion. *Canadian Journal of Law and Society, 24*, 301.

Raz, J. (1986). *The morality of freedom*. New York: Oxford University Press.

Reay, T., & Hinings, C. R. (2009). Managing the rivalry of competing institutional logics. *Organization Studies, 30*, 629–652. https://doi.org/10.1177/0170840609104803

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*, 2749–2767.

Schreiner, J. H., Thurston, D. L., & Willemsen-Dunlap, A. (2020). Readmission risk assessment technologies and the anchoring and adjustment heuristic. *Journal of Medical Systems, 44*, 61. https://doi.org/10.1007/s10916-020-1522-z

Shaw, J., Rudzicz, F., Jamieson, T., & Goldfarb, A. (2019). Artificial intelligence and the implementation challenge. *Journal of Medical Internet Research, 21*, e13659. https://doi.org/10.2196/13659

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*, 47–53.

Simion, M. (2020). Testimonial contractarianism: A knowledge-first social epistemology. *Noûs*. https://doi.org/10.1111/nous.12337

Sutrop, M. (2019). Should we trust artificial intelligence? *Trames, 23*, 499–522.

Taddy, M. (2019). The technological elements of artificial intelligence. In A. Agrawal, J. Gans, & A. Goldfarb (Eds)., *The economics of artificial intelligence: An agenda* (pp. 61–87). University of Chicago Press.

Tallant, J. (2019). You *can* trust the ladder, but you shouldn't. *Theoria*. https://doi.org/10.1111/theo.12177

Vaesen, K., et al. (2013). Artefactual norms. In M. J. de Vries (Ed)., *Norms in technology. Philosophy of engineering and technology* (Vol. 9, pp. 119–136). Springer.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy of Technology, 28*, 107–124. https://doi.org/10.1007/s13347-014-0156-9

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds & Machines, 30*, 385–409.

Voerman, S. A., & Nickel, P. J. (2017). Sound trust and the ethics of telecare. *Journal of Medicine and Philosophy, 42*, 33.

Wolfensberger, M., & Wrigley, A. (2019). *Trust in medicine: Its nature, justification, significance, and decline*. Cambridge University Press.