

What Are Humans Doing in the Loop? Co-Reasoning and Practical Judgment When Using Machine Learning-Driven Decision Aids

Sabine Salloch & Andreas Eriksen

To cite this article: Sabine Salloch & Andreas Eriksen (20 May 2024): What Are Humans Doing in the Loop? Co-Reasoning and Practical Judgment When Using Machine Learning-Driven Decision Aids, The American Journal of Bioethics, DOI: [10.1080/15265161.2024.2353800](https://doi.org/10.1080/15265161.2024.2353800)

To link to this article: <https://doi.org/10.1080/15265161.2024.2353800>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 20 May 2024.



Submit your article to this journal [↗](#)



Article views: 329



View related articles [↗](#)



View Crossmark data [↗](#)

What Are Humans Doing in the Loop? Co-Reasoning and Practical Judgment When Using Machine Learning-Driven Decision Aids

Sabine Salloch^{a*} and Andreas Eriksen^{b*}

^aHannover Medical School; ^bOslo Metropolitan University

ABSTRACT

Within the ethical debate on Machine Learning-driven decision support systems (ML_CDSS), notions such as “human in the loop” or “meaningful human control” are often cited as being necessary for ethical legitimacy. In addition, ethical principles usually serve as the major point of reference in ethical guidance documents, stating that conflicts between principles need to be weighed and balanced against each other. Starting from a neo-Kantian viewpoint inspired by Onora O’Neill, this article makes a concrete suggestion of how to interpret the role of the “human in the loop” and to overcome the perspective of rivaling ethical principles in the evaluation of AI in health care. We argue that patients should be perceived as “fellow workers” and epistemic partners in the interpretation of ML_CDSS outputs. We further highlight that a meaningful process of integrating (rather than weighing and balancing) ethical principles is most appropriate in the evaluation of medical AI.

KEYWORDS

Clinical Decision Support Systems; Machine Learning; ethical principles; human in the loop



INTRODUCTION

Computerized Clinical Decision Support Systems (CDSS) have gained increased importance in health-care since their first use in the 1980s. According to a widely shared definition, CDSS provide “clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care” (Osheroff et al. 2007). Progress in data science and in the availability of clinical data has recently fueled a rapid development and the introduction of Machine Learning-driven CDSS (ML_CDSS) in various clinical workflows (London 2018). Radiology (Hwang et al. 2019), pathology (Kiani et al. 2020; Bulten et al. 2022) and ophthalmology (Gulshan et al. 2016; Ting et al. 2017) are among the pioneering disciplines in the introduction of ML_CDSS because they deal with large quantities of visual data. But also surgical disciplines (Standiford et al. 2022), dermatology (Esteva et al. 2017), pediatrics (Liang et al. 2019) and other clinical fields (Neugebauer et al. 2020; Jia et al. 2020) potentially profit from ML_CDSS.

In response to the rapid development of AI-based tools, there has been an increased concern with

regulatory frameworks and responsibility. The legal slogan is that there must always be a “human in the loop” when decisions affect important interests. This means that crucial decisions must be subject to human control instead of being processed fully automatically. Yet it remains fundamentally unclear what humans are supposed to be doing in the loop. Regulatory frameworks often seem to assume that simply adding a human will create the best of both worlds (i.e., machines will do what they do best and humans will do what they do best), ignoring the wealth of issues that call for new kinds of judgment and responsibility mechanisms (cf. Crootof, Kaminski, & Price, 2023). Humans can potentially serve a wide range of functions, but good execution requires clear expectations regarding the kinds of reasoning required in the role. This will include expectations of ethical judgment, but how can we get beyond abstract hortatory notions in this regard?

In the literature on ethical use of ML_CDSS, there is the demand for a “mental model of the range of relevant questions and ethical considerations that should guide design, evaluation, and implementation decisions” (Char et al. 2020). Usually, the models that guide ethical evaluation of ML_CDSS and other AI

CONTACT Sabine Salloch  salloch.sabine@mh-hannover.de  Institute of Ethics, History and Philosophy of Medicine, Hannover Medical School, Hannover Germany.

*Equally contributing authors.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

tools in health care use *principles* as major reference points. The spectrum of principle-based AI guidelines has become so rich that the “ethical tableau” of recommendations is already the subject of review articles. Jobin et al. (2019), for example, show in their comprehensive analysis of the global landscape of AI ethics guidelines that five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy) very much dominate the debate. The guidelines included in their analysis were not restricted to AI as applied in healthcare but aimed to provide ethical guidance on AI more generally. To an extent, however, the principles identified in this sample mirror principles traditionally used in the bioethical domain. Nevertheless, they also differ significantly, for example with respect to the importance of autonomy (Beauchamp and Childress 2019). According to Jobin et al.’s analysis autonomy—at least quantitatively—does not seem to have the same prominence in AI ethics guidelines as it traditionally has in bioethics frameworks. Consequently, there is a vivid and ongoing scientific debate on the selection of appropriate principles for the ethical evaluation of AI in healthcare. In particular, the status of explicability as an epistemic or normative guidance remains contested (Ursin et al. 2022; Adams 2023; Floridi and Cowsls 2022). This article will connect to the ongoing debates on principle-based approaches to AI ethics but take it in two novel directions.

First, it will not be occupied with the selection of an appropriate set of ethical principles nor with the differences between AI ethics and general medical ethics. Instead—and as a critical corrective to the currently ongoing debates on the ethics of AI in healthcare—this article takes one step back and considers the preconditions of responsible reasoning and judgment in further detail. While it has been common to think of principles for ethical AI as primarily pertinent to the design stage, we argue that the principles also apply to the conditions of responsible clinical reasoning. They do so partly in the sense that clinicians must implicitly or explicitly vouch for the tools they use and thereby the normative tradeoffs involved. But clinicians can also be active in shaping the tools themselves. Humans in the clinical stage should not simply be “in the loop” in the sense of mere recipients at a one-way conveyor belt. Rather, when supported by institutional processes they may continually form judgments that can *feed back* into the loop.

Second, whereas the debate on ML_CDSS to date has captured several important points of concern, there has been tendency to focus exclusively on the

physician aspect. Patients are often referred to in light of their individual values and preferences that need to be incorporated in AI-driven clinical decision-making (McDougall 2019). Beyond expressing their values and preferences, the role of patients is usually depicted as restricted to highly specific tasks, such as being “co-managers” of the clinical data that is used by ML_CDSS (Braun et al. 2021). In response, we aim to revive a more foundational account of co-reasoning and integrative judgment, where patients are more fully recognized as epistemic agents in their own right. This can facilitate a new and systematic appreciation of an ethically legitimate clinical deployment of AI in health care, but it can also contribute to the detection and diagnosis of “moral dangers” associated with such technologies.

