JPE

How AI Can Aid Bioethics

WALTER SINNOTT-ARMSTRONG
Duke University

JOSHUA AUGUST SKORBURG
University of Guelph, Canada

This paper explores some ways in which artificial intelligence (AI) could be used to improve human moral judgments in bioethics by avoiding some of the most common sources of error in moral judgment, including ignorance, confusion, and bias. It surveys three existing proposals for building human morality into AI: Top-down, bottom-up, and hybrid approaches. Then it proposes a multi-step, hybrid method, using the example of kidney allocations for transplants as a test case. The paper concludes with brief remarks about how to handle several complications, respond to some objections, and extend this novel method to other important moral issues in bioethics and beyond.

Keywords: artificial intelligence, machine learning, bioethics, moral conflicts, medical decision making, ideal observer theory.

Morality and computing are often seen as opposite ends of a spectrum: Computing is hard science, formal, and unfeeling, whereas morality is soft, situational, and emotional. They could not be further apart, according to this common view. This contrast raises questions of whether we could ever bring morality and computers together—and if so, how. These questions become all the more pressing as Artificial Intelligence (AI) and Machine Learning (ML) are increasingly used to make or guide medical decisions that raise moral issues. But what is the point of bringing computing and morality together in the medical context? One goal

Contact: Walter Sinnott-Armstrong <walter.sinnott-armstrong@duke.edu>,
https://orcid.org/0000-0003-2579-9966
Joshua August Skorburg <skorburg@uoguelph.ca>
https://orcid.org/0000-0002-3779-5076

https://doi.org/10.3998/jpe.1175

is to use AI to improve human moral judgments in bioethics. In this article, we propose a way to do this. But first we need to see why humans need help.

Some Problems

A kidney transplant surgeon told us that he was woken up at 3 a.m. and told that a car crash had killed a kidney donor, so he had to decide which of his patients would receive this kidney. He had only a few minutes to make this life-changing decision, because the kidney would not remain viable for long, and the chances of a successful transplant were going down every minute. The staff needed to prepare the chosen patient for surgery, and the doctor needed to leave as soon as possible for the hospital to do the surgery. He was still groggy from sleep, he had no time to review patient charts, and he presumably liked some of his patients more than others. This kind of situation is far from ideal for forming moral judgments about who should receive a kidney.

Other moral judgments are formed in better settings. When there is time for a hospital ethics committee to exchange views, and its members have had time to review information about the relevant patients, one might expect moral judgments to be more trustworthy. That is why hospitals have ethics committees. However, members of hospital ethics committees rarely have time to review all of the relevant information carefully, and they often meet too briefly for everyone on the committee to be able to present their perspectives fully. Moreover, some members of the group might be more willing to listen to certain members rather than others with equally valid perspectives. And groupthink is also a danger. Of course, hospital ethics committees try to avoid these distortions, but they often fail, like other committees.

In addition, a committee of experts often has different values than the local community. Sometimes experts' judgments are superior to those of the public, but the public is not always as misguided as many elites assume. Experts sometimes learn from the public. In any case, what the public thinks still matters, both because public moral concerns might be different but still valid, and also because the public pays for public hospitals, which gives them some right to have their values represented. Moreover, ignoring the public's values leads to miscommunication and misunderstanding and might make the public less inclined to support the hospital. Involving healthcare service users in planning and research can help clinicians and administrators identify potential problems before they arise. In these ways, aligning healthcare services with the values of stakeholders is not only fair but also useful.

Sources of Error

Of course, the public is not always right. Sometimes they are misinformed, forgetful, confused, emotional, or biased. In such cases, we can still ask what the public would say about a moral issue if they were not misled in these ways. This method of extrapolation was suggested by Justice Thurgood Marshall in his opinions about capital punishment:

> In *Furman [v. Georgia*, 1972], I observed that the American people are largely unaware of the information critical to a judgment on the morality of the death penalty, and concluded that if they were better informed they would consider it shocking, unjust, and unacceptable. 408 U.S. at 360–69. (Marshall, 1976)

This "Marshall Hypothesis" has been confirmed (Sarat & Vidmar, 1976) and suggests more generally that we can use statistical methods to find out not only what the public *does* believe now but also what they *would* believe under more ideal circumstances. Public policies can then be based on these idealized views, although doing so will be controversial.

