Medical AI, Categories of Value Conflict, and Conflict Bypasses

Gavin Victor, Jean-Christophe Bélisle-Pipon
Simon Fraser University
gavin_victor@sfu.ca, jean-christophe_belisle-pipon@sfu.ca

Abstract

It is becoming clear that, in the process of aligning AI with human values, one glaring ethical problem is that of *value conflict*. It is not obvious what we should do when two compelling values (such as autonomy and safety) come into conflict with one another in the design or implementation of a medical AI technology. This paper shares findings from a scoping review at the intersection of three concepts—AI, moral value, and health—that have to do with value conflict and arbitration. The paper looks at some important and unique cases of value conflict, and then describes three possible categories of value conflict: *personal* value conflict, *interpersonal* or *intercommunal* value conflict, and *definitional* value conflict. It then describes three general paths forward in addressing value conflict: additional ethical theory, additional empirical evidence, and bypassing the conflict altogether. Finally, it reflects on the efficacy of these three paths forward as ways of addressing the three categories of value conflict, and motions toward what is needed for better approaching value conflicts in medical AI.

1 Introduction

As artificial intelligence (AI) becomes a new frontier in health technology development, ethicists are hard at work developing approaches that outline the necessary actions to be taken for the responsible integration of AI in healthcare. One approach to ethical analysis and design of AI centers technology innovation on human values (Aldewereld and Mioch 2021, Smits et al. 2022, Umbrello and van de Poel 2021). Taking human values as signifiers of what needs to be considered ethically, value-based approaches to AI ethics are premised on the idea that an AI technology more aligned with contextually relevant human values is a more ethical technology. Such an approach requires value identification, arbitration, and translation into guidelines and instructions for AI development and implementation.

Adding to the quickly growing efforts to align AI with values, we set out to offer an inventory of values and value-based discussions evoked in AI design for healthcare. Doing so helps to build a foundation for better value alignment of medical AI. This inventory also helps to identify important discussions occurring in a wide range of medical sub-domains, as well as identify not only the core set of values discussed, but also emerging and underdiscussed values. Such a map aims to unearth and bring to light a broad range of challenges and pathways for the value alignment of medical AI.

The present paper centers on a particular subset of findings based on a systematic scoping review of moral values in the medical AI context (Victor, Salem, and Bélisle-Pipon 2023, Victor, Barbu, and Bélisle-Pipon 2024). The scoping review is meant to survey the diverse range of values, approaches, challenges, and opportunities present in moral value centered medical AI design literature. Alongside providing a map of values at play in this domain, the review also surveyed *procedural* components of value-grounding medical AI; namely, it looked at processes such as identification of values, substantiation of values, and arbitration between values. This paper presents findings regarding this last topic: value arbitration. In cases of value conflict, when values such as explainability and accuracy are at odds, the question of what to do is not yet settled. This paper offers a characterization of distinct kinds of value conflict and ways of responding to them.

After describing our methodology in section 2, we move on to results. We first describe three examples of value conflict, to make the challenge more palpable and to serve as examples for the remainder of the paper (subsection 3.1). We then describe three general categories of value conflict found in sample papers: *personal* value conflict, *interpersonal* or *intercommunal* value conflict, and *definitional* value conflict (subsection 3.2). We then turn to prescriptive arguments found in the review regarding pathways for arbitration given that authors find that pre-established guidelines are not always useful for conflict arbitration (subsection 3.3). Scholars advocate for adding *ethical theory*, adding *empirical research,* or *bypassing* value conflicts altogether. In the discussion section, we note benefits and drawbacks to these three pathways in addressing each of the three categories of value conflict and note the directions in empirical research that this discussion suggests (section 4).

2 Methodology

A protocol for the scoping review was published to OSF (Victor, Salem, and Bélisle-Pipon 2023). The research question was: "What are the best practices, challenges, opportunities, and recommendations for identifying, arbitrating, and substantiating the values at the core of responsible and trustworthy AI development in the health sector?" The present paper follows a prior analysis of the sample in terms of specific values to create an inventory of values (which is presented in line with the PRISMA-ScR checklist) (Victor, Barbu, and Bélisle-Pipon 2024, Tricco et al. 2018).