THE ETHICS OF CDSS

In parallel to (or at least in the early aftermath of) the technological progress in automated clinical decision support, the debate on the promises and pitfalls of the development and clinical use of such systems is already in full swing. The potential pitfalls identified relate to, among other things, the context of patient safety, the fragmentation of clinical workflows, negative impact on user skills (“de-skilling” of physicians) or lack of interoperability (Sutton et al. 2020). Empirical research indicates that health care professionals’ perspectives on such systems vary but loss of professional autonomy and difficulties integrating AI in clinical workflows are the most dominant concerns (Lambert et al. 2023). From an ethical perspective, epistemic issues (e.g. misguided or inconclusive evidence) need to be considered together with normative aspects (e.g. unfair outcomes) and problems of traceability (e.g. explainability) of the CDSS functioning and concrete recommendations (Morley et al. 2020). Other debates refer to the impact of the use of CDSS on patient autonomy, for example, when debating “computer paternalism” (McDougall 2019) or discussing whether patients have a right to refuse ML-driven diagnostics and treatment (Ploug and Holm 2020a).

How far automated support truly contributes to its aim of enhancing the quality of health care depends greatly on the systems’ concrete use in practice and—more explicitly—on the integration of the automated recommendation into physicians’ clinical judgment. Various modes of interaction with ML_CDSS have been distinguished in the literature. According to Braun et al. (2021, 2), “conventional” systems deliver a “statement for consideration” to

the physician based on patient data as input. “Integrative” systems, by contrast, request and gather patient data autonomously, present it to the physician and document it in the electronic health records. As a last step, “fully automated” ML_CDSS not only supplement professionals’ decision-making but alter the decision authority. Fully automated ML_CDSS, thus, do not merely augment physicians’ decision-making but replace human reasoning, to a certain extent. How far a partial replacement of physicians might be realistic and desirable is subject to debate: some authors hold that ML_CDSS could replace human doctors in certain tasks, for example, in providing additional diagnostic services (Kempt and Nagel 2022). Some even suggest that in time machine intelligence will be sufficiently sophisticated regarding ethical reasoning and that it may “replace human ethical decision-making in certain settings” (Meier et al. 2022, 17, emphasis in original). Others are more skeptical even regarding ostensibly non-ethical issues and argue that the development and implementation of ML_CDSS should always target physician *support* rather than replacement (Taylor-Phillips and Freeman 2022).

The great interest in human-machine comparison is also mirrored in the theoretical and ethical debates on ML_CDSS that, for example, analyze situations of “peer disagreement”, where the automated suggestion deviates from physicians’ own clinical judgment. Grote and Berens (2020) point to the fact that the deployment of ML_CDSS might shift the epistemic authority and evidentiary norms for medical diagnosis. They argue that physicians’ normative justification for their clinical decision-making might become blurred against this background. Other authors convincingly argue that, from a methodological and research standpoint, studies on the concordance between physicians’ judgment and automated suggestions cannot provide evidence for the system’s positive effect on the quality of health care (Tupasela and Di Nucci 2020).

The relationship between “intelligent” CDSS and physicians is often depicted in a way that implicitly builds on a “competitive” picture of human-machine interaction. This competitive understanding is mirrored, for example, in comparative studies assessing the performance of human and computers in diagnostic and therapeutic tasks that have become vast and manifold in recent years. Such “outperforming studies”, for example, address the diagnostic precision (often: specificity and sensitivity) in the evaluation of images, for example, of skin lesions (Esteva et al. 2017; Brinker et al. 2019), electrocardiograms

(Hannun et al. 2019) or retinographies (Asiri et al. 2019). A systematic review and meta-analysis of 82 comparative studies with physicians and deep learning models in the assessment of medical images found the diagnostic performance of such models to be widely equivalent to that of health-care professionals (Liu et al. 2019). Even if methodological details of such research and its “ecological validity” demand further scrutiny, the mere fact of the multitude of such empirical studies is remarkable and sheds light on the prospective development of the field.

In legal analyses, by contrast, the competition between humans and machines is usually of less interest, but cooperative forms of interaction are explored. Concepts such as “meaningful human control” (Braun et al. 2021) or “effective human oversight” (Haselager et al. 2023) serve as key ideas—not only in the health-care domain—signifying that the control of the decisive steps lies with the human agent and is not fully left to machines. This does not necessarily mean a physical control of each step of the execution of tasks, but could also refer to the introduction of regulatory frameworks ensuring that the human agent can overrule and control the machine at important tipping points. The demand to have a human involved at the relevant steps can also be understood against the background of Article 22 of the EU General Data Protection Regulation stating that automated individual decision-making must not be performed based solely on automated processing in procedures that significantly affect the individual.

Appeals to the “human in the loop” as the final arbiter are intuitively plausible, especially in light of high-risk applications in health care. As noted, however, the “human factor” here remains thoroughly vague—particularly as comparative studies show that “intelligent” machines outperform physicians in certain isolated clinical tasks. Beyond this, it is unclear *which* humans need to be in control. A recent review notes that clinicians are most frequently mentioned, but that patients are also sometimes referred to with a particular emphasis on conditions for exercising informed consent (Hille et al. 2023). But perhaps being “in the loop” calls for more active involvement of patients as *co-reasoners* rather than mere consenters?

We believe that a closer look at the ethical frameworks that have been proposed for AI in healthcare reveals a need for reconceptualizing both the patient role and the meaning of moral clinical judgment. In the next section, we clarify the current mismatch between ethical frameworks and conceptions of moral reasoning.

PRINCIPLES FOR THE ETHICAL EVALUATION OF AI IN HEALTHCARE

The field of “ethical AI” is dominated by lists of principles that are relevant to the field. On the one hand, the pervasiveness of such lists is unsurprising. In line with the longstanding tradition of “principlism” and the dominance of principle-based accounts in bioethics, a natural step for “ethical AI” is articulate its own canonical lists tailored to its distinct problems. The *AI4People’s Ethical Framework*, for example, presents a synthesis of five ethical principles (Beneficence, Non-Maleficence, Autonomy, Justice, Explicability) that should guide AI development and adoption (Floridi et al. 2018). Explicability, in this account, is understood as a synthesis of (epistemological) intelligibility and (ethical) accountability and complements and enables the other four principles.

