Impacts on Trust of Healthcare AI

Emily LaRosa[1] **& David Danks**[1,2]

[1]-Department of Philosophy; [2]-Department of Psychology

Carnegie Mellon University

5000 Forbes Ave., Pittsburgh, PA 15213, USA

elarosa@andrew.cmu.edu ddanks@cmu.edu

Abstract

Artificial Intelligence and robotics are rapidly moving into healthcare, playing key roles in specific medical functions, including diagnosis and clinical treatment. Much of the focus in the technology development has been on human-machine interactions, leading to a host of related technology-centric questions. In this paper, we focus instead on the impact of these technologies on human-human interactions and relationships within the healthcare domain. In particular, we argue that trust plays a central role for relationships in the healthcare domain, and the introduction of healthcare AI can potentially have significant impacts on those relations of trust. We contend that healthcare AI systems ought to be treated as assistive technologies that go beyond the usual functions of medical devices. As a result, we need to rethink regulation of healthcare AI systems to ensure they advance relevant values. We propose three distinct guidelines that can be universalized across federal regulatory boards to ensure that patient-doctor trust is not detrimentally affected by the deployment and widespread adoption of healthcare AI technologies.

ACM Reference format:

Emily LaRosa and David Danks. 2018. Impacts on Trust of Healthcare AI. In *Proceedings of 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), AIES '18, February 2–3, 2018, New Orleans, LA, USA,* ACM, NY, NY, USA, 6 pages. https://doi.org/10.1145/3278721.3278771

Roles for Healthcare AI

Artificial Intelligence and robotics are rapidly moving into healthcare, and these technologies will play key roles in strategically and intelligently supporting diverse medical functions. Healthcare AI and robotics are proposed for, either now or in the near-future: diagnosis of patients; performance of simple surgeries; well-defined tasks within more complex procedures (e.g., closing incisions with sutures or staples); monitoring of patients' health and mental wellness in short- and long-term care facilities; basic physical interventions to improve patient independence during physical or mental deterioration (e.g., physical aid, or reminders to take medications); independent patient mobility (e.g., voice command wheelchairs); and even particular tasks requiring physical interventions in dynamic contexts (e.g., blood draws).

Much of the focus in healthcare technology development has been on human-machine interactions: How do we ensure that a home healthcare robot does not harm the patient? How do we develop diagnostic systems that provide superior performance to human doctors? And a host of related technology-centric questions. Moreover, there have been multiple analyses of methods and techniques for establishing doctor- or patient-machine trust (e.g., [9]).

We instead focus on the impacts of these technologies on human-human interactions, and particularly on patient-doctor trust relationships. A presupposition of our paper is that healthcare AI and robotics have the potential (though not necessity) to reshape human-human relations of all types and levels in healthcare. For instance, the decision to use an AI or robot to care for ourselves or loved ones can potentially have physical and psychological impacts on the patient, familial unit, and bonds created by human caregiving. Thus, the decision to use a home healthcare robot for an elderly parent should not depend "merely" on the safety of the robot, but should incorporate other potential impacts. For example, it might strengthen the parent-child bond by enabling more meaningful interactions (since the robot handles menial tasks); or alternatively, weaken it by reducing the number of interactions (as the child is not needed).

Healthcare AI and robotics also have the potential to reshape human-institution relationships and interactions. Patients frequently gain support from informal institutions (e.g., support groups), but the socially implied roles of individuals in such informal contexts may need to be explicitly or formally codified if healthcare AI provides emotional or behavioral aids, rehab support, or other similar functions. In a different setting, healthcare AI and robotics can be expected to alter doctor-institution relationships through impacts on the insurability of doctors and healthcare institutions, and through potential changes in the social and institutional nature (and necessity) of the primary care physician. At yet a higher level, relationships with and between international organizations (e.g., World

AIES'18, February 2–3, 2018, New Orleans, LA, USA.

ACM ISBN 978-1-4503-6012-8/18/02...$15.00.