2.1 Search Strategy

We searched five databases (Embase, Medline, IEEE, Web of Science, and Scopus) for papers that contained three necessary concepts. The first is a "health" concept, for which our search terms included ones like "medical" and "clinical." The second was an "AI" or "Artificial Intelligence" concept. The third was a "moral/ ethical value" concept, using search terms like "value sensitive design," to capture values (in the relevant sense) discussed in this domain. A precise description of the search strategy is written in our pre-print relaying findings regarding specific moral values (Victor, Barbu, and Bélisle-Pipon 2024).

2.2 Data Extraction

Database searches were conducted in March 2023. With duplicates excluded, a total of 2568 studies were screened in Covidence. 2338 were deemed irrelevant in the screening stage, and 227 papers were assessed in full for eligibility. 115 studies were ultimately deemed to meet eligibility requirements.

Papers were then imported to NVivo (14), and a coding strategy was developed and tested in consecutive meetings with three members of the team. Two reviewers coded half of the papers each. Our coding sections were broken into themes including discussions of specific values and discussions of conflict arbitration. In coding, we inductively gathered cases of described conflict between two values, as well as when papers made arguments concerning challenges, opportunities, best practices, frameworks, and recommendations regarding value arbitration in cases of conflict.

3 Results

We found 44 value conflicts discussed in the sample. We found that 27 articles discussed the more abstract task of arbitration between values in cases of conflict. After three examples of value conflict found in articles included in the review, we articulate categories of value conflict and ways of responding to value conflict illustrated by discussions in the sample.

3.1 Notable Value Conflicts

Conflict One: Accuracy and Transparency

A pervasive conflict in AI ethics is that between *accuracy* and *transparency* (in any number of its forms: explainability, explicability, understandability, interpretability, and more—generally, we take transparency to signify making the algorithmic processes accessible and/or understandable to stakeholders). Our sample reflects this—it was the conflict we found most represented, being discussed by 16 papers included in the review. At its most general, this conflict captures the fact that often, as an algorithm becomes more *transparent*, it also becomes less *accurate* (Hatherley et al. 2022, McCoy et al. 2022). This is because algorithms tend to need to be of reduced parameter complexity if they are to be accessible to a human mind (Bezemer et al. 2019). This loss in complexity tends to correlate, scholars argue, with losses in accuracy (i.e., higher likelihoods of false judgments).

For example, take an algorithm that makes judgments based on the input of radiological images. To make such an algorithm transparent and explainable, a feature may be designed such that the algorithm is asked to produce a "saliency map" that shows what areas of the image contributed to the judgment (Bousquet and Beltramin 2022). Doing so, a clinician can thereby get some understanding of the algorithm's decision. The problem is that, by adding such a feature that asks the algorithm to "think" in human-like ways (or at least *explain* its thinking in human-like ways), we risk increasing the likelihood of false positives and negatives in judgments the algorithm makes.

The value we ought to prioritize in this conflict isn't obvious. Scholars argue that we should not monolithically value *accuracy* as the sole epistemic value, because opacity breeds its own epistemic concerns (Afnan et al. 2021). Defending desires for accuracy over transparency, some argue that predominant framings of this conflict misconstrue values of transparency, or at least their history. In the words of a Sendak et al. (2020) contributor, "The human body is a black box." Sendak et al. (2020) take this quotation to capture the way in which calls for transparent AI may fail to appreciate that the lack of transparency is a historically accepted norm in medicine.

Conflict Two: Efficacy and Privacy

Value conflicts can also present themselves in cases where the *efficacy* (how effective it can be in benefiting health outcomes) of a particular intervention or the desired outcome comes into conflict with *privacy*. Buruk, Ekmekci, and Arda (2020) offer a succinct description of the origin of the problem: algorithms are only as good as their data are extensive. Bak et al. (2022) articulate a similar point, noting

that "we sometimes seem to forget that data access is the most important prerequisite for any AI innovation." At the same time, medical algorithms are trained on sensitive health data. So, medical AI feeds on data that individuals may be reluctant to share.