On the other hand, documents that promote principles have not been uncontroversial. For example, there is fear of “ethics washing” if private companies set up ethics codes and initiatives with limited practical impact which, however, might prevent policy makers from pursuing adequate legal regulation. Moreover, bioethicists have drawn attention to the fundamentally different character of the medical profession compared to AI development (Véliz 2019; Seger 2022). Brent Mittelstadt (2019) highlights four features of medicine that are lacking or appear in a different guise in AI development (common aim and fiduciary duties, professional history and norms, methods to translate principles into practice and legal and professional mechanisms of accountability). This divergence makes it unlikely that a principle-based approach will be satisfactory. As potential remedies Mittelstadt suggests, *inter alia*, to clearly define sustainable pathways to impact and to understand ethics rather as a process than as a technological issue.

While this institutional and procedural critique of principle-based accounts is important, it must be complemented by a more foundational conception of responsible *reasoning with principles*. While principle-based frameworks articulate *domains* where we need responsible decision-making, it is unclear how the various domains are supposed to relate to each other and form a coherent framework of reasoning. Typically, the frameworks speak of the need to “weigh” or “balance” principles against each other. The mentioned *AI4People’s Ethical Framework* highlights the task of weighing efficacy against control over decision-making (Floridi et al. 2018). And as we will discuss in more detail below, there is also sometimes talk of balancing accuracy against explainability when it comes to the clinical use of AI.

Naturally, the notions of “weight” and “balance” are metaphorical, shorthand for saying that reasons of a particular kind are especially salient in the relevant context. However, the metaphors do not say anything substantive about the kind of reasoning needed (Rawls 2020, 30; Richardson 2000). Standard theoretical accounts claim that balancing is a matter of “practical wisdom” and an associated set of virtues such as “practical astuteness, discriminating intelligence, and sympathetic responsiveness” (Beauchamp and Childress 2019, 22). In short, there is a need for good judgment. But pointing out this does not in and of itself provide guidance. Hence, the clinical setting is provided with a batch of complex ethical tasks, but not adequate conceptual tools for discharging them. Is there anything substantive yet sufficiently general to be said about reasoning and applying the principles of ethical AI and thereby ensuring responsible use of ML_CDSS?

In the upcoming sections, we argue that the two shortcomings identified have a common source. Both the lack of clarity about what humans are doing “in the loop” and the lack of substantive guidance on how to reason with principles are in part due to the absence of an account of how principles can be integrated through co-reasoning. Our response is not to provide a blueprint for meaningful control and a recipe for weighing principles. Instead, we offer a conceptual reorientation toward new ideals.

As already indicated, our approach is committed to what is sometimes called the “collaborative” approach to AI ethics in medicine, which emphasizes that AI principles need to be operationalized through mutual engagement between clinicians and designers. According to this approach, frameworks for ethical AI are not just for designers but also for clinicians. Hence, a core idea is that clinicians need to reason with AI principles as well as with the traditional bioethical principles. Often, this will take the form of mutual explanation of needs, potentials and problems: “Medical doctors could communicate to designers what levels of accuracy are needed for specific tasks and the tradeoffs between principles and standards in real-time decision-making, for instance between accuracy and urgency in emergency situations” (Gundersen & Bærøe, 2022, p. 11). However, as we will argue below, the collaborative model may work best when patients are also included in the loop of co-reasoning across domains of expertise.

A Neo-Kantian Framework

Although principlism and similar approaches are often described as a “mid-level theories” that are consistent with different foundational starting points, there is a

tendency for underlying theoretical commitments to affect first-order judgment. In this article, we build on the work of Onora O'Neill. She has of course been an influential and direct contributor to many topics of bioethics. However, our approach is to draw on her more foundational work on reasoning and judgment. In particular, our argument is that her exposition of the Kantian ideas of reasoners as “*fellow workers*” and of judgment as a matter of *integration* can shed light on the challenges of ethically responsible use of ML_CDSS.

It is worth noting that O'Neill's (1989, 82) work is neo-Kantian in the sense that her ambition “goes far beyond a concern with reading Kant accurately”. It is primarily concerned with expounding valid normative perspectives. We could perhaps say the same of our reading of O'Neill: we are not primarily concerned with interpreting what she would say about the case at hand herself, but rather with the *conceptual tools* her work has delivered for dealing with ethical AI. In particular, by reviving and conciliating two separate strands of her work—co-reasoning and integration—we aim to show that her neo-Kantian framework delivers important ways to think about ethical clinical reasoning with ML_CDSS.

While our approach takes the form of applying ideas from O'Neill to the case of ML_CDSS, we should add that the resulting framework may have much broader application. A neo-Kantian perspective on “humans in the loop” regarding this technology can shed light on more general ethical concerns regarding respectful medical interaction. Current debates on, for example, the informational tasks of regulatory bodies (e.g., Svirsky, Howard, & Berman, 2022) presuppose some notion of how clinicians should integrate external advice into the clinical context. Similarly, debates on “patient expertise” (e.g., Watson, 2024) presuppose a conception how doctors should interact with patients as co-reasoners. While a neo-Kantian framework does not deliver any ready-made solutions for such debates, it nevertheless highlights modes of reasoning that apply to them. In other words, the idea of operationalizing principles for ethical use of ML_CDSS through the concepts of co-reasoning and integration can be considered an illustration of a conceptual framework that has broader ramifications.

A) Co-Reasoning: The Tasks of “Fellow Workers”

AI guidelines for the medical context are seldom addressed directly to patients as active epistemic subjects, they mostly concern the tasks of clinicians. In a rare exception, however, the American Medical

Association's (2023) statement on “Augmented Intelligence in Health Care” encourages “education for patients, physicians, medical students, other health care professionals, and health administrators to promote greater understanding of the promise and limitations of health care AI.” There are two ways of interpreting this. One is that patients are included on this list merely as a way of becoming more aware of their own preferences and opportunities. This is arguably the sense intended by the drafters. A more ambitious interpretation, however, is that patients are to be included as epistemic participants in the quest for ethical AI. We will argue that the second interpretation fits with the challenges posed by ML_CDSS, even though this may diverge from the conceptions that underlie current normative frameworks.

By way of establishing this, the first strand of O'Neill's (1989) work that we draw on is her notion of a “political” conception of reason, which connects reasoning to specific attitudes toward other epistemic subjects. It is political in the sense that it is likened by Kant to a process of obtaining agreement among “free citizens” (Kant 1998, 643). O'Neill frames the political reading by paying attention to how Kant's imagery circles around metaphors like “lawgiver” without “dictatorial authority,” reason as a reflexive “tribunal,” and not least social and egalitarian metaphors like “fellow workers.”