Marshall emphasized factual ignorance, but other factors also distort moral judgments. We all sometimes forget or fail to attend adequately to morally relevant facts. When many complex considerations support each side of an issue, we all sometimes get overwhelmed and confused. In addition, intense anger, disgust, and fear often lead us to fixate on a small subset of the morally relevant facts or even to base our moral judgments on factors that are morally irrelevant, such as physical appearance or height. Yet another source of moral error is bias in a broad sense that includes cognitive biases and favoritism towards oneself and one's family and friends as well as racial or gender prejudice.

Most people (including many moral anti-realists) admit that moral judgments should not be based on ignorance, forgetfulness, inattention, confusion, excessive emotion, or bias. They cite such influences to criticize other people's moral judgments, even while they fail to apply the same standards to themselves. Thus, these factors are distortions even according to people who are subject to them. To call them errors is not to impose external standards that they reject.

Disputes still arise over whether a particular moral judgment is based on such sources of error. Opponents often accuse each other of ignorance, forgetfulness, inattention, confusion, powerful emotion, and bias. How can we settle such disputes? In our view, the best way is to predict which moral judgments each side would make if they were not ignorant, forgetful, inattentive, confused, overly emotional, or biased. Accurate predictions can point to corrected moral judgments that reflect people's deeply-held values rather than their fleeting

foibles. The point is not to impose external standards that they might reject but instead to help them apply their own standards.[1]

But how can we predict what humans would endorse in such idealized circumstances? The answer might lie in AI. When properly programmed, AI has the potential to find, store, and use a lot more information than humans, even if not all information. AI also will not forget or fail to attend to any information, will not get confused by complex information, will not be misled by intense emotions, and could potentially reduce bias (see below).[2] Thus, many of the most common sources of moral errors by humans might be avoided by computers that are properly programmed.

How can we program human morality into computers? Following Wallach & Allen (2009), we can distinguish three main ways: top-down, bottom-up, and a hybrid. We will discuss these three in turn.

Top-Down

The top-down approach starts with principles at a high level of generality, then programs them into a computer along with information about particular situations, and finally applies the principle to the situation to infer a moral judgment. One crucial question for this approach is: Which principles?

A popular contender is Isaac Asimov's three laws of robotics (Asimov 1950):

> A robot may not injure a human being or, through inaction, allow a human being to come to harm.
>
> A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
>
> A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

These simple rules strike many people as plausible, but they quickly run into trouble. Should a robot roam the world preventing any harm to any human?

1. This distinction between corrected and raw moral judgments is analogous to the distinction between competence and performance in a language (Mikhail, 2007). Just as a language can be characterized by competence rather than performance, so human morality can be characterized by the moral judgments we would make if we were not misled by ignorance, forgetfulness, inattention, confusion, excessive emotion, or bias.

2. This is not meant to downplay the fact that AI applications can and do produce new forms (or reproduce and exacerbate existing forms) of bias against marginalized groups. A growing body of popular and scholarly work addresses these questions in detail, and so our strategy here is to acknowledge this work, while also exploring the possibilities of AI to counteract these biases in some medical contexts.

What is a robot to do when one human attacks another human, and the robot cannot prevent harm to the victim without injuring the aggressor? And what is a robot to do when one human tells the robot to go to the United States now, but another human tells the robot to go to the United Kingdom now? Such examples show that Asimov's laws are inadequate.

Perhaps philosophers can do better. Consequentialist moral theorists propose that we should do whatever maximizes good consequences and minimizes bad consequences for everyone in the long run (Driver, 2012). This principle sounds plausible, but it has been criticized vigorously, and consequentialism might be especially dangerous in the context of AI. The best way to minimize suffering and death in the long run might be to kill all humans now (because after that no humans would die or suffer or hurt other animals). Thus, if we program AI to minimize bad consequences, it might try to kill us all (Bostrom, 2014). Moreover, a computer cannot apply a consequentialist principle to a concrete case without having information about how all options would affect all people in the long run. It is not clear how to obtain this information or whether applying it could be computationally tractable. And if consequentialists try to ameliorate this problem by focusing on only some people in the short run, then it is not clear how they could justify choosing those people or that time frame. Some theorists have attempted to overcome these problems (Gips, 1995) but without much success (Wallach & Allen, 2009, 86–90).