As an example, consider "care robots" that can "monitor the health conditions of patients, give medication, manually lift and aid the movements of disabled patients, and provide social companionship" (Yew 2021). The desire for a robot to be efficacious in, for instance, monitoring a healthcare condition might lead to advocating for saving and logging data gathered in relation to the patient. But this also means that the patient is always being watched, listened to, and recorded. The more that the robot does this, the less private the life of the individual appears.

Conflict Three: Autonomy and Safety

As a final case here, there can be a conflict between patient *autonomy* and *safety* in the use of medical AI, especially for individuals who are at risk for dangerous behavior. This conflict captures the tension between desires for freedom and the inherent risk that goes along with increased freedom.

For example, Burmeister (2016) looks at a technology that helps individuals with dementia to navigate through a city. Such a technology, for Burmeister (2016), embodies the conflict between *safety* and *autonomy*. Allowing the individual more autonomy in their navigation can mean that they may become lost or disoriented, and like Burmeister writes, walk "in the midst of [a] four lane road." But if we prioritize *safety*, then the individual will sacrifice some level of autonomy as they are guided step-by-step to their destination.

Peters et al. (2020) note that in cases of extreme mental illness, respect for autonomy may directly oppose the safety of the patient or others, if the patient is at risk for behaving harmfully. They take this to indicate that sometimes respect for autonomy should be overridden by values that are more important in a given situation.

These three value conflicts serve as examples of those that are being debated and arbitrated in creating medical AI technologies. In this paper, they will serve as touchstones for examples, and will make more concrete the relatively abstract project of value arbitration.

3.2 Three Categories of Value Conflict

Values can conflict in a number of distinct ways. What follows is a list of three different general categories of value conflict, though this list is likely not exhaustive of all possible categories of value conflicts.

Personal Value Conflicts

First, there can be a conflict in values for any one given stakeholder. In these cases, a technology cannot respect two important values held by a particular stakeholder, likely due to technological constraints. Think here of *conflict three*, in which safety and autonomy conflict in navigation assistance for those with dementia. Just about any person will care about *both* autonomy and safety. In some cases, we may only be able to be safe if we sacrifice autonomy, or only be fully autonomous if we sacrifice safety. In this sense, then, a user's own personal values can come into conflict in how a given technology is able to actualize them. For the values of any one stakeholder, it can be the case that a technology can only maximize one at the cost of the other.

Interpersonal and Intercommunal Value Conflicts

There are also important concerns regarding differences in value preferences *between* and *among* stakeholders. Consider how this can surface in a case like *conflict one*: If a radiologist is invested in the *transparency* of an algorithm that makes judgments about images, and their patient in question cares deeply about accuracy (and if accuracy and transparency come into conflict in the technology at hand), then we have an *interpersonal* conflict between different stakeholders. Or in *intercommunal* conflict, one group of stakeholders desires a value whose actualization would preclude the actualization of a value held by another group. We can imagine that radiologists may, as a group, express a strong desire for transparency to review and evaluate for themselves the judgment that the algorithm makes (Ursin et al. 2022). Moreover, we can imagine that patients may, as a group, care more deeply for the accuracy of the algorithm, giving them a better chance at the best possible diagnosis and treatment. (Though, this may be an oversimplication. See e.g. Zhang et al. 2021).

Naturally, stakeholders in different positions in relation to a technology are going to have differing hierarchies of values. But this is not the only way that values can conflict interpersonally and between communities. In addition, stakeholders in the same position may *just have differing values*. Consider *conflict two*, in which a socially assistive robot is more *efficacious* the less it respects *privacy*. People's desires for *privacy*, we should expect, are not all the same. So, even between two *users* of a socially assistive robot that inhabit the same stakeholder group, values may differ (and conflict) *interpersonally*.

Definitional Value Conflicts

A third form of value conflict in which some arbitration is necessary is when different individuals have different understandings of some given value (Jacobs 2020). Jacobs (2020) uses the example of the value of health, which one stakeholder group may define as "ability to realize one's vital goals" and another as "absence of disease" (Jacobs 2020). Similarly, what a value means, how it is important, and how trade-offs are made may be relative to a domain, as König et al. (2022) assume to be possible between predictive policing and skin cancer prediction. The point here is that how a value is defined, and what gives it its normative weight, may vary not only between the medical field and

policing, but perhaps between medical sub-domains like gerontology and radiology—as well as from person to person, even within a given domain. Thus, there arise cases where we must arbitrate between different notions of a given value.