Against this background, she notes that “the Kantian vindication of reason presupposes plurality-without-preestablished-harmony” (O'Neill 2015, 13). In this picture, reasoning is about openness to contestation and acceptance of valid claims. It is the opposite of one-sided power, where claims are backed by force rather than reasons: “Only those who try to think from the standpoint of everyone else and strive to listen to and interpret others and to see the point of their contributions are genuinely aiming to be ‘fellow workers’ and to avoid maxims to which others cannot agree” (O'Neill 1989, 26).

The idea that reasoning depends on attitudes toward “fellow workers” finds resonance in more recent literature that highlights the positive epistemic role of patients. For example, the “partnership model” sees patients as having a constructive role to play in supporting the reasoning of physicians. Patients can help physicians avoid narrow or simplistic reasoning by being attentive and asking questions (i.e., acting as co-reasoners). As physician and writer Jerome Groopman describes it: “a few pertinent and focused questions [can] protect me from the cascade of cognitive pitfalls that cause misguided care” (Groopman 2008, 268).

Importantly, this ideal goes beyond conceptions of Shared Decision-Making. Shared Decision-Making is an approach of physicians and patients making decisions together while using the best available evidence (Elwyn et al. 2010). This approach emphasizes “empowering” patients (Emanuel and Emanuel 1992, 2222) or enabling them to communicate informed “preferences” regarding screening, management, or treatment options (Elwyn et al. 2010, 971). The notion of “empowering” suggests an epistemic asymmetry where the clinician’s task is to promote patients to become sufficiently responsible agents. Similarly, the notion of communicating “preferences” lacks connotations of patients as sources of valid cognitive claims. Naturally, empowerment of informed preference communication is an important task with regards to clinical decision making. And we are not wishing away inevitable asymmetries in knowledge. Nevertheless, we believe the rise of ML_CDSS requires a revival of the partnership model at a more foundational level. Patients are not only to be empowered by clinicians, they should also empower them in return.

Importantly, the status of patients as epistemic partners does not necessarily rely on them bringing in significant medical expertise. Rather, their epistemic potential lies in playing a discursive role, asking questions and testing their own understanding. The push to articulate reasons can make reliance on ML_CDSS more reflective and responsible. The epistemic mechanism at work here is familiar from accountability studies, where it has been shown that agents who know they have to provide reasons—but do not know the exact kinds of reasons—engage in more reflective and balanced reasoning (Bovens and Schillemans 2014, 678-679).

How does this translate to the clinical setting and the use of ML_CDSS? A clear example is the prevalence of “automation bias.” It has been documented that clinicians often substitute their own judgment with CDSS recommendations—even in cases where their original judgment is superior (Goddard et al. 2012). Interventions that have either increased user accountability or provided doctors with information about automation bias have had little success in counteracting this (Lyell & Coiera, 2017 p. 430). Experiments where clinicians experience failures of automated systems (e.g., misdiagnosis) lead to some reduction of automation bias, but this tends to be confined to the specific kind of medical recommendation in question (Bahner, et al, 2008, p. 697). By contrast, bringing patients in as “fellow workers” introduces a social and dialogical dimension that potentially exerts a different and more dynamic kind

of pressure. By encouraging patients to take active part in a process of reasoning, doctors need to articulate reasons to support their reliance on automated decisions. This may not be realistic in certain contexts, given that doctors can lack access to sufficient information through the interface. But such issues are receiving more attention, for example with suggestions that interfaces should support independent verification by providing relevant data side-by-side with decision support (Lyell & Coiera, 2017, p. 430). It is unlikely that the mere presence of such information will sufficiently counteract automation bias, given the lack of success with other informational interventions. But combining this interface feature with an institutionalized expectation that patients should be actively involved in the reasoning process would provide an important additional incentive for doctors to consult their own professional judgment and to articulate an independent justification.

A second type of example concerns the role of patients as co-reasoners regarding the role and potential of ML_CDSS. Consider Rosalind McDougall’s notion of AI systems as “discussion prompts”: “Well-designed AI systems could be used as a tool to prompt doctors and patients to discuss treatment goals and articulate the patient’s values relevant to the decision at hand” (Zhou 2019, 158) The key concept here is “well-designed,” because it points to an area where clinicians and patients almost literally need to become “fellow workers.” Effective contestability (as explicability might be described from a patient’s perspective, cf. Ploug & Holm 2020b) clearly contributes to enhancing the patient’s opportunities for critically assessing an AI system’s quality and usefulness for their own clinical situation. It can be anticipated, however, that such high demands toward information on, e.g., use of data, potential biases and system performance might not be consistently available in a form that is accessible to each individual patient.

For the systems to be clinically meaningful and effective, both clinicians and patients need to provide feedback to designers. This might especially hold for AI that is applied in the immediate clinical encounter but likewise in disciplines such as radiology or pathology where we see some chance that feedback is provided by both physicians and patients, e.g. about their ways of dealing with inconclusive or uncertain diagnostic outcomes. The notion that the clinical context should stand in a collaborative relationship with design of AI tools reflects much actual practice, as both academic institutions and industry rely on feedback from medical expertise (Gundersen & Børøe, 2022, p. 10).

However, in many cases, the notion of medical expertise that informs this collaborative process should be rooted in co-reasoning between clinician and patient. That is, neither clinicians nor patients can assess whether a tool is well-designed without some *joint understanding* of what the clinical process needs. Consider this quote from a study that shows practitioners' reflections when user-testing a tool for Shared Decision-Making that is linked to evidence summaries: "Clinician: I am not quite convinced that 'uncertainty' is a concept that patients can grasp or that the way it is presented in the tool is all that helpful" (Heen et al. 2021, 8). Clearly, this is a kind of statement that would benefit from engagement with patients as epistemic resources in their own right. The clinician claims that patients cannot grasp the concept of "uncertainty," which is a factual statement that can be tested through discussion with the relevant patient group.

The founding principle at work here is not anti-paternalism or empowerment of patients as such (although it speaks to these concerns). Rather, it is the neo-Kantian idea of *reasoning* as an inherently political or social enterprise. As noted, the notion of a "fellow worker" is an epistemic ideal—with moral implications—that involves striving "to listen to and interpret others and to see the point of their contributions" (O'Neill 1989, 26). It is not primarily a matter of registering desires or preferences, but of being *responsive to valid considerations* concerning what the situation demands.