The problem of inadequate information could be solved if simple rules were programmed into the computer. A computer could determine whether an act violates the rule "Keep our promises" without having to calculate any consequences of that act. To determine whether an act violates the rule "Don't kill", a computer would have to determine whether the act causes death, but it still would not have to calculate all of the consequences of the act for other people far in the future. So it would be easier (though still not easy) to program such rules into computers.

Rules like these seem to have exceptions, however. Despite disputes over interpretation, Immanuel Kant seems to argue that it is morally wrong to lie even when lying is the only way to save a friend from an assassin (Kant, 1797). Absolute rules like this might be simple enough to program into an AI, but most people would reject the resulting judgments.

Other rule-based moral theorists are not absolutists. W. D. Ross, for example, provided a list of seven basic duties: fidelity, reparation, gratitude, justice, beneficence, non-maleficence, and self-improvement (Ross, 1930). In medical ethics, Tom Beauchamp and James Childress proposed principles of Respect for Autonomy, Beneficence, Non-maleficence, and Justice (Beauchamp & Childress, 2019). They all admit that these duties or principles can be overridden, so they allow lying or violating another duty or principle to save a life in some cases. This flexibility makes this approach plausible, but it also makes it hard to program into a computer. Concepts like justice, fidelity, and autonomy are too vague to

program into a computer. Moreover, these duties and principles are bound to conflict sometimes, and then how is a computer supposed to tell which reasons are adequate to override a duty or principle? Without some way of resolving conflicts between overridable duties, the computer will not be able to reach any conclusion about what an agent ought to do overall. Yet another problem is how to determine which rules and weights to deploy. Again, some theorists have attempted to solve these problems (Anderson & Anderson, 2011) but without much success (Arkin, 2009, 106–108; Wallach & Allen, 2009, 127–129).

The biggest problem for the top-down approach is that it requires us to choose which principles to use at the top. As we saw, competing moral theorists want different principles, and it is hard to see how to justify picking one set of principles instead of another. That is why others try another approach that does not assume any principles, rules, or duties.

Bottom-Up

A second approach works in the opposite direction. Instead of applying general principles to particular cases, the bottom-up approach derives principles from cases. The most promising examples of this approach use unsupervised deep learning.

The method is *deep* insofar as it uses multiple levels of nodes connected by edges. Reinforcement *learning* occurs when the weights of these edges—that is, the probability that activation of one node leads to activation of a connected node—increase or decrease in light of success or failure at a task. This process is *unsupervised* insofar as the programmer does not impose or assume any theory or even categories in advance. The AI itself determines which categories work best for the task.

A plausible application of unsupervised deep learning to moral judgment would require several steps. First, researchers ask many (perhaps 1000) participants to describe many (perhaps 100) moral problems. Each problem must be described in the participant's own words, and each participant needs to indicate which option in the scenario is morally wrong. Next, each participant needs to read many (perhaps 100) of the moral scenarios that other participants wrote and indicate which actions are morally wrong in those other scenarios. The researcher thus accumulates moral judgments by the thousands.

With enough data of adequate quality, the AI could perhaps construct a model that will predict the human participant's moral judgments on the basis of the words in the scenarios. Using a simplistic form of cross-validation, it could, for example, use 90% of the scenarios as a training set to develop its model, use the remaining 10% of the scenarios to test how well that model can predict moral judgments in new cases, revise the model to the extent that its predictions are

inaccurate in that testing set, and test it again on a completely new set separate from the old training set. Each time it is tested and revised, the computer learns. That is why this method is machine *learning*. To the extent this is successful, the computer becomes able to predict human moral judgments accurately. The AI need not endorse or even understand the moral judgments, but it can tell us which judgments humans would endorse in a moral dilemma. The computer would be predicting not what a doctor and others would say when aroused at 3 o'clock in the morning but, instead, what the doctor would say in experimental circumstances when less tired, bothered, and hurried.

This method has several advantages. Whereas the top-down method begins with a moral principle that some others are bound to question, researchers using the bottom-up method do not tell the computer which categories are relevant to predicting human moral judgments. That is what makes it *unsupervised*. The bottom-up method thus avoids serious objections to the top-down method.