DOI: https://doi.org/10.1145/3278721.3278771

Health Organization, multinational pharmaceutical companies) will presumably shift as they influence who will have access to, be the focus of, and potentially benefit from, these novel technologies.

Healthcare AI and robotics clearly have the potential for far-reaching impacts on diverse human-human relationships. In the remainder of this paper, we exemplify this potential through a close focus on arguably the most fundamental and intimate of human-human relationships in the healthcare domain: the patient-doctor relationship. In particular, we primarily focus on potential impacts of healthcare AI on patient-doctor *trust*. We begin by characterizing the general multidimensional notion of trust, and then consider (partly informed by historical changes) the current ways in which patient-doctor trust develops and is maintained. At that point, we are positioned to explore the specific challenges and opportunities presented by healthcare AI technologies for relations of trust that ought—for both pragmatic and ethical reasons—obtain between patient and doctor. We conclude with a brief discussion of potential regulatory and policy implications.

The Nature of Trust

At a high level, human trust involves the trustor making herself vulnerable based on expectations about the trustee's likely actions, intentions, or capabilities (see, e.g., [7] for an overview of research on trust). This high-level characterization is ambiguous about the reason for the trustor's expectations. In particular, there are roughly two distinct, not mutually exclusive, types of trust.

First, trust can be grounded in reliability, in the sense of the trustee being predictable by the trustor. For example, one can trust that one's car will start in the morning because it has reliably started on past occasions. More generally, this type of trust depends on the trustor's beliefs about what the trustee will do in known contexts. We refer to this type of trust as "behavioral trust," as the trustor's knowledge is essentially behavioral: she does not know the mechanisms by which the trustee's "behavior" is generated, but only the likely behaviors in particular situations. This type of trust provides no basis for generalization to truly novel situations, since it is grounded simply in the actual past experiences. Nonetheless, behavioral trust can provide the requisite bases —predictions and expectations—for coordinated trustor-trustee action. In general, this type of trust is appropriate for artifacts or other machines, as well as contexts in which the trustor and trustee have only limited knowledge of one another (e.g., in game-theoretic experiments or situations).

Second, trust can be grounded in an understanding of the "mechanisms" (again, broadly construed) by which the behavior or actions are generated. That is, this type of "understanding trust" is based on the trustor's beliefs about why the trustee acts as it does in a particular case. For example, one's trust of a close friend is based on understanding that friend's beliefs, desires, values, and intentions, rather than the ability to perfectly predict what that friend will do. Understanding trust carries the significant advantage (compared to behavioral trust) of generalizing to novel situations, precisely because the trustee can use her "why"-knowledge to predict the trustee's behaviors and intentions in previously unexperienced settings. Although understanding trust can be established with an artifact, it is far more commonly found with other humans, precisely because we have rich "theories of mind" to explain other people's actions using our knowledge of their beliefs and values.

As suggested by the earlier example of a close friend, these types of trust can come apart. On the one hand, understanding trust need not provide the level of predictive ability required to establish behavioral trust. On the other hand, behavioral trust can be established without knowledge of the mechanisms that generate the trustee's behavior, and hence without understanding trust.

We frequently come to trust another—whether artifact or human—because of the role that the trustee plays in a larger system or organization. For example, our trust in the auto mechanic at our local garage is grounded in her inhabiting a particular role within the organization, coupled with our expectation (or assumption) that the garage would not employ her if she did not have appropriate knowledge and skills.

Role-based trust is not a third kind of trust that is distinct from behavioral or understanding trust, but rather acts as a vehicle to establish one or the other kind. In particular, the exact nature of particular role-based trust depends on the role that grounds the trust. If the role is defined by (reliable) behavior, then it will yield behavioral trust since the trustor can expect (all else being equal) that the trustee will perform the behaviors that define her role. For example, one can trust that a garbage collector will reliably (though not always!) pick up the trash every Wednesday morning, as the role is largely defined by the performance of that behavior. That same role-based trust does not, however, provide any insight into the values, beliefs, or intentions of our garbage collector. In contrast, if the role is defined by function or skill, then said role typically yields understanding trust as the trustor can, unless countervailing reasons are provided, assume that the trustee has the relevant skills and "ways of thought" that define the particular role.