These are some of the ways that value conflicts can arise, and in which conflicts need to be assessed and arbitrated.

3.3 Pathways for Addressing Value Conflicts

Unsurprisingly, a number of authors consider already established frameworks or approaches designed for arbitration. The AIHLEG's seven key requirements, for example, might offer a path forward in value arbitration (Bak et al. 2022, Borsci et al. 2022, High-Level Expert Group on Artificial Intelligence 2019). But, Bak et al. (2022) argue that the guidance given by the AIHLEG does not take a broad enough perspective to account for value conflicts. Other, more specific guidelines may be more helpful, according to some authors. But, even in response to domain-specific guidelines, authors also claim that guidance can be further developed to aid in resolving value conflicts (e.g., Burmeister 2016). Beyond guidelines published thus far, frameworks and approaches can be developed to more productively arbitrate value conflicts, and papers in our sample make strides in this direction.

Supplemental Ethical Theory

Jacobs (2020) develops an approach termed "Capability Sensitive Design" (CSD) in order to address a number of challenges (including addressing conflicts and resolving trade-offs). Drawing upon the value-alignment methodology "Value Sensitive Design" (VSD) (Friedman and Hendry, 2019) and Martha Nussbaum's Capability approach, CSD provides "substantive normative foundation" for formulating and defending moral claims in cases like value arbitration (Jacobs 2020). Roughly, such a view draws attention to a person's capabilities of value—potential ways of being that are preferable—to make ethical arguments about value arbitration. It gives ethical support for valuing autonomy over safety, for example, if autonomy maximizes the user's capabilities of value. This addition gives us the theoretical material needed for arbitration and does so by adding onto the VSD framework. So, one way of developing approaches to dealing with value conflicts is by tacking on an ethical theory to which we can defer in cases of value conflict arbitration. Though Jacobs' (2020) paper defends this capability theory specifically, her argument leaves open the possibility of other ethical theories to provide a similar grounding for value arbitration.

Supplemental Empirical Grounding

Much like Jacobs (2020) supplements VSD with "capability theory," Burmeister (2016) considers not a *theoretical* addition to VSD in order to arbitrate values, but additional *empirical* modes of investigation into values. This would be to conduct empirical investigations, a core component of the VSD framework (Friedman and Hendry, 2019) so as to directly investigate value conflicts. "Inter-rater reliability," using experts to find a consensus in a conflict, though promising, has the problem of not involving the diverse range of stakeholders implicated in a trade-off (Burmeister 2016). So, Burmeister gives more attention to the "weighted-point evaluation method," a systematic framework for evaluating decisions and alternatives. He argues that this framework is amenable to multi-stakeholder group participation, and so better offers empirical grounding for value-arbitration decisions. So, for example, enabled by the opinions and values of stakeholders, designers can determine when and how much to sacrifice efficacy for the sake of privacy in a socially assistive robot.

Somewhat similarly, van der Veer et al. (2021) offer a method of arbitrating between accuracy and explainability via a "citizen's jury." A method originally developed as part of the "Deliberative Democracy" approach to policymaking, the citizen's jury engages a group of stakeholder jurors in an education and deliberation process to come to a conclusion about some important topic. Van der Veer et al.'s (2021) paper investigates the trade-off between *accuracy* and *explainability* in different contexts (including healthcare). Such an approach offers a method that gathers pertinent empirical evidence about how well-informed and engaged individuals come to conclusions about value conflicts.

Bypassing (Apparent) Value Conflict

Two general approaches have therefore come forward from views in our sample: one ethical-theoretical and one empirical, that can help us perform arbitration in cases of value conflict. These approaches make headway in answering questions of how to approach value conflicts and/or which value to favor when we are forced to make a trade-off. But, some authors describe pathways by which we can bypass the need to choose one value over another in the first place.