Although we have noted that this does not rely on patients bringing in significant medical expertise, it is worth highlighting that the framework also lends support to recent accounts that emphasize seeking out patient "testimony" as a counterweight to AI-generated recommendations (McCadden et al. 2023; Slack and Barclay 2023). A key claim in these accounts is that patients are knowers in a medically relevant sense and that doctors commit a kind of "epistemic injustice" (Fricker 2007) when they fail to respect patients' status as knowers. The notion of testimony refers to claims to epistemic validity as opposed mere preferences, as per the partnership model's ambition of treating patients as "fellow workers" in an epistemic sense. In line with this, the idea of patients as co-reasoners helps illuminate the dangers of epistemic injustice that are potentially triggered by ML_CDSS. Clinicians faced with a decision of believing either their patient or automated outputs provide paradigm illustrations. Prediction Drug Monitoring Programmes (PDMP), for example, predict a patient's risk of misusing opioids but come

along with the risk that patients are misjudged and denied helpful (e.g. analgesic) treatment. PDMPs might relativize patients' status as knowers and exacerbate preexisting social inequalities (Pozzi 2023). Similarly, in psychiatry there is potential for certain AI tools to be used in ways where inferences from digital data undermine patients' status as knowers about their psychological states in a clinically relevant sense (McCadden et al. 2023; Slack and Barclay 2023). Considering the patient as a "fellow worker" in the interpretation of computerized outputs might prevent clinicians from unjustly undermining the patient's credibility due to diverging suggestions from automated decision-support.

So far, we have argued that ML_CDSS triggers reasons for rejuvenating a partnership model of clinical reasoning, because patients can play genuine epistemic roles in avoiding dangers such as automation bias and in providing feedback on how systems should be designed. However, we now want to argue a further claim: patients should play a role in exercising the practical judgment that enacts the principles of ethical AI.

B) Integration: Connecting the Principles of Ethical AI

The sets of ethical principles that have been developed for ethical AI in guidelines do not come as a puzzle where all the pieces fit immediately and intuitively together. Instead, they contain tensions and the need for continued evaluative judgment. Even if some may have an incentive to make them appear as a plain and transparent structure, Mittelstadt (2019) argues that we should "hesitate to celebrate consensus around high-level principles that hide deep political and normative disagreement." This resonates with reviews of AI guidelines that highlight pervasive "uncertainty as to which ethical principles should be prioritized and how conflicts between ethical principles should be resolved" (Jobin et al. 2019, 396). However, the uncertainty does not simply concern how to resolve *particular* tensions, it is a more general uncertainty regarding *how to reason* about conflicts. Talk of "balancing" or the need for "practical wisdom" are ways of delineating the need for judgment, but not ways of supplying substantive standards.

Consider again the claim that principles such as efficiency or accuracy may "outweigh" considerations of human control in the domain of AI-based tools (e.g., London 2019; AI4People 2020, 18-19). By what measure do we decide that efficiency is more

important than control? As O'Neill puts it: "There is no metric for balancing or trading-off different types of norms" (O'Neill, 2018, p. 84). Her alternative neo-Kantian conception of judgment emphasizes the task of integration: "Practical judgment is an aspect of practical reasoning because it aims to integrate rather than to prioritize or trade off a plurality of norms" (O'Neill 2018, 84).

The need for an account of judgment that highlights integration is in part due to the nature of principles. When they appear in ethical codes or other statements of professional associations, principles take the form of abstract and general considerations. While rules usually instruct some specific kind of action, principles appeal to considerations that should be part of complex assessments. Typically, this makes talk of principles being "outweighed" or "defeated" exaggerated or misplaced. The standard case is one of agents having to satisfy a multitude of principles at once. Maximizing one principle and downgrading others is likely to be blameworthy in the medical context: "Great success in seeking patients' informed consent does not compensate for providing them with sub-standard care" (O'Neill 2018, 190–191). Of course, an urgent operation on an unconscious patient makes the principle of informed consent less salient than principles of beneficence, but in such cases the work of judgment is not really a matter of "balancing" but of recognizing the *limits* of informed consent. It is not as if informed consent suddenly loses normative significance qua principle, but rather that the consequences and concerns that the principle protects are not present in the standard way.

The already noted debate on explainability versus accuracy provides a clear example of why this is important with regards to ML_CDSS. It is often argued that tradeoffs need to be made between explainability and clinical validity or—in general—the performance of a ML_CDSS (Amann et al. 2020). Moreover, some empirical research is framed as demonstrating that patients "value explainability of AI systems in healthcare less than in non-healthcare domains and less than often assumed by professionals, especially when weighed against system accuracy" (van der Veer et al. 2021). However, this framing, and the ethical conclusions that downgrade explainability, presuppose that it is a static notion, having a fixed meaning in all practical domains (e.g., giving an account of causal factors identified in diagnostic claims).

By contrast, O'Neill's (1989, 229) suggestion is that we treat principles as more dynamic and context-dependent phenomena: "[T]he exact demands of

justice must vary with circumstances. For example, what constitutes coercion will depend on the vulnerability of those who would be victimized. Activity that might be normal bargaining or negotiating procedure in interaction with an equal may coerce the vulnerable." Arguably, we could say the same about explainability (cf. Zhou & Danks, 2020). What constitutes explainability will depend not only on the system properties of ML_CDSS, but also on the *addressee* of the explanation. It may count as a substantial explanation for a patient to learn, for example, that the diagnosis was aided by system that detects patterns in data, that data is classified by experts before being fed into the system, or that certain features are weighted by designers to avoid false negatives or positives. Such rudimentary features can be enough to establish trust, which is the success condition of explanations in this context. In other words, although patients neither want nor need a detailed causal account of how the AI derived its conclusions, that does not mean explainability loses its importance. The normative effect is a change of meaning rather than weight. Note, however, that such explanations are not *devoid* of causal accounts, although they refer to mechanisms that are external to the first-order diagnostic assessment of data.

But how can we get this account to say something substantive beyond merely pointing to the context-dependent nature of integration? Integration, as an alternative to balancing, is at its core a matter of using principles as mutually supporting elements of judgment. That is, the operative meaning of principles is discovered by interpreting them *in light of each other*. For example, the demands of explainability can help define the demands of accuracy, and further principles like efficiency or informed consent. This process of specifying principles by seeing them as mutually constraining is nicely illustrated by O'Neill (2018, 23): "Just as equations can often be solved only when we know a sufficient number of constraints, so questions about how to act are often resolved only by taking account of a number of constraints." However, should this process of determining constraints also be a process of co-reasoning?

