Addressing the Complexity in Applying Ethical Principles to Artificial Intelligence within Healthcare and Data Governance

(Draft version. Work in Progress. Do not cite without permission.)

A version of this draft was presented at the DICB AIM AHEAD Workshop at Hilton Garden Inn, Orange Beach, Alabama, July 19-21, 2024.

Dr. Brett Coppenger, Tuskegee University
Dr. Sam Taylor, Tuskegee University

Abstract

This paper attempts to make salient some of the complexities in the application of ethical principles to artificial intelligence (AI) by using W.D. Ross's pluralistic ethics as inspiration for analyzing AI's role in healthcare. While AI introduces new scenarios, the general ethical principles and the complexities that arise during their application are not entirely novel. Ross's presentation of a pluralist framework can provide a robust foundation for addressing these challenges. His seven ethical duties—beneficence, self-improvement, nonmaleficence, fidelity, gratitude, reparation, and justice—are employed to assess moral obligations in AI-driven healthcare decisions. However, equally important is the way Ross's discussion of a moral decision procedure can aid in navigating conflicting duties in specific cases. The application of this framework is illustrated through examples involving AI-assisted clinical decisions, highlighting the complexities of physician autonomy, AI transparency, and accountability. The study underscores the necessity of thoughtful ethical analysis and emphasizes the potential for reasonable disagreement in moral judgments, advocating for a Ross inspired framework as a valuable tool for contemporary ethical dilemmas in AI healthcare integration.

1. Introduction

As philosophers it is an exciting (and sometimes worrying) time to be alive. Advances in artificial intelligence have made *real* many of the different types of situations that we were only able to *imagine* before.

However, it is our contention that while many of the specific situations that have arisen (or will soon arise) due to advances in artificial intelligence are new, these very same situations are the kinds of things that philosophers have already been thinking about in the abstract. And, importantly, many of the views that philosophers have developed have the resources needed to address these new situations. In essence our claim is that there is no need to reinvent the wheel. The wheel is simply covering new terrain!

In what we take to be the seminal chapter, "What Makes Right Acts Right", of his seminal text, *The Right and the Good*, W.D. Ross presents and defends seven ethical principles[1] that are meant to be able make sense of our moral obligations.[2] The purpose of this paper will be to apply this Rossian framework to some paradigmatic examples of the use of artificial intelligence in the healthcare setting.

2. The Rossian Framework

[1] On Ross' view what I am calling ethical principles are what he calls prima facie duties. We will be using the terminology of 'ethical principle' and 'duty' interchangeably in this paper. And we will avoid a discussion of what 'prima facie' is supposed to imply.

[2] By Ross's own admission the list of seven principles he develops is not necessarily exhaustive.

While it should be admitted that the Rossian Framework discussed below was originally developed as a Deontological ethical system, it should also be admitted that the kinds of principles that Ross defends are also the kinds of principles that others have tried to develop along Consequentialist or Virtue Theoretic systems. Because our goal in this present context is not the justification of these principles but the application of the principles to situations that arise from the use artificial intelligence in the health care setting, we will not defend or commit ourselves to any single ethical system whether it be Deontological, Consequentialist, or Virtue Theoretic. However, we do want to make clear that the philosophical project of figuring out why a specific situation is morally permissible (or impermissible) is not complete until the principles themselves are justified. We are simply saying that this is a project for another day (or another conference on the beach!).

With this preliminary statement of the incompleteness of the present project addressed it is reasonable to ask what principles Ross argues make sense of our moral judgments. In what follows we will present the principles Ross develops and discuss how those principles apply to specific situations.

2.1. Step 1: Identifying a Plurality of Principles that Govern Morality

According to Ross there are, broadly speaking, two different kinds of principles that govern morality. First, there are Value Based Duties; these are duties that depend on some conception of consequences. And, second, there are Duties of Special Obligation; these are duties that depend on past actions we have made or the past actions of others.

Ross presents as the Value Based Duties the duty of beneficence, the duty of self-improvement, and the duty of nonmaleficence. Beneficence requires that we produce as much good as possible, self-improvement requires that we make ourselves better, and nonmaleficence requires that we do not do bad things to other people.

Ross presents as the Duties of Special Obligation the duty of fidelity, the duty of gratitude, the duty of reparation, and the duty of justice. Fidelity requires that we stay committed to the promises we make, gratitude requires that we acknowledge the service of others, reparation requires that we make things right because of previous wrongful acts, and justice requires that we treat people fairly and that we reward virtue and punish vice.[3]

It is helpful to think about this list of seven ethical principles as identifying the kinds of things that should concern as about the moral status of a particular situation. When we are trying to figure out what the right thing to do is (or when we are trying to avoid doing the wrong thing) we should be worried about our duties to beneficence, self-improvement, nonmaleficence, fidelity, gratitude, reparation, and justice. However, it is also helpful to notice that just because we know what features of a particular situation should catch our attention morally it is also clear that simple awareness of these principles is not enough. We also need to know how to apply these principles in a particular situation.