With mind to conflicts that go along with "black box" models, Sendak et al. (2020) argue that the values that underly pushes for interpretable and explainable AI can be respected in a given technology without making the AI interpretable or explainable. That is, for example, if we want accountability, we may view the explainability of a technology as *the* way to grant accountability. But then we are failing to realize how we can attain accountability via procedures, practices, and requirements independent from explainability (Sendak et al. 2020). The opportunity Sendak et al. (2020) offer here is to take a step back from cases of value conflict and understand that, when value conflicts implicate an instrumental value that we desire for the sake of some deeper value(s) (i.e., transparency for the sake of accountability), we can avoid the value conflict by focusing on alternative ways of respecting core values. In such cases,

we can circumnavigate or bypass the conflict and need not accept a trade-off between the core values at play.

Another way of avoiding conflicts is by designing *value flexibility* into algorithms and medical AI use cases (McDougall 2019). According to McDougall (2019), a value flexible system "acknowledges value diversity and attempts to accommodate differences among individuals." So, with such a system, interpersonal value differences don't need to be arbitrated because the algorithm can consider and respond to unique user values. To return to the example of navigation guidance for those with dementia, designing the product to have different settings for those who value (or need) safety over autonomy or vice versa offers a solution by turning what may be a conflict into different options. It also allows for a change in values embodied in the technology as a person's condition and values shift.

4 Discussion

As argued above, approaches to value arbitration can be divided into kinds that add *ethical theory* to existing frameworks (such as Value Sensitive Design), kinds that add *empirical investigations* for value arbitration, and kinds that *bypass* value conflicts.

Now, we will consider these three ways of responding to (apparent) value conflict in light of the three categories of value conflict (*personal, interpersonal/intercommunal, and definitional*). We begin with a discussion of the two *additive* approaches to value conflict, adding ethical theory and adding empirical investigation. Then, we discuss categories of (apparent) value conflict and conflict bypassing.

4.1 Supplementary Approaches and Conflict Categories

Those approaches that *add ethical theory* also add the grounds for making normative claims about values (i.e. we *should* support the capability of autonomous decision-making), but risk failing to represent the values of some or most stakeholders if their personal preference differs from the conclusion that follows from the ethical theory. For example, we might not want to accept a deontology-informed solution to a trade-off (valuing, perhaps, autonomy over safety) if choosing autonomy over safety in such a case is radically unintuitive.

Moreover, it is unclear how convincing relying on an ethical theory will be in cases of *personal* value conflict. Most likely, the solution to the conflict being convincing will be a function of the theory's convincingness itself. If I am not persuaded by the theory that one relies on for addressing a conflict, then the pathway illuminated by the theory is unlikely to be convincing. The same goes for *interpersonal* and *intercommunal* conflicts—we cannot expect all to accept the solution just because it is backed by theoretical rationale.

Because reaching a solution through *increased empirical investigation* is only accurate insofar as it is representative of its stakeholders, it shouldn't have the problem of failing to capture the views of those impacted. It does, though, still have the often-noted problem of deriving normative judgments from empirical claims: "taking stakeholder values as leading values in the design process without questioning whether what *is* valued by stakeholders also *ought* to be valued" (Jacobs 2020). One approach is therefore vulnerable to criticisms about its ability to accurately represent stakeholders, and the other is vulnerable to criticisms about its ability to make grounded normative claims.

Increased empirical investigation appears especially useful in the case of *definitional* value conflict, as these investigations will help to examine what exactly is at the core of values at play in the scenario. Knowing the diverse range of ways in which people conceptualize privacy in a given context allows for more clarity about where disagreements lie. In turn, having clarity about disagreements enables them to be better accounted for in the design of a particular technology.

There is room for further discourse on these two paths for building approaches, and though they may differ, there is no reason why the approaches should be mutually exclusive; we don't necessarily need to make a trade-off between these two pathways (See Freedman et al. (2020) for an example of a similar kind of hybrid approach).

However frameworks are developed to address value conflicts, it is important that they not take all value conflicts to be of one form. *Interpersonal* value conflicts are likely the predominate form, but what it takes to arbitrate between people is different from what it takes to arbitrate *personal* value conflicts. *Personal* value conflicts have to do with personal values hierarchies and willingness to sacrifice certain values for others, whereas *interpersonal* or *intercommunal* value conflict requires balancing the values of certain stakeholders or populations against others. Given this, frameworks should take into account the diverse ways in which value conflicts can surface and provide guidance in approaching each. This means accounting for the particularities of not only the three categories of conflict alluded to here, but also any others that aren't captured here.