Yes, there is no way of identifying the relevant constraints without listening to and interpreting the standpoints of patients. Hence, it is helpful to bring the two strands of O'Neill's neo-Kantian approach together. Sound integration of principles is not exclusively a matter of individual clinical judgment, but rather a joint process of establishing the demands of the situation. Again, co-reasoning is not primarily about the patient being empowered to express

reasoned preferences or desires. Rather, qua fellow workers, patients are in epistemic partnership with the doctor, which entails a joint commitment toward practicing good judgment. With regards to principles such as explainability, patients provide constructive input concerning the appropriate ethical meaning and its place in a broader network of principles.

This is not to imply that the process of integration always results in a complete resolution of value conflicts. Doctors, patients, managers or other stakeholders may have views that resist unification, for example if the potential for efficiency and reliability with the help of ML_CDSS is deemed by doctors and managers to warrant forms of decision-making that are opaque to patients and thereby constitute a genuine cost to patient empowerment. Although such dilemmas may have a resolution from an abstract vantage point, real-life decisions come with tradeoffs that are subject reasonable disagreement. While we have simplified the discussion by framing it as a matter of individual doctors reasoning with individual patients, responsible case-specific reasoning should track reason-giving procedures at higher levels that provide legitimacy to value tradeoffs. In this regard, our approach is in line with versions of AI deployment ethics that emphasize having representative bodies at the institutional level that ensure both voice and quality: “decisions about whether and how to deploy advanced technology in medical institutions should be the outcome of a deliberative process among diverse stakeholders that include patient groups and/or their advocates, healthcare workers, and administrators” (Palmer & Schwan, 2024, p. 126). These representative processes cannot fully preempt the need for judgment in concrete cases, but they can provide ways of reasoning that are anchored in good-faith attempts to treat medicine as an enterprise of different kinds of “fellow workers.”

Objections

As a way of further clarifying our framework, we will attempt to respond to some anticipated objections. An obvious objection is that this is too ambitious. Can we really expect clinicians and patients to act together as co-reasoners, not least as there are number of patients whose capacity to self-determination is limited for various reasons? We have two responses to this. First, our approach above draws on a range of empirical examples where the co-reasoning approach is at least incipiently at work or where failures are related to inadequate processes of co-reasoning. Hence, we see this framework as a normative reconstruction of

principles at work in AI-based medical practice rather than a top-down imperative based on ideal theory. Second, like the other principles, the principle of co-reasoning should be interpreted in a way that can be integrated with a range of further concerns, such as efficiency, accuracy, and more. In other words, the demandingness objection is already anticipated by the holistic methodological approach.

A second objection is that our framework is too vague. What does co-reasoning actually demand? Although we have referred to a range of examples that indicate what this implies more concretely, we want to acknowledge that the framework is deliberately vague in certain regards. It is vague with regards to how this should be implemented institutionally because this will depend on context-sensitive issues such as the stage of clinical consultation (e.g., screening or treatment options), nature of treatment (e.g., preparatory surgery, medication program), the patient group affected (e.g., children, cognitively incapacitated), and so on. Some contexts may call for generic checklists for clinical consultations, other contexts require that the demands co-reasoning influence the design of ML_CDSS, yet other contexts may need further mechanisms such as inter-professional procedures or patient educational programs.

A third objections is that we are placing too much responsibility on the patient. We believe this is an especially important objection to address because it prompts us to clarify how this is a framework for epistemic practices rather than individual liabilities. Promoting the patient to co-reasoner as opposed to mere source of preferences does not imply adding to patients' liability for process or outcome. When we speak of “partnership” or “co-ownership” of process, the concern is to create epistemically sound and non-alienating processes, it is not about establishing patient accountability. As Gary Watson (1996) has expounded with great clarity, being addressed as *answerable* is distinct from being held *accountable*. The former is to be addressed as a person capable of responding to reasons and reflecting on claims and commitments, the latter is about being the appropriate subject of sanctions or other consequences when decisions fail to meet established standards. The framework of co-reasoning about issues involving ML_CDSS does not aim to add further to patient liabilities or to further pulverize accountability in this field, but rather to identify opportunities for epistemically and morally responsible reasoning and judgment. As already noted in response to the vagueness objection, we leave it open how these opportunities should be realized institutionally.

SUMMARY AND CONCLUSIONS

The ongoing ethical debates on ML_CDSS are essentially shaped by the implicit meta-ethical presupposition that decision-making needs to be guided by a plurality of ethical principles that are weighed against each other in judging concrete cases. Little effort has been exerted so far in using established philosophical approaches to moral judgment to better understand the appropriate use of ethical principles for evaluating AI-driven applications in health care. A reference to neo-Kantian accounts of ethical judgment serves for providing both a fuller estimation of the interrelatedness of ethical principles on the abstract level and a better conception of patient-physician co-reasoning in the concrete clinical encounter.

In particular, our approach highlights that—beyond patient empowerment and Shared Decision-Making—patients can serve as co-reasoners and help physicians in asking the right questions and adequately interpreting ML_CDSS outputs. Seeing patients as “fellow workers” might also prevent physicians from walking into some traps associated with automated decision-support such as automation bias or epistemic injustice toward marginalized groups. With respect to ethical principles, a neo-Kantian approach in line with O’Neill clarifies that diverging ethical principles must not be “silenced,” “defeated,” or “outweighed” in concrete situations but adjusted to the circumstances and interpreted in light of each other.

Further work needs to be directed toward the concrete interpretation of “competing” ethical principles in light of AI applications in health care. Strengthening the position that ethical principles do not work as “algorithmic decision procedures” but need interpretation by the genuinely human faculty of judgment also helps us to understand what the demand for a “human in the loop” actually refers to in moral terms. The concept of moral co-reasoning in the immediate patient-physician-encounter might help to avoid dangers that arise from automated recommendations undermining physicians’ professional judgment.

ACKNOWLEDGMENTS

The authors would like to thank four anonymous reviewers and the Expertise, Politics and Ethics group at the Center for the Study of Professions (Oslo Metropolitan University) for valuable comments.