Doctors provide a canonical example of role-based trust, at least on first encounter. The "role" of doctor is defined largely in terms of skills and knowledge, and so we can trust that a doctor will have particular beliefs (e.g., knowledge about human physiology), values (e.g., patient health supersedes the doctor's personal preferences), and intentions (e.g., act to improve the patient's health to the best of her ability). Ergo, patient-doctor understanding trust can be initially grounded in the role that doctors presently occupy, with the recognition that this trust can shift over time.

This form of trust also extends to medical teams, in which different medical professionals must interact and coordinate in support of patient care. In many modern care scenarios, patient-doctor trust extends beyond the scope of the one-to-one interactions of general practitioners to patient trust in the entirety of the team, many of whom might engage with the patient only occasionally or sporadically. In these cases, coordinated role-based trust becomes even more vital; if one of the team fails, the trust for the whole may be damaged.

The Nature of Patient-Doctor Trust

A critical foundational relationship in healthcare interactions is trust between patients and doctors. This trust is clearly not merely behavioral; patients frequently cannot predict their doctor's particular actions, precisely because the doctor has a wealth of knowledge and skills that the patient does not. Rather, patients place understanding trust in their doctors: they make themselves vulnerable because of their expectations and beliefs about their doctor's knowledge, skills, values, and intentions. At the same time, those expectations have shifted significantly over the past forty years, as the dominant narrative of doctors has shifted from a paternalistic model to a collaborative, patient-centric model of interaction. This shift in role has produced a corresponding shift in what is attributed to the doctor, and that thereby grounds the patient's understanding trust.

The medical role of doctor was initially laden with paternalism. A given patient would trust his[1] doctor to make decisions on his (the patient's) behalf, and in his best interests. Moreover, the doctor's decisions were made independently of the patient's interpretations of, or ability to adhere to, the care decisions. "Doctor knows best" was a mantra repeated loudly and often. The understanding trust was thus grounded in the doctor's paternalistic role in which she (the doctor) had the patient's health as the highest value, regardless of his other preferences. The doctor was attributed largely infallible skills and knowledge, and so relatively unquestioned understanding trust developed (see [1, 5]).

In the late 1980s, partly spurred by the AIDS pandemic, the doctor's role shifted in the West with the emergence of the model of an "active" patient who collaborates on care with his doctor. Patients became partners in diagnostic and care processes, and understood themselves as decision-making equals; the patient recognized his role as the primary stakeholder in his own care [2, 3]. Patient input and influence in the medical encounter is now a critical part of the nature and model of doctor-patient interactions, though with notable exceptions. This reconceptualization led to increasing focus on patients' rights, democratization of care, and reduction in medical paternalism [2].

More importantly for our present purposes, the shift towards an "active patient" also changed the doctor's role, and thereby the grounding for the patient's understanding trust in his doctor. In particular, the doctor was no longer assumed to know the patient's core values. Thus, the doctor must elicit her patient's values and preferences through discussion and interaction, rather than simply imputing her beliefs (about patient values) to the patient. Understanding trust is still possible in this new situation, but it is predicated on the doctor successfully learning the patient's values. If she fails to do so, then the patient will likely exhibit reduced, or absent, trust in his doctor.

One general, though significant, challenge to patient-doctor trust is exactly these differences between the patient's and doctor's goals and desired outcomes. By default, humans typically "project" their own values, preferences, and beliefs onto others, at least in the absence of evidence to the contrary [10]. Patients may thus default to assuming that their doctors share their values. Both sides can agree that the overarching goal and desired outcomes of care are to support and advance the patient's best interests, but that agreement can mask differences in interpretation of the meaning of 'best interests'.