4.2 Conflict Bypassing and Conflict Categories

Authors have also given ways of circumnavigating value conflicts via cutting out values that are *instrumental* to some other end, or via implementing *value flexibility*. But the efficacy of these options is a technical question to do with particular values in the case of the former, and particular algorithms in the case of the latter. On the one hand, we need

answers to questions such as: *Which* values are instrumental to some end, and which are irreducible? In which contexts are certain values reducible vs. not? On the other hand, we need answers to questions like: What kinds of algorithms can implement value flexibility, and how? If it turns out that value flexibility is relatively easy to build into many algorithms, then our strategy in approaching value conflicts can differ greatly compared to if we know we must make sacrifices in the form of trade-offs. In this way, greater research into the promise of avoiding value conflicts by way of reducing instrumental values to irreducible values and building value-flexible algorithms will enable better bypassing methodologies in facing value conflicts.

The *value flexibility* of an algorithm, though, isn't necessarily helpful in approaching conflicts of each of the three categories described above. It thrives in the case of *intercommunal* value conflict, where one group cares about autonomy and another group cares about safety. In such a case, a technology that can adapt to the differences in desires for autonomy and safety—providing minimal or robust guardrails—to mitigate the conflict.

Engineering *value flexibility* into a technology may also be useful in cases of *definitional* value conflict, if the technology can have different mechanics as a function of different notions of a value. For example, someone may desire "privacy" from a socially assistive robot in the sense that it doesn't enter her bedroom, or in the sense that it doesn't save any information gathered by a camera or microphone. Corporeal privacy and informational privacy can be two distinct flexible parameters under the larger concept of privacy. If so, the algorithm can adapt to different notions of a broader value.

Value flexibility, though*,* isn't useful in cases where technological constraints preclude sufficient actualization of two or more *personal* values held by a given individual, like if a stakeholder values both transparency and accuracy, and both cannot be captured sufficiently in a given technology. That is, *value flexibility* is only an option in cases where parameters can vary as to adequately respect both values implicated in a conflict.

Reducing instrumental values to their end-values, similarly, is only useful in some cases. It is only useful in cases where a conflict implicates a value that is merely instrumental, and not itself an end value. This means that in cases where the conflict is between core, irreducible values, this strategy is no help. In apt *personal value conflict* situations where a given stakeholder has two values that come into conflict in the technology, the individual is likely to be satisfied with sufficient actualization of the values to which one of the conflicting values is a means.

Reducing instrumental values can be useful in addressing *interpersonal* value conflicts if it is the case that a given subset of stakeholders value one of the implicated values only for the sake of other values that aren't in conflict. This group would then be satisfied by the instrumental value being avoided but the values to which it is a means being actualized.

In cases of *definitional* value conflict, moreover, if one notion of the value is instrumental and reducible, then stakeholders would likely accept their definition of the value being flouted for the sake of the values that they care about at their core. The reduction strategy, then, applies to a number of cases of seeming value conflict; but, these value conflicts are still of a particular form. They concern reducible instrumental values. This strategy is not useful in cases that don't concern values that are desired for other, independently attainable values.

The *reduction* strategy therefore provokes the question: which values *are* instrumental, to what ends, and in what cases? We should not assume that any value that is instrumental in one case is instrumental in every other case. This is to say, if we care about transparency for the sake of clarifying accountability, and if we are satisfied with no transparency given that accountability is ensured via other means, this does not mean that there aren't cases where we care about transparency for its own sake. Further, we should not assume that a value that is instrumental to some other set of values in one context is instrumental to the same set of values in another context. Likely, what transparency is a means to in one context varies in comparison to what it is a means to in another context. In one case it may be that we want transparency as a means to clarify who is accountable for a judgment, but in another case, we may want algorithmic transparency simply to give patients the dignity of understanding the system that is impacting them. This signifies that context-specific research about what values are held instrumentally and to which ends is key for the efficacy of addressing apparent value conflicts through *reduction*.