Ross's ethical framework is a version of what philosophers refer to as *ethical pluralism* as opposed to *ethical monism*. Monistic views attempt to provide a single moral principle capable of capturing all moral considerations. Pluralistic views, on the other hand, hold that it is necessary to utilize several moral principles to sufficiently capture the fundamental moral considerations. The influence of Ross's ethical pluralism is evident in the dominant ethical framework within bioethics. The traditional bioethical framework posits four

[3] Ross's gloss on the duty of justice explains it in terms of rewarding virtue and punishing vice, but it is more common today to think of justice in terms of 'distributive justice' that best fits with an ordinary concept of fairness. Insofar as one aspect of distributive justice is that rewards/benefits and punishments/harms are distributed according to merit, or that we give to each person according to what they deserve or merit, Ross's gloss of rewarding virtue and punishing vice can be seen as capturing one important aspect of justice even if it leaves out others.

fundamental ethical principles: beneficence, non-maleficence, autonomy, and justice (see Beauchamp and Childress, 2019).

While there is obvious overlap between the fundamental ethical principles identified in Ross's framework and the traditional bioethics framework, there are also important differences. This is largely because each set of duties was developed for different purposes – Ross's aimed to provide a comprehensive ethical theory for all aspects of our lives while the simplified list of four duties was developed for the ethical needs of those working in institutional medical context. Given the different purposes, it is unsurprising that there are slight differences in the list of fundamental duties. One issue this brings to light is the fact that the increased use of artificial intelligence is causing large, rapid changes to the medical context compared to 1979 when Beauchamp and Childress wrote their influential, foundational bioethics text. As such, it might be necessary to modify or supplement the four principles identified in the traditional Bioethics framework. We will revisit this idea below when we consider potential implications of our analysis.

However, the differences between Ross's set of seven duties and traditional bioethics framework simplified set of four duties need not concern us now. We want to focus on issues related to the similarly *pluralistic* structure of these frameworks. In any pluralist ethical framework, the first step requires that we *identify* the relevant ethical principles that ought to guide our conduct. However, in each framework, there is an additional step for learning how to *apply* those principles to specific situations. This second step introduces moral complexities that need to be taken into consideration when we attempt to create ethically sound policies and procedures with the ethical.

2.2. Step 2: The Decision Procedure

Ross goes beyond identifying the ethical principles that govern morality by developing a decision procedure to help us identify what the right thing to do is in a particular situation. And we think Ross is exactly right when he argues that in many (and perhaps most) specific situations we will typically have more than one applicable duty, and often, these duties can conflict with one another.

The key to Ross's view is figuring out, in a specific case, which duty one should follow. According to Ross,

> when I am in a situation, as perhaps I am always am, in which more than one of these … duties is incumbent on me, what I have to do is to study the situation as fully as I can until I form the considered opinion (it is never more) that in the circumstances one of them is more incumbent than the other; then I am bound to that to do this … duty is my duty … [without qualification] in the situation.[4]

Thus, importantly, Ross's decision procedure requires that we can do two different things; first, we must figure out in a particular situation which of the ethical principles are applicable and *how* they apply to that situation, and then once we have done that, in the case of competing ethical principles, we must weigh the competing pressures of those applicable principles.

2.3. Interesting Results from the Rossian Framework

There are many different interesting results that fall out of the Rossian Framework, and it would, again, be helpful at this point to make a couple of them explicit.

[4] Ross, W.D. 1930. p. 19.

First, Ross admits that "Our judgments about our actual duty in concrete situations have none of the certainty that attaches to our recognition of the general principles of duty."[5] Notice what Ross is saying here, even though our recognition of ethical principles can be certain, our judgments about what is right in the specific cases is far from certain. We can be sure that justice is good, and that maleficence is bad, but the real world is messy. When we try to apply these duties to actual (i.e. concrete) situations we should not expect that things will still be clear.

However, this also doesn't mean that we should just throw up our hands. As moral agents when we are faced with a moral situation, we should be studying the situation to identify the applicable duties. We should be forming a considered opinion about which applicable duty (or duties) is most incumbent on us. And we should realize that once we have formed a conclusion about which duty (or duties) is incumbent on us, we should act accordingly. We can honestly say that given our understanding of the situation, our recognition of the applicable ethical principles, and our judgement about what is most morally significant, we are doing as much as could be expected from us.