These divergences can be particularly acute when the patient's care needs differ from his near-term desires. For example, patients diagnosed with bipolar disorder and treated with lithium need to have weekly blood tests to ensure that detrimental levels of lithium do not build up in their system. However, many patients find this constant monitoring to be overly burdensome and disruptive, resulting in them preferring a local desire for convenience over long-term need for monitoring. If the doctor (correctly) focuses on the patient's needs, the patient will judge his doctor to have the "wrong" (for purposes of trust) values, leading to a potential barrier to patient-doctor trust. This challenge can be further exacerbated if the patient defers entirely to his doctor's judgment, asking, "what would you do?" and simply accepting the doctor's answer without further reflection. The "active patient" model depends on the patient having understanding trust in his doctor for the right reasons: correct beliefs, whether grounded in the doctor's role or significant experiences, that the doctor has appropriate skills, knowledge, and understanding of the patient's values and preferences.

Healthcare AI & Patient-Doctor Trust

Healthcare AI and robotics can impact patient-doctor trust by influencing any of the multiple ways in which patients develop this understanding trust in their doctors. We focus here on three key routes to trust, all of which are potentially supported or disrupted by the introduction of healthcare AI or robotic systems.

First, doctors are explicitly certified and licensed to practice medicine. Licensure indicates that particular individuals have specific skills, knowledge, and high-level values (e.g., "do no harm"). Grantors of licenses serve to ensure that individuals satisfy certain public, objective criteria, and thereby justify lay people's corresponding expectations of those holding a license. In the healthcare domain, grantors of licenses provide one key basis for understanding trust, as they provide grounding for patients' justified expectations about the reasons for doctors' actions, even if those actions are not *a priori* predictable. As such, the license grantor must articulate public standards that can be reviewed, understood, and critiqued by the larger community or society (as opposed to granting licenses based on judgments of a "black box" decision-making system).

Now consider the introduction of a healthcare AI or robotics system for a particular treatment or diagnosis task. To the extent that this system replaces a task traditionally performed by the doctor, it potentially threatens to displace some patient-doctor trust: licensure of the doctor can no longer ground

[1] For the remainder of the paper, we use masculine pronouns to refer to patients, and feminine pronouns to refer to doctors.

understanding trust for that task, so the patient needs to know whether the system is appropriately approved or "licensed" for the functions it performs. That is, the impact will depend on whether appropriate regulatory approval mechanisms and frameworks are in place for the particular function performed by the AI or robotic system.

On the one hand, the function might be defined purely behaviorally (e.g., apply stitches to an incision in the AI's camera field), where the doctor must judge whether the present context is suitable for its use. In this case, regulation can proceed along the same lines as for non-autonomous medical devices, using ordinary performance standards. On the other hand, if the function is defined by successful outcomes or values (e.g., determine the treatment that balances the patient's interests), then the system itself will need to judge which actions are appropriate for the present context. In that case, we cannot specify clear performance standards (since we do not necessarily know the contexts of operation), and so should instead regulate the AI or robotic system in a manner similar to novel drugs or other medical interventions [4].

The second route to patient-doctor understanding trust is through the particular social role that doctors play as part of an active dyad (or larger system) charged with ensuring care that supports the patient's values. This social role justifies a default attribution of various knowledge and values to the doctor, namely those that are necessary to serve this social role (which may be different from the licensed role). The impact of healthcare AI and robotics on this route to trust depends on whether such systems change the social role of 'doctor' in ways that impact the social assumptions and expectations about doctors' knowledge and understanding of the AI systems. To the extent that doctors are viewed as "mere users" of the AI, we would expect the social role to shift away from the doctor being a distinctive repository of knowledge and skills, thereby undermining a trust-promoting element of the social role. If doctors are instead viewed as intelligent users of AI systems, then the social role should likely shift towards higher degrees of perceived or imputed expertise, thereby promoting role-based understanding trust.