However, the bottom-up method introduces its own problems. First, the bottom-up method devours a great deal of data. Its tremendous appetite arises because very many words or groups of words might be relevant to predicting human moral judgments, and it cannot limit what it considers without assuming which categories are relevant. Thus, it must consider every word, such as definite versus indefinite articles, since such apparently irrelevant words still might predict patterns of moral judgment. The system then needs a tremendous body of data in order to determine precisely which words or combinations of words are relevant to predicting moral judgments.

Second, an unsupervised deep learning system can predict *which* acts humans will judge as wrong, but it cannot tell us *why* humans judge those acts and not others as wrong, much less why they are wrong. When an AI does not use human categories, it cannot reveal reasons for the moral judgments that humans would recognize as reasons. This opacity is an especially serious problem when AI is used in legal, military, and medical contexts where people's lives are at stake, and they deserve to know why the AI decided against them.

As before, some theorists have attempted to solve these problems (Guarini, 2006, 2011) or provide limited kinds of interpretability (Wachter et al., 2018). Their success has been questioned (Arkin, 2009, 108; Wallach & Allen, 2009, 132–133), but their attempts along with others (Anderson & Anderson, 2011) do point the way toward a more promising method.

A Hybrid

These problems for the top-down and (unsupervised) bottom-up methods leave us wanting a better alternative. The method that we propose here starts

with morally relevant categories that resemble those in the principles of the top-down method, and it also uses machine learning like the bottom-up method. That is why we will call our method a hybrid, although it can also be classified (perhaps more accurately) as a form of a supervised bottom-up method.

Our method is inspired by the ideal observer tradition in moral philosophy (Firth, 1951). According to ideal observer theorists, very roughly, we are justified in believing that an act is morally wrong *if and only if* ideal observers would disapprove of that act and other acts of the same kind.[3] This central claim of ideal observer theories is plausible, because an ideal observer should know what is morally wrong if anyone does. This approach assumes nothing about whether ideal observers would disapprove of acts on the basis of consequences, rules, or whatever, so it avoids the heated controversy between consequentialist and deontological moral theories.

Still, ideal observer theories need to be filled out in at least two ways. First, each ideal observer theory needs to specify which observers are ideal. We require ideal observers to be informed, rational, and impartial, because these standards imply that ideal observers avoid ignorance and forgetfulness, confusion and distorting emotions, and bias and prejudice—the common sources of moral error discussed above.

Second, each ideal observer theory also needs to specify which features of acts matter. An ideal observer would be arbitrary and, hence, irrational or incoherent if it disapproved of killing a bald person but approved of killing an otherwise similar person who was not bald. In contrast, it would make sense for an ideal observer to disapprove of killing out of hatred but not disapprove of killing in self-defense. Thus, some features of acts are morally relevant, but others are not. The morally relevant features are what guide the reactions of ideal observers and determine when acts are "of the same kind."

In order to apply this framework, we need to determine which features of acts are and are not morally relevant. Some philosophers have attempted to argue for their own lists of morally relevant features (Gert, 2004). Our approach is more empirical. We use surveys and experiments to try to come up with a list of features that are or are seen as morally relevant.

3. Sometimes the disapproval of ideal observers is seen as *constituting* moral wrongness, but other ideal observer theorists claim only that disapproval by ideal observers is *evidence* of moral wrongness. What matters for our purposes here is only the latter, epistemological version. It is also worth noting that some versions of this general approach ask whether ideal observers disapprove of social norms that permit acts of the same kind. This variation is what matters for hospital policies and professional codes of ethics.

Kidney Transplants

As a test case, our team focuses on kidney transplants. A kidney can come from a cadaver or a live donor, since most of us have two kidneys but need only one. Some donors offer a kidney to a needy stranger, but this is rare. Live donation is more common among donor recipient pairs. Imagine that a wife needs a kidney, and her husband is willing to donate one of his, but their blood types are not compatible. Suppose also that a brother needs a kidney, and his sister is willing to donate one of hers, but they are also not compatible. If the sister is compatible with the wife, and the brother is compatible with the husband, then they can exchange kidneys, and both patients get what they need. Such kidney exchanges become extremely complex when they involve large numbers of potential donors and recipients. That complexity makes AI useful.