Second, from what we have already pointed out, it should come as no surprise that if one adopts the Rossian Framework we should expect that reasonable and well-intentioned people might disagree about the morally right action in a particular situation. But again, given the account that Ross defends, this is not necessarily a bad thing. Given how common moral disagreement is, we should expect that an accurate ethical system to account for this kind of disagreement.

On Ross's view, while we should all recognize the legitimacy of the seven ethical principles that govern morality, the ability to recognize which principles apply in a specific case and the ability to weigh competing principles is difficult. However, even in the case of moral disagreement, the Rossian framework presented here should get us a step closer to understanding why right acts are right, and how we ought to proceed.

3. Our First Application of the Rossian Framework

One use of artificial intelligence systems in healthcare is to support clinical decisions for treatment. A relatively famous example (with well-known problems) was IBM's "Watson for Oncology" (Watson).[6]

As we all know, AI can be trained to make recommendations for medical treatment. Incorporating those recommendations to help support medical decisions, however, requires human interaction with the AI. Ethical difficulties are made obvious when AI recommendations differ from human recommendations; clinicians must determine how to respond to a conflicting source of information.

Relying on the Rossian framework will help clinicians figure out what they should do. First, we must determine which duties are relevant to the situation. In this kind of situation beneficence and non-maleficence are immediately obvious.

3.1. Step 1: Identify the Applicable Principles

Option A: Defer to the AI System

[5] Ibid. p. 30.

[6] This project was discontinued by IBM in 2020 and various parts of the larger Watson for Health system have been sold off to a private equity firm. We aren't concerned with this specific AI system, but rather with some of the ethical issues this example highlights for AI in a healthcare context.

The AI system's recommendation has the *potential* to be beneficent by saving time, increasing efficiency, and utilizing a must larger data set for forming evidence-backed conclusions. When an AI disagrees with the human recommendation it's *potentially* because the AI is making use of a relevant dataset that is simply too large for any individual to consider. Similarly, considerations of non-maleficence are relevant. AI's conclusions may consider potential risks of which the clinician was either unaware or simply failed to consider.[7]

Option B: Defer to the Human

The same duties can push against AI recommendations as well. There may be reason to think that the training data of the AI system is biased towards specific populations. For instance, many of the problems known with Watson involved worries about biases in training data that prevented its applicability to many real-world scenarios. So, human recommendations may be sensitive to local features/context of an individual patient that are not well represented within the AI's training data. Human recommendations may thereby have the potential for the patient to benefit as well. Moreover, insofar as these AI systems contain bias, overconfidence in AI recommendations has the potential to cause disproportionate harm to populations that are not as well-represented in its training data thereby worsening inequalities.

3.2. Step 2: The Decision Procedure and a New Complication

Recall that the decision procedure requires that we *identify* applicable duties, but also that we consider how to weigh duties against one another in our specific situation. This has important implications for incorporating AI recommendations in an ethically responsible manner. Initially it might seem reasonable to treat AI systems as something analogous to considering second opinions. However, the 'black box' problem for AI systems makes using their recommendations during this aspect of the Rossian decision procedure difficult.

When I disagree with another person, they can explain their underlying reasoning to me which I can then consider in comparison to my own reasoning. The problem with the black box is that the factors that go into the AI's reasoning and how those factors are weighted is often hidden from the user's view. For example, Watson gave links to studies for supporting evidence, but "it obfuscates the scoring criteria that it uses to value some studies over others. In other words, the platform 'black boxes' the values that are coded into its scoring system"[8]. Similarly, insofar as the reasoning is hidden from view, it isn't clear to the user how an AI system may be weighing beneficence and non-maleficence against one another when making its recommendation. This 'black box' creates a critical obstacle for users' ability not only to trust the recommendations, but also their ability to incorporate AI's recommendations when weighing moral considerations against one another while trying to responsibly form a considered judgment. Moreover, this 'black box' raises difficult ethical issues about respecting physician and patient autonomy insofar as the reasoning for an AI recommendation is opaque. To this extent, the Rossian ethical framework we are applying may lend some credence to recent calls for 'explainable AI.'[9]

---

[7] However, even on the *assumption* that an AI tool has these benefits, they can only be realized if those systems can earn the trust of medical professionals – both the history of failures such as those associated with IBM's Watson for Oncology and the 'black box' problem we discuss below may serve as obstacles to such adoption.

[8] Tupesela & Nucci, 2020.

[9] Explainability could help allow clinicians to make more ethically informed decisions while incorporating AI recommendations into their decision procedures. As such, AI developers may have moral reason to adopt procedures that focus on providing increased explainability. However, those procedures would require some method for balancing this against other values since increased explainability may come with its own trade-offs (such as with reliability or accuracy).