Third, and finally, a patient's experiences with his doctor are a significant driver—potentially positive or negative— for patient-doctor understanding trust. As the patient has repeated interactions with his doctor, understanding trust will shift as he gains additional evidence about his doctor's distinctive abilities, competences, and knowledge of values and desires [7]. For example, if the patient has an open line of communication with his doctor and engages in conversation about care and treatment, then the patient should experience increasing understanding trust. Inversely, if the doctor repeatedly ignores the patient's wishes, then these actions will have a negative impact on understanding trust. When a doctor engages with a patient, she builds social and experiential "capital" with him, resulting in increased understanding trust.

Consider now the incorporation of an AI diagnostic system that reduces: the likelihood of misdiagnosis; the lack of diagnosis by the doctor (but caught by the AI); and maltreatment such as over-, under-, or inappropriate prescription of medications or other legitimate medical alternatives. This type of AI system should naturally lead to improved (diagnostic) experiences with the doctor, and thereby increased trust. Moreover, these successes should improve the patient's care management plan, further increasing his trust in the healthcare system. Of course, to the extent that both the AI system is suboptimal (e.g., large numbers of false positives) and the doctor delegates diagnostic or decision-making authority to the AI, we should expect the patient's experiences to be negative, so reduce trust in his doctor. The reliability of accurate diagnosis and beneficial treatment regimens, as well as appropriate use by the doctor, are critical in strengthening a patient's understanding trust.

This analysis presupposes a stable power dynamic between the doctor and patient. If the patient sees the use of AI in his care management plan as impeding his ability to act as a partner with his doctor, it will become vital for the physician to right the dynamic back to a stable equilibrium through discussions, education, and reconfirmed consent.

This example also shows how healthcare AIs can have complex impacts on patient-doctor trust. The use of a diagnostic support system could simultaneously lead to improved experiences on the part of the patient, while also undercutting the position of the doctor as an authority on medical matters (if she delegates too much authority to the AI). That is, these systems can have distinct positive and negative impacts on the development of patient-doctor understanding trust, and it is an empirical matter whether such systems thereby provide a net benefit or net detriment.

Healthcare AI and robotic systems can impact all three of these "routes to trust" in distinct ways, and so assessment will necessarily be quite complex (and dependent on contingent details of the setting). As a final example, consider a healthcare monitoring AI that dynamically presents appropriate information to the doctor. People are often more willing to provide information to an AI than a human [8], particularly when that information is socially negative (e.g., failure to take medication). Thus, this monitoring AI has the potential to gather more objective information, thereby improving the patient's experiences and outcomes (the third route to trust). In making her patient's willingness to follow specific care regimens and reception to treatment more transparent, an AI can support a doctor in her pursuit of the best course of treatment, potentially eliminating costly rounds of medical testing.

At the same time, the use of such an AI may lead to a change in social role (second route) if the doctor is perceived as "offloading" important work to the AI, rather than engaging in a collaboration with the patient or others on their medical team (to gain information, learn what has been happening, and so forth). By using an AI to monitor behavior, the doctor has changed the sourcing of information and altered the collaborative nature of the patient-doctor relationship. This change in social role potentially damages patient-doctor trust, precisely because an important type of knowledge—namely, of the patient's experiences—can no longer be assumed solely by virtue of the doctor inhabiting a particular social role. Direct patient-

doctor communication is an important part of the (current) social role, and grounds part of the understanding trust. The AI's role in this communication must be understood and agreed upon by both parties to avoid damaging that trust.

Finally, licensure and regulation (first route) become critically important if the AI dynamically presents only the "appropriate" information to the doctor. Simple behavioral measures are insufficient to capture the notion of 'appropriate' in these contexts [4]. Hence, if this system is approved and regulated solely using such measures (as with medical devices), then licensing of a particular doctor must ensure that she can judge what information is "appropriate," perhaps by having sufficient understanding of the AI monitoring system. An alternative, more practical path would be to regulate this type of AI as a novel medical intervention, as that incremental, dynamic approach can better determine relevant performance profiles and suitable contexts [6]. For purposes of trust development, the key is that overall regulation and licensure of doctor-plus-AI must ensure "appropriateness judgments" are evaluated, wherever those judgments are made.