DISCLOSURE STATEMENT

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

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

REFERENCES

- Adams, J. 2023. Defending explicability as a principle for the ethics of artificial intelligence in medicine. *Medicine, Health Care, and Philosophy* 26 (4):615–23. doi:10.1007/s11019-023-10175-7.
- AI4People 2020. AI4People’s 7 AI Global Frameworks. Accessed May 04, 2023 <https://ai4people.eu/wp-content/pdf/AI4People7AIGlobalFrameworks.pdf>.
- Amann, J., A. Blasimme, E. Vayena, D. Frey, and V. I. Madai. 2020. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making* 20 (1):310–6. doi:10.1186/s12911-020-01332-6.
- American Medical Association 2023. Augmented intelligence in medicine. Accessed September 19, 2023 <https://www.ama-assn.org/practice-management/digital/augmented-intelligence-medicine>.
- Asiri, N., M. Hussain, F. Al Adel, and N. Alzaidi. 2019. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artificial Intelligence in Medicine* 99:101701. doi:10.1016/j.artmed.2019.07.009.
- Bahner, J. E., A. D. Hüper, and D. Manzey. 2008. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies* 66 (9):688–99. doi:10.1016/j.ijhcs.2008.06.001.
- Beauchamp, T. L., and J. F. Childress. 2019. *Principles of Biomedical Ethics*. 8. ed. Oxford: Oxford University Press.
- Braun, M., P. Hummel, S. Beck, and P. Dabrock. 2021. Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics* 47 (12):e3–e3. doi:10.1136/medethics-2019-105860.
- Brinker, T. J., A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schandendorf, T. Holland-Letz, et al. 2019. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer* 113:47–54. doi:10.1016/j.ejca.2019.04.001.
- Bulten, W., K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. van Boven, R. Vink, , et al. 2022. Artificial intelligence for diagnosis and gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine* 28 (1):154–63. doi:10.1038/s41591-021-01620-2.
- Char, D., M. Abramoff, and C. Feudtner. 2020. A framework to evaluate ethical considerations with ML-HCA applications—valuable, even necessary, but never comprehensive. *The American Journal of Bioethics* 20 (11):W6–W10. doi:10.1080/15265161.2020.1827695.
- Crootof, R., M. E. Kaminski, and N. W. Price. II. 2023. Humans in the Loop. *Vanderbilt Law Review* 76:429–510.

- Elwyn, G., S. Laitner, A. Coulter, E. Walker, P. Watson, and R. Thomson. 2010. Implementing shared decision making in the NHS. *BMJ (Clinical Research ed.)* 341 (oct14 2):c5146–c5146. doi:10.1136/bmj.c5146.
- Emanuel, E. J., and L. L. Emanuel. 1992. Four models of the physician-patient relationship. *Jama* 267 (16):2221–6. doi:10.1001/jama.1992.03480160079038.
- Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542 (7639):115–8. doi:10.1038/nature21056.
- Floridi, L., and J. Cowls. 2022. A unified framework of five principles for AI in society. In *Machine learning and the city: Applications in architecture and urban design*, ed. S. Carta, 535–45.
- Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al. 2018. AI4People—an ethical framework for a good AI society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28 (4):689–707. doi:10.1007/s11023-018-9482-5.
- Fricker, M. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.
- Goddard, K., A. Roudsari, and J. C. Wyatt. 2012. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19 (1):121–7. doi:10.1136/amiajnl-2011-000089.
- Groopman, J. 2008. *How Doctors Think*. Boston: Mariner Books.
- Grote, T., and P. Berens. 2020. On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics* 46 (3):205–11. doi:10.1136/medethics-2019-105586.
- Gulshan, V., L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316 (22):2402–10. doi:10.1001/jama.2016.17216.
- Gundersen, T., and K. K. Bærøe. 2022. The future ethics of artificial intelligence in medicine: Making sense of collaborative models. *Science and Engineering Ethics* 28 (2):17. doi:10.1007/s11948-022-00369-2.
- Hannun, A. Y., P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* 25 (1):65–9. doi:10.1038/s41591-018-0268-3.
- Haselager, P., H. Schraffenberger, S. Thill, S. Fischer, P. Lanillos, S. van de Groes, and M. van Hooff. 2023. Reflection machines: Supporting effective human oversight over medical decision support systems. *Cambridge Quarterly of Healthcare Ethics: The International Journal of Healthcare Ethics Committees* 10:1–10. doi:10.1017/S0963180122000718.
- Heen, A. F., P. O. Vandvik, L. Brandt, F. Achille, G. H. Guyatt, E. A. Akl, S. Treweek, and T. Agoritsas. 2021. Decision aids linked to evidence summaries and clinical practice guidelines: Results from user-testing in clinical encounters. *BMC Medical Informatics and Decision Making* 21 (1):202. doi:10.1186/s12911-021-01541-7.
- Hille, E. M., P. Hummel, and M. Braun. 2023. Meaningful human control over AI for health? A review. *Journal of Medical Ethics* :jme-2023-109095. doi:10.1136/jme-2023-109095.
- Hwang, E. J., J. G. Nam, W. H. Lim, S. J. Park, Y. S. Jeong, J. H. Kang, E. K. Hong, T. M. Kim, J. M. Goo, S. Park, et al. 2019. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology* 293 (3):573–80. doi:10.1148/radiol.2019191225.
- Jia, P., P. Jia, J. Chen, P. Zhao, and M. Zhang. 2020. The effects of clinical decision support systems on insulin use: A systematic review. *Journal of Evaluation in Clinical Practice* 26 (4):1292–301. doi:10.1111/jep.13291.
- Jobin, A., M. Ienca, and E. Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1 (9):389–99. doi:10.1038/s42256-019-0088-2.
- Kant, I. 1998. *Critique of Pure Reason*, translated and edited by Paul Guyer and Allen W. Wood. Cambridge: Cambridge University Press.
- Kempt, H., and S. K. Nagel. 2022. Responsibility, second opinions and peer-disagreement: Ethical and epistemological challenges of using AI in clinical diagnostic contexts. *Journal of Medical Ethics* 48 (4):222–9. doi:10.1136/medethics-2021-107440.
- Kiani, A., B. Uyumazturk, P. Rajpurkar, A. Wang, R. Gao, E. Jones, Y. Yu, C. P. Langlotz, R. L. Ball, T. J. Montine, et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digital Medicine* 3 (1):23. doi:10.1038/s41746-020-0232-8.
- Lambert, S. I., M. Madi, S. Sopka, A. Lenés, H. Stange, C. Buszello, and A. Stephan. 2023. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. *NPJ Digital Medicine* 6 (1):111. doi:10.1038/s41746-023-00852-5.
- Liang, H., B. Y. Tsui, H. Ni, C. C. S. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen, et al. 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine* 25 (3):433–8. doi:10.1038/s41591-018-0335-9.
- Liu, X., L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, et al. 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet. Digital Health* 1 (6):e271–e297. doi:10.1016/S2589-7500(19)30123-2.
- London, A. J. 2018. Groundhog day for medical artificial intelligence. *The Hastings Center Report* 48 (3):inside back cover. doi:10.1002/hast.842.
- London, A. J. 2019. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *The Hastings Center Report* 49 (1):15–21. doi:10.1002/hast.973.
- Lyell, D., and E. Coiera. 2017. Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association* 24 (2):423–31. doi:10.1093/jamia/ocw105.
- McCadden, M., K. Hui, and D. Z. Buchman. 2023. Evidence, ethics and the promise of artificial intelligence in psychiatry. *Journal of Medical Ethics* 49 (8):573–9. doi:10.1136/jme-2022-108447.