5 Conclusion

This paper has outlined three broad categories of value conflict: *personal conflicts* in which a technology cannot actualize two values held by a given individual, *interpersonal* or *intercommunal conflicts* in which values held by different stakeholders or stakeholder communities come into conflict, and *definitional conflicts* in which stakeholders or contexts disagree on what is meant by a given value. Guidelines and frameworks are still developing guidance in addressing such conflicts. We pointed to three general ways of developing solutions to value conflicts: adding *ethical theory*, adding *empirical research*, and by bypassing the conflict altogether in one of two ways—first, by *reducing instrumental values* implicated in conflicts (such as transparency) to their end values (such as accountability), and second, by building

value flexibility into the algorithms and their scenarios of use. In light of these categories of value conflict and ways of responding to them, we have made the case for a number of future developments in research for more efficacious solutions to value conflicts including research into how empirical and theoretical methods can approach different kinds of value conflict, research into which kinds of algorithms can be value-flexible (with regard to which values), and research into which kinds of values are instrumental, to which ends.

Acknowledgments

We thank Andreea Barbu, Guido Calderini, Antoine Boudreau Leblanc, Sherif Salem, Jimin Rhim, Marie-Françoise Malo, Vardit Ravitsky, and many other members of the Ethics Pillars from both "Cell Maps for AI" (CM4AI) and "Bridge2AI-Voice Consortium" for help in the process of the scoping review. This work is supported through the Bridge2AI program, NIH Grant Number: 1OT2OD032742-01.

References

Afnan, M..; Liu, Y.; Conitzer, V.; Rudin, C.; Mishra, A.; Savulescu, J.; and Afnan, M. 2021. Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Human Reproduction Open*, *2021*(4) hoab040. doi.org/10.1093/hropen/hoab040

Aldewereld, H., and Mioch, T. 2021. Values in Design Methodologies for AI. In Advanced Information Systems Engineering Workshops. CAiSE 2021. Lecture Notes in Business Information Processing, vol 423. Springer, Cham. doi.org/10.1007/978-3-030-79022-6_12

Bak, M.; Madai, V. I.; Fritzsche, M.-C.; Mayrhofer, M. T.; and McLennan, S. 2022. You Can't Have AI Both Ways: Balancing Health Data Privacy and Access Fairly. *Frontiers in Genetics*, *13*. doi.org/10.3389/fgene.2022.929453

Bezemer, T.; de Groot, M. C.; Blasse, E.; Ten Berg, M. J.; Kappen, T. H.; Bredenoord, A. L.; van Solinge, W. W.; Hoefer, I. E.; and Haitjema, S. 2019. A Human(e) Factor in Clinical Decision Support Systems. *Journal of Medical Internet Research*, *21*(3) e11732. dx.doi.org/10.2196/11732

Borsci, S.; Lehtola, V.; Nex, F.; Yang, M.; Augustijn, E.; Bagheriye, L.; Brune, C.; Kounadi, O.; Li, J.; Moreira, J.; Van der Nagel, J.; Veldkamp, B.; Le, D.; Wang, M.; Wijnhoven, F.; Wolterink, J.; and Zurita-Milla, R. 2022. Embedding artificial intelligence in society: Looking beyond the EU AI master plan using the culture cycle. *AI & SOCIETY*. doi.org/10.1007/s00146-021-01383-x

Bousquet, C., and Beltramin, D. 2022. Machine Learning in Medicine: To Explain, or Not to Explain, That Is the Question. *Studies in Health Technology and Informatics*, *294*: 114–115. doi.org/10.3233/SHTI220407

Burmeister, O. 2016. The development of assistive dementia technology that accounts for the values of those affected by its use. *ETHICS AND INFORMATION TECHNOLOGY 18*(3): 185–198. doi.org/10.1007/s10676-016-9404-2

Buruk, B.; Ekmekci, P. E.; and Arda, B. 2020. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy 23*(3): 387–399. doi.org/10.1007/s11019-020-09948-1

Freedman, R.; Borg, J.; Sinnott-Armstrong, W.; Dickerson, J.; and Conitzer, V. 2020. Adapting a kidney exchange algorithm to align with human values. *ARTIFICIAL INTELLIGENCE*, *283*. doi.org/10.1016/j.artint.2020.103261

Friedman, B., and Hendry, D. G. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. 10.1007/s00146-021-01383-x. Cambridge and London: MIT Press.