Regulatory Policy Recommendations

Policy and regulation can potentially play powerful roles to ensure the development and maintenance of patient-doctor trust, even as AI and robotic systems are introduced into the healthcare ecosystem. We need to proactively establish direct, comprehensive, scientifically-based policies that are decipherable by the layperson. These measures should (on ethical grounds) be focused on the patient's welfare, rather than privileging "mere" technological development or the business case. The healthcare system is already highly regulated, but it is important that AI and robotic systems not fall outside of this apparatus. At the same time, AIs function in highly diverse capacities and roles, and so actual regulation requires specificity about each domain or technology. We here articulate only general principles that present a challenge to AI developers and deployers to yield technologies that benefit the patient and healthcare ecosystem, rather than developing without focus on wider impacts.

Our first regulatory principle is suggested by our earlier observations that patient-doctor trust will likely be damaged if doctors are perceived (socially) to abdicate their current roles to AI systems. We thus propose: Doctors using AI systems and their results must have educational training that is overseen, measured, and approved by an independent outside group. This principle blocks the social role of 'doctor' from being weakened towards doctors being (perceived as) "mere button-pushers," thereby supporting patient-doctor understanding trust. This principle could be implemented through existing mechanisms of continuing education, though we emphasize that is not the only such mechanism. Regulators (or even insurers) could alternately compel certain types of education or knowledge as a precondition for use of a healthcare AI or robotics technology.

At the same time, it is not sufficient to ensure knowledge *only* by the doctor; the patient also needs to understand what a healthcare AI or robot can, and cannot, do so that he can be appropriately informed about its use, and how it potentially changes both the doctor's role and the patient's subsequent experiences.

This leads us to our second regulatory proposal: AI should not be used for patient care without the *educated* consent of the patient or caregiver. Educated consent is more stringent that informed consent, which only requires that a patient be supplied the care-relevant information. Instead, educated consent involves patients in a conversation about these protocols and procedures, and requires more active forms of consent. This education could potentially take many forms, ranging from passive information transmission to direct patient-AI/robot interactions prior to use in their care. These efforts would undoubtedly be influenced by changing social perceptions of AI capabilities, as people will transfer beliefs about AI capabilities in one domain to its capabilities elsewhere. The establishment of a healthy trust relationship with the AI requires proper implementation by the healthcare professional who has the patient's primary trust.

We emphasize that the motivation for this second principle is to support patient-*doctor* trust, not patient-AI trust (though it surely would also help with that). This educational effort will thus likely have the impact of further shifting the doctor-patient dyad towards a team dynamic, as there will now be shared knowledge and understanding of the technology. Of course, while patients or their caregivers have a right to be fully educated prior to making decisions, this requirement presents a substantial burden on the use of AI and robotic technologies, since these can be difficult to understand. Nonetheless, there is a clear need for such education to support patient-doctor trust. In this regard, healthcare AI and robotic systems are no different from any other medical intervention, where patient-doctor trust would be jeopardized by its use without educated consent of the patient, both because of the negative experience of betrayal (third route) and also the resulting shifts in social role for 'doctor' away from trusted advisor (second route).

A related observation leads to our third principle. If a patient perceives that a technology's use is taking priority over his wellbeing, then he will likely experience significantly reduced trust of his doctor, as well as the whole healthcare system. Such a perception (that needs of technology are dominant) may well result from a lack of presented options: if the patient does not perceive an alternative, then his "decision" is based on blind faith, which leads back to the paternalistic paradigm that was less consistent with understanding trust. Lack of codified, accepted, viable alternatives threatens the patient's understanding trust in his doctor. We thus propose: Until a healthcare AI is accepted as "standard of care," the doctor must provide the alternative of a human performing the assigned task or function. This principle implies that many AI systems should be regulated as medical interventions, not devices, precisely because they should be evaluated against a "standard of care" criterion that encompasses not just the technology, but also methods and application contexts. Such a move would require a staged, dynamic regulatory system; such a framework already exists (e.g., the U.S. Food & Drug Administration, or FDA), but it would

require treating AI and robotic systems as interventions, not devices. We contend, though, that this approach is necessary to ensure continued patient-doctor trust in light of the autonomous capabilities in these systems.