- McDougall, R. J. 2019. Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics* 45 (3):156–60. doi:10.1136/medethics-2018-105118.
- Meier, L. J., A. Hein, K. Diepold, and A. Buyx. 2022. Algorithms for ethical decision-making in the clinic: A proof of concept. *The American Journal of Bioethics: The American Journal of Bioethics* 22 (7):4–20. doi:10.1080/15265161.2022.2040647.
- Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1 (11):501–7. doi:10.1038/s42256-019-0114-4.
- Morley, J., C. C. Machado, C. Burr, J. Cows, I. Joshi, M. Taddeo, and L. Floridi. 2020. The ethics of AI in health care: A mapping review. *Social Science & Medicine* (1982) 260:113172. doi:10.1016/j.socscimed.2020.113172.
- Neugebauer, M., M. Ebert, and R. Vogelmann. 2020. A clinical decision support system improves antibiotic therapy for upper urinary tract infection in a randomized single-blinded study. *BMC Health Services Research* 20 (1):185. doi:10.1186/s12913-020-5045-6.
- O'Neill, O. 1989. *Constructions of reason: Explorations of Kant's practical philosophy*. Cambridge: Cambridge University Press.
- O'Neill, O. 2015. *Constructing authorities: Reason, politics and interpretation in Kant's Philosophy*. Cambridge: Cambridge University Press.
- O'Neill, O. 2018. *From principles to practice: Normativity and judgement in ethics and politics*. Cambridge: Cambridge University Press.
- Osheroff, J. A., J. M. Teich, B. Middleton, E. B. Steen, A. Wright, and D. E. Detmer. 2007. A roadmap for national action on clinical decision support. *Journal of the American Medical Informatics Association: JAMIA* 14 (2):141–5. doi:10.1197/jamia.M2334.
- Palmer, A., and D. Schwan. 2024. More process, less principles: The ethics of deploying AI and robotics in medicine. *Cambridge Quarterly of Healthcare Ethics: The International Journal of Healthcare Ethics Committees* 33 (1):121–34. doi:10.1017/S0963180123000087.
- Ploug, T., and S. Holm. 2020b. The four dimensions of contestable AI diagnostics - A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine* 107:(101901. doi:10.1016/j.artmed.2020.101901.
- Ploug, T., and S. Holm. 2020a. The right to refuse diagnostics and treatment planning by artificial intelligence. *Medicine, Health Care, and Philosophy* 23 (1):107–14. doi:10.1007/s11019-019-09912-8.
- Pozzi, G. 2023. Testimonial injustice in medical machine learning. *Journal of Medical Ethics* 49 (8):551–2. doi:10.1136/jme-2022-108630.
- Rawls, J. 2020. *A Theory of Justice* : Revised Edition. Harvard: Harvard University Press.
- Richardson, H. S. 2000. Specifying, balancing, and interpreting bioethical principles. *The Journal of Medicine and Philosophy* 25 (3):285–307. doi:10.1076/0360-5310(200006)25:3;1-H;FT285.
- Seger, E. 2022. In defence of principlism in AI ethics and governance. *Philosophy & Technology* 35 (2):45. doi:10.1007/s13347-022-00538-y.
- Slack, S. K., and L. Barclay. 2023. First-person disavowals of digital phenotyping and epistemic injustice in psychiatry. *Medicine, Health Care and Philosophy* 26 (4):605–14. doi:10.1007/s11019-023-10174-8.
- Standiford, T. C., J. L. Farlow, M. J. Brenner, M. L. Conte, and J. E. Terrell. 2022. Clinical decision support systems in otolaryngology–head and neck surgery: A state of the art review. *Otolaryngology-Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery* 166 (1):35–47. doi:10.1177/01945998211004529.
- Sutton, R. T., D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. 2020. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine* 3 (1):17. doi:10.1038/s41746-020-0221-y.
- Svirsky, L., D. Howard, and M. L. Berman. 2022. E-cigarettes and the multiple responsibilities of the FDA. *The American Journal of Bioethics: The American Journal of Bioethics* 22 (10):5–14. doi:10.1080/15265161.2021.1907478.
- Taylor-Phillips, S., and K. Freeman. 2022. Artificial intelligence to complement rather than replace radiologists in breast screening. *The Lancet. Digital Health* 4 (7):e478–e479. doi:10.1016/S2589-7500(22)00094-2.
- Ting, D. S. W., C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, et al. 2017. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multi-ethnic populations with diabetes. *Jama* 318 (22):2211–23. doi:10.1001/jama.2017.18152.
- Tupasela, A., and E. Di Nucci. 2020. Concordance as evidence in the Watson for Oncology decision-support system. *AI & SOCIETY* 35 (4):811–8. doi:10.1007/s00146-020-00945-9.
- Ursin, F., C. Timmermann, and F. Steger. 2022. Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics* 36 (2):143–53. doi:10.1111/bioe.12918.
- van der Veer, S. N., L. Riste, S. Cheraghi-Sohi, D. L. Phipps, M. P. Tully, K. Bozentko, S. Atwood, A. Hubbard, C. Wiper, M. Oswald, et al. 2021. Trading off accuracy and explainability in AI decision-making: Findings from 2 citizens' juries. *Journal of the American Medical Informatics Association: JAMIA* 28 (10):2128–38. doi:10.1093/jamia/ocab127.
- Véliz, C. 2019. Three things digital ethics can learn from medical ethics. *Nature Electronics* 2 (8):316–8. doi:10.1038/s41928-019-0294-2.
- Watson, G., 1996. Two faces of responsibility. *Philosophical Topics* 24 (2):227–48. doi:10.5840/philtopics199624222.
- Watson, J. C. 2024. Patient expertise and medical authority: Epistemic implications for the provider–patient relationship. *The Journal of Medicine and Philosophy* 49 (1):58–71. doi:10.1093/jmp/jhad045.
- Zhou, Y., and D. Danks. 2020. Different "intelligibility" for different folks. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 194–199. Accessed May 10, 2024. <https://dl.acm.org/doi/abs/10.1145/3375627.3375810>