Hatherley, J.; Sparrow, R.; and Howard, M. 2022. The Virtues of Interpretable Medical Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees*, 1–10. dx.doi.org/10.1017/S0963180122000305

High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines For Trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed: 2024-06-14.

Jacobs, N. 2020. Capability Sensitive Design for Health and Wellbeing Technologies. *SCIENCE AND ENGINEERING ETHICS 26*(6): 3363–3391. doi.org/10.1007/s11948-020-00275-5

König, P. D.; Felfeli, J.; Achtziger, A.; and Wenzelburger, G. 2022. The importance of effectiveness versus transparency and stakeholder involvement in citizens' perception of public sector algorithms. *Public Management Review 0*(0): 1–22. doi.org/10.1080/14719037.2022.2144938

McCoy, L. G.; Brenna, C. T. A.; Chen, S. S.; Vold, K.; and Das, S. 2022. Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology* (*142*): 252–257. doi.org/10.1016/j.jclinepi.2021.11.001

McDougall, R. J. 2019. Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics 45*(3): 156–160. doi.org/10.1136/medethics-2018-105118

Peters, D.; Vold, K.; Robinson, D.; and Calvo, R. A. 2020. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society 1*(1): 34–47. doi.org/10.1109/TTS.2020.2974991

Sendak, M.; Elish, M.; Gao, M.; Futoma, J.; Ratliff, W.; Nichols, M.; Bedoya, A.; Balu, S.; and O'Brien, C. 2020. *"The Human Body is a Black Box":* Supporting Clinical Decision-Making with Deep Learning. 99–109. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*: 99-109. New York: Association for Computing Machinery. doi.org/10.1145/3351095.3372827.

Smits, M; Nacar, M.; Ludden G. D.S.; and van Goor, H. 2022. Stepwise Design and Evaluation of a Values-

Oriented Ambient Intelligence Healthcare Monitoring Platform. *VALUE IN HEALTH 25*(6): 914–923. hdoi.org/10.1016/j.jval.2021.11.1372

Tricco, A. C.; Lillie, E.; Zarin, W.; O'Brien, K. K.; Colquhoun, H., Levac, D.' Moher, D.' Peters, M. D. J.; Horsley, T.; Weeks, L.' Hempel, S.' Akl, E. A.; Chang, C., McGowan, J.; Stewart, L.; Hartling, L.; Aldcroft, A.; Wilson, M. G.; Garritty, C.; Lewin, S.; Godfrey, C. M.; Macdonald, M. T.; Langlois, E. V.; Soares-Weiser, K.; Moriarty, J.; Clifford, T.; Tunçalp, Ö.; and Straus, S. E. 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of internal medicine 169*(7): 467–473. doi.org/10.7326/M18-0850

Umbrello, S., and van de Poel, I. 2021. Mapping value sensitive design onto AI for social good principles. *AI and Ethics 1*(3): 283–296. doi.org/10.1007/s43681-021-00038-3

Ursin, F.; Timmermann, C.; and Steger, F. 2022. Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics*, 36: 143–153. doi.org/10.1111/bioe.12918

van der Veer, S. N.; Riste, L.; Cheraghi-Sohi, S.; Phipps, D. L.; Tully, M. P.; Bozentko, K.; Atwood, S.; Hubbard, A.; Wiper, C.; Oswald, M.; and Peek, N. 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): 2128–2138. doi.org/10.1093/jamia/ocab127

Victor, G.; Barbu, A.; and Bélisle-Pipon, J-C. 2024. Moral Values in Medical AI: A Scoping Review. Research Square preprint. doi.org/10.21203/rs.3.rs-4391239/v1

Victor, G.; Salem, S.; and Bélisle-Pipon, J-C. 2023. A Scoping Review of Relevant Moral Values in Health Sector AI Development. *Open Science Foundation*. 10.17605/OSF.IO/BCVK3

Yew, G. C. K. 2021. Trust in and Ethical Design of Carebots: The Case for Ethics of Care. *International Journal of Social Robotics 13*(4): 629–645. doi.org/10.1007/s12369-020-00653-w

Zhang Z.; Citardi D.; Wang D.; Genc Y.; Shan J.; and Fan X. 2021. Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data. *Health Informatics Journal* 27(2). doi.org/10.1177/14604582211011215