We have presented three high-level regulatory principles, and we close by sketching some possibilities for implementation, though we emphasize that our focus here has been on an analysis of impacts on trust, and not on the particular legal or political paths to implementation. We focus on the United States as we are most familiar with it, but also because it is one of the most complex medical systems due to its highly decentralized nature.

The need to ensure doctor knowledge of the AI's capabilities is naturally addressed by licensing bodies, such as the American Medical Association. These groups are ideally positioned to ensure that doctors have sufficient knowledge and information to not abdicate their social and licensed roles to the AI or robot. The focus on patient education can be addressed in large measure by insurers, perhaps led by the Centers for Medicare & Medicaid Services (CMS). CMS operates the U.S. federal branches of medical insurances, and many private insurers take their lead from it. A standard promulgated by CMS could outline conditions under which healthcare AIs and robots could be used (for particular procedures and conditions). In particular, this standard could include substantial patient education components as preconditions for AI and robot use. Such a standard would likely have large impact on practice, as actions contrary to it would be fiscally difficult given loss of insurance payments. Finally, the FDA could direct that systems with autonomous capabilities should be evaluated as medical interventions, not medical devices, which would help ensure that patients and doctors all recognize that appropriate evaluation standards are being used.

Conclusions

As current AI and robotic technologies unfold and permeate aspects of healthcare, the nature of the patient-doctor relationship and its foundational trust will be challenged, and likely changed. All of the typical "routes to trust" are potentially altered by the introduction of healthcare AI or robotic systems into the healthcare ecosystem. In order to ensure cohesive and effectual care based on the standards and values of both the patient and his doctor, the medical community and AI developers need to work together to establish expectations and standards that work within the democratic-care paradigm to help preserve trust between patients and their doctors. We propose that three high-level principles can guide this effort: education and licensure of medical professionals on AI systems by an external party; determined, educated consent given by the patient or caregiver prior to an AI's implementation in care; and providing alternate methods of care until AI is accepted as the "standard of care". In prioritizing these functions in regulatory measures, the industry and medical community will begin to ensure the societally and interpersonally proper deployment and implementation of such healthcare technologies.

Acknowledgements

This paper was made possible in part by a grant from Carnegie Corporation of New York. DD is the recipient of an Andrew Carnegie Fellowship. The statements made and views expressed are solely the responsibility of the authors.

References

[1] Deborah S. Ballard-Reisch. 1990. A model of participative decision making for physician-patient interaction. *Health Communication* 2(2), 91-104.

[2] Philippe Batifoulier, Jean-Paul Domin, and Maryse Gadreau. 2011. Market empowerment of the patient: The French experience. *Review of Social Economy* 69, 143-162.

[3] Stephen Buetow, Annemarie Jutel, and Karen Hoare. 2009. Shrinking social space in the doctor-modern patient relationship: A review of forces for, and implications of, homologisation. *Patient Education and Counseling* 74, 97-103.

[4] David Danks and Alex John London. 2017. Regulating autonomous systems: Beyond standards. *Intelligent Systems* 32(1), 88-91.

[5] R. Kaba and P. Sooriakumaran. 2007. The evolution of the doctor-patient relationship. *International Journal of Surgery* 5, 57-65.

[6] Jonathan Kimmelman and Alex John London. 2015. The structure of clinical translation: Efficiency, information, and ethics. *Hastings Center Report* 45(2), 27–39.

[7] Roy J. Lewicki, Edward C. Tomlinson, and Nicole Gillespie. 2006. Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management* 32, 991-1022.

[8] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37, 94-100.

[9] Enid N. H. Montague, Woodrow W. Winchester, and Brian M. Kleiner. 2010. Trust in medical technology by patients and health care providers in obstetric work systems. *Behaviour & Information Technology* 29, 541–554.

[10] Lee Ross, David Greene, and Pamela House. 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology* 13: 279-301.