The discussion of the present type of scenario would also be incomplete without paying attention the duties of justice and reparations. When things go wrong, considerations of reparation requires that we identify responsible parties and that we attempt to repair any harms done. The potential of using AI to support medical decisions raises difficult ethical issues about how to apply these moral duties. While surely various responsibilities land on the medical professionals as *users* of AI, we might also ask questions about the responsibility of those *developers* of AI – notice that it is the *developers* that would have a responsibility to incorporate an appropriate degree of "explainability" as this is a necessary pre-condition for the *users* to be able to responsibly engage with the AI system's suggestions.[10]

4. Our Second Application of the Rossian Framework

We are currently working to choose an example explicitly dealing with data governance that we think will be most useful for illustrating the application of a Rossian, pluralistic ethical framework and the issues it helps make salient.

4.1 Step 1: Identify the Relevant Principles

Once we have chosen which example would be most useful as an illustration within the data governance context, we will distinguish the available choices that could be made and identify the different Rossian duties that would apply to that context.

4.2 Step 2: The Decision Procedure

Once we have chosen which example would be most useful as an illustration within the data governance context, we will apply the decision procedure.

5. Implications for the Ethical Incorporation of AI in Healthcare and Data Governance

We will be adding a section that attempts to make explicit some of the implications we think our earlier discussion has for how to develop an ethical framework for policies, procedures, and documentation for AI in healthcare and data governance.

*We will emphasize that the framework needs to be flexible enough to accommodate the possibility that different people and different organizations might *rationally* decide to weigh competing ethical duties/considerations differently. This is especially important when considering the differences between large medical organizations with many resources at their disposal vs. smaller medical organizations with much less resources available. However, to be able to hold people and organizations accountable for their choices regarding data governance policies and procedures despite that flexibility, the ethical framework would likely need to require that people or organizations document how they have decided to weigh those competing considerations and the processes they used to come to that decision. That documentation would help auditors better evaluate whether the difference in governance structures/policies/procedures/etc. is a *rational* disagreement, an ethical breach, or a post-hoc*

[10] Smith *et al.* (2024) cites an unnamed IBM executive claiming that "Watson does not make decisions on what a doctor should do. It makes recommendations based on hypothesis and evidence based [sic]" (p. 79). As they make clear in their paper, this seems to be a way of putting full ethical responsibility on the clinician. This may seem to be a way of showing respect for the clinician's autonomy in their decision making. However, as Smith *et al.* point out, "[t]his is clearly advantageous for SDCs [software development companies] ... The clinician in this role insulates the SDC from the consequences of their system's errors" (2024, p. 79). Such an arrangement could cause various harms (conflicting with our duty of non-maleficence) insofar as this "shield: can encourage SDCs to adopt more lax governance policies for AI development than if they are forced to be cognizant of risks for their company when potential errors of that system cause harm.

rationalization. This could also thereby allow for flexibility and rational ethical disagreement between people/organizations without devolving into an 'anything goes' approach.

We will also reiterate and more fully develop how our discussion may lend credence to the push for explainable AI as a critical ethical consideration and how there will thereby need to be data governance processes put into place aimed at facilitating greater explainability. This will then lead into the possibility that our discussion may suggest that AI is leading to sufficient changes within the healthcare context that we may need to expand the traditional four principles of bioethics to a more complex but more robust set of principles.

6. Conclusion

We have tried to accomplish three tasks; first, to point out that while the actual situations that arise from artificial intelligence are new there is no need to entirely reinvent the wheel. Second, to present the ethical principles and decision procedure developed by Ross with an eye towards illustrating the kinds of questions that should be on our minds as we address the issues at hand. And third, the Rossian framework can be clearly, and helpfully, applied to cases involving artificial intelligence in the health care setting. These cases can then be used to create helpful examples that can serve to inform how we ought to proceed.

Hopefully what is clear from what we have said is that if we hope to find the right answers, we must know the right questions to ask.

7. References

Beauchamp TL, Childress JF. 2019. *Principles of Biomedical Ethics*. Oxford University Press, New York. 9th edition.

Ross, W.D. 1930. *The Right and the Good*, Oxford: Oxford University Press, New York.

Smith H, Birchley G, Ives J. "Artificial intelligence in clinical decision-making: Rethinking personal moral responsibility." *Bioethics*. 2024 38 (1): 78-86.

Tupasela, Aaro & Ezio Di Nucci. 2020. "Concordance as evidence in the Watson for Oncology decision-support system." *AI and Society* 35 (4): 811-818.