The general problem is that there are not enough donors (live or dead) to supply all patients in need. Roughly 100,000 people in the US alone are waiting for kidney transplants.[4] As a result, doctors or hospitals often have to decide which one of two patients should receive a kidney. Currently, most kidney exchanges make these decisions on the basis of medical compatibility, age, health, organ quality, and time on the waiting list. However, many people in our surveys (Doyle et al., in progress) report that transplant centers should also consider other factors, such as number of dependents, record of violent crime, and misbehavior causing the kidney disease. Which features should determine who gets a kidney is thus a controversial issue where the public seems to disagree with hospitals.

This issue is not purely medical but also moral, both because some features of patients are morally tinged and because it affects the welfare of other people. Identifying a recipient can affect the chances of other patients receiving a kidney soon or ever. It seems unfair to base such decisions on race, gender, religion, or certain other features. It also strikes some as unfair to base such decisions on features like history of violent crime or alcohol abuse, which others want to treat as negative indicators. Thus, some features seem clearly morally relevant, others seems clearly morally irrelevant, and others are controverisal.

This problem becomes complex when a large number of features of each patient are considered and a large group of potential donors and recipients are candidates for kidney transplant. No matter how we prioritize some patients over others, the puzzle of figuring out the best way to distribute the inadequate supply of kidneys becomes too complex for any human being. For this reason,

4. https://www.kidney.org/news/newsroom/factsheets/Organ-Donation-and-Transplantation-Stats [accessed 24 November 2020]

algorithms are used to help kidney transplant centers decide who gets a kidney first (e.g., Roth, Sönmez, & Ünver, 2004).

Nonetheless, many people react with horror to the idea of AI deciding who lives and who dies. They do not object in the same way to using AI to diagnose diseases, even though both uses of AI are difficult and affect lives. Why is there more opposition to using AI to determine who gets a kidney? Moral subjectivists might answer that there is no objectively correct judgment about which act is morally right, whereas there is an objectively correct judgment about which disease a patient has. Another reply might be that AI (at least in the near future) cannot know what it is like to be a human being, much less a patient with kidney disease fighting to stay alive. Such sensitivity to human concerns might seem to be a prerequisite for making life-and-death decisions. However, some of these objections can be overcome if we can figure out some way to build human concerns, especially human morality, into an AI without assuming that these concerns and moral judgments are either universal or objectively correct.

How to Build Morality Into AI

That is what our team is trying to do. Our research involves several steps and methods, including survey, experiment, theory, and computation. This interdisciplinary enterprise is too complex and too preliminary to provide full details here, but a general outline should convey the basic idea.

Gather Features

Our project begins by crowdsourcing opinions about which features of patients should and should not influence who gets a kidney. We plan to survey both the general public and also doctors and hospital administrators, including those who are engaged in kidney transplants.

It is important that we ask them not only what *should* but also what *should not* influence who gets a kidney. The first question will provide morally relevant features or categories to supervise the learning by the AI. The second question will show us which features we should leave out of the AI in order to avoid biases that people themselves recognize as biases (but see below for complications).

No mere survey can tell us which features of patients really should or should not influence who gets a kidney. Our goal is only to find out which features people think ought to influence that moral decision. By starting with features that other people deem morally relevant, we avoid imposing our own assumptions.

In addition, survey participants might—and did—mention some features that we overlooked.

Edit Features

Despite such advantages, open-ended surveys also have disadvantages. Participants' descriptions of features are usually vague or ambiguous. Different participants often describe the same feature in different terms. Our next step is then to clarify and remove redundancy in the features that participants supplied.

In addition, participants who spend only a few minutes in our surveys might fail to mention features that they would see as relevant if they thought of them. To fill in such gaps, we can add features that philosophers and other ethicists have proposed as morally relevant (Gert, 2004). Each moral theory in effect picks out different features as morally relevant.

Test Features

This editing process requires assumptions and can allow bias to creep in. To reduce these problems, we need to try to remain as neutral and open to criticisms as we can. We also need to test our revised list. After editing our features for clarity, redundancy, and completeness, we can determine whether the edited list is accurate by asking new survey participants whether they agree that those edited features are morally relevant and whether they want to add any features to the list or remove any. This process of refinement and testing might need to be repeated in multiple stages. The result will be three lists: features that are seen as morally relevant, features that are seen as morally irrelevant, and features that are controversial.

We have already gathered some preliminary information about these features (Doyle et al., in progress). Some good news is that the vast majority of our participants agree that race, gender, sexual orientation, religion, political beliefs, wealth, and reliance on government assistance should not influence who gets a kidney (although a different sample always might yield different results on this). It is also not surprising that almost all of our participants agree that the urgency of need, time on the waiting list, and likelihood of transplant success should affect who gets a kidney, along with age, current health, and life expectancy as well as quality of life after a transplant. Most also said that smoking and drug and alcohol abuse currently (after diagnosis with kidney disease) as well as historically (before diagnosis) should matter. It was more controversial whether mental health, record of violent or non-violent crime, or number of children or

elderly dependents should affect who gets a kidney when there are not enough for everyone.

These preliminary results are only the beginning of a long experimental program. Many of these features need to be refined: Which kinds of current health problems or crimes count? Does it matter whether past smoking, drinking, and drug abuse caused the current need for a kidney transplant? Are age and current health as well as exercise habits really just proxies for life expectancy? Is life expectancy what matters or is it rather quality of life? We will need separate experiments to answer such questions, such as by varying life expectancy independently of age and health to determine which of these factors drives moral judgments in this area. Unfortunately, no feasible set of experiments could ever answer all of the relevant questions. Still, a preliminary list of morally relevant features can get us started.

Conflicts

After constructing a list, we need to determine how much weight is put on various features on the list. We could simply ask people in a new survey. Unfortunately, people seem to be better at identifying which general features are morally relevant than they are at reporting how much weight those features should have in conflicts. A more promising method constructs conflicts among features on the list and then asks participants which patient should receive a kidney in those conflicts.

It is easy to construct conflicts from a short list of features. If features 1–3 are on the list, we can ask participants to decide who gets a kidney when feature 1 favors patient A but features 2–3 favor patient B, when features 1–2 favor patient A but feature 3 favors patient B, when features 1 and 3 favor patient A but feature 2 favors patient B, and so on. We can also vary the degree of the difference between Patients A and B, such as the difference between their life expectancies, alcohol consumption levels, or number of dependent children. How a participant distributes the kidney in these conflicts will reveal how those features interact in producing the judgments of this participant.

One obvious problem is that the number of comparisons grows exponentially along with the number of relevant features. It quickly becomes practically infeasible to ask participants so many conflicts as well as computationally intractable to determine how all of the features interact. Nonetheless, some progress can be made by analyzing a limited number of features at a time and by asking only about a subset of conflicts that provide the most information about how people weigh features.

We have begun to gather data of this kind on our website (whogetsthe kidney.com). One feature that we added was an option to flip a coin instead of giving the kidney to either patient. This coin flip option gives extra information that helps us derive weights and interactions. For example, if participants flip a coin when they see no significant difference between the patients, then we can tell which differences participants see as significant.[5]

Analysis

After gathering enough data of the right kind, machine learning can help to determine: (A) Which features really do influence participants' moral judgments about who should get a kidney, (B) How these features interact to produce an overall judgment, and (C) Which model best predicts each individual's moral judgments. The AI can be trained on one set of data, tested on another, and refined in light of the test results. In this way, it can learn how to predict human moral judgments about distributing kidneys.

This machine learning is supervised because it uses the categories that participants judged to be morally relevant in previous surveys. That minimal supervision reduces the amount of data that is required. It also makes the results more interpretable. The AI will, hopefully, be able to spell out a model that predicts an individual's moral judgments. By inspecting that model, we can find out which features affect that person's moral judgments. Those features (at least sometimes) correspond to reasons for the judgment, such as when a patient's old age is a reason to give the kidney to the other, younger patient. If the features that predict a moral judgment correspond with the individual's reasons for endorsing that moral judgment, then the AI can reveal not only *that* but *why* this person makes that moral judgment. This transparency is an advantage over the bottom-up method described above (using unsupervised deep learning), both because it makes the theory explanatory and also because people often deserve to know the reasons behind moral decisions that affect their lives.

Promise

One payoff from this procedure is that we can compare what people say *should* affect their moral judgments and how they say those reasons *should* interact against which factors really *do* affect their moral judgments and how. We can

5. Significance here is not the same as relevance, because relevance applies to individual features, whereas significant difference is a relation between features.

learn how much insight humans have into the computations that produce their moral views. We can also compare the models that best predict moral judgments by different individuals, and different groups. This method can, thus, help move moral psychology forward in fruitful directions.

Complications

Although this general plan might look simple, several complications arise quickly.

Philosophers often speak of weighing one moral consideration against another. Reality is not so simple. It is unlikely that each consideration can be given a simple weight: five kilograms of dependents (three children and two elderly parents?) minus four kilograms of fault for causing one's own kidney disease (because one drank four drinks a day for forty years?) equals one kilogram in favor of this patient. No way!

Instead, moral considerations are more likely to interact in complex ways that philosophers are only beginning to map (Horty, 2014; Snedegar, 2017). One complication is that the force of a moral consideration can vary with context, as particularists emphasize. Having five or ten child dependents strikes many as a reason to give the parent a kidney, but the same number of children might not count at all or as much if that patient was convicted of child abuse. Again, having been on the waiting list only a short time might count less against a patient if that patient's need is as urgent as others who were diagnosed earlier, because they had access to better doctors. And the quality of a patient's life after transplant might matter less if that patient is not responsible for facing a deprived life. Different factors will surely interact in different ways for different individuals in different situations. Machine learning might be able to capture more of these complex interactions than humans do, but we should not expect perfection. There are too many subtleties and contexts to figure out.

Another complication is probability. Both risk (known probabilities) and uncertainty (unknown probabilities) pervade kidney transplants. It is unrealistic to claim that someone's life expectancy is precisely 42 years. Instead, there is a probability that this individual will die within 10 years, another probability that she will die in 10–20, 20–30, 30–40, 40–50, etc. The reality is a complex array of probabilities rather than a single precise figure, and we do not know any of these probabilities precisely. The same complication arises for other factors. If someone smoke and drank, there is a probability that his smoking and drinking caused his kidney problems, but nobody can be certain. The probability might vary with how much he smoked or drank and during which period of life.

Most people are not knowledgeable or sophisticated enough to understand probabilities (more on this below). Still, we will need to do our best to figure

out how probabilities affect people's moral judgments, accommodate legitimate aversion to risk and uncertainty, and then correct for clear errors in probabilistic reasoning. These tasks will not be easy, but careful training and research can make some progress (Gigerenzer, 2015).

Yet another complication is that humans are not consistent. They make different judgments of the same act in the same circumstances when they judge on different days and in different frames (order and wording, for example). We need to understand such foibles and correct for them. For example, if we can understand how the order or wording of scenarios influences moral judgments about them, then an AI can predict which moral judgments would be made by humans who saw both frames. This research will again be difficult and imperfect, but some progress can be made if we try.

Ideal Observers?

Can we also correct for other sources of error? Recall that an ideal observer is supposed to be informed, rational, and unbiased. If an AI is supposed to play the role of an ideal observer, as we propose, then we need to construct it so that it predicts not only which moral judgments humans actually make but also how humans ideally would judge acts if they were informed, rational, and unbiased.

To say that a moral judgment or decision is biased is to say that it results from a cognitive bias or some feature—such as race, religion, gender, sexual orientation, wealth, or attractiveness—that should not affect the moral judgment or decision. In our surveys, most people agreed that these features should not affect kidney distribution. This consensus gives us reason to avoid basing moral judgments on those features (though we could exclude these features on other grounds as well).

One might think that these biases could be avoided simply by imposing a kind of veil of ignorance (Rawls, 1952, 1971). If participants in a survey do not know whether one or the other of two patients is black, Muslim, female, gay, poor, or ugly, then those features cannot directly influence their moral judgments about which patient should get a kidney. Nonetheless, their moral judgments could still be influenced indirectly when they use other features that are correlated with the features whose influence we and they want to avoid. For example, geographic information (such as postal codes) will correlate with race and wealth, so using geographic information might implicate forbidden features. Such indirect bias will be hard to detect and remove completely. Still, ignorance of forbidden attributes can enable us to reduce bias by better approximating which moral judgments humans would make if they were less biased. An AI

could perhaps then reflect those less-biased judgments if it also has no information about such morally irrelevant factors.

Moreover, when bias does indirectly creep in, we might be able to determine how much bias occurs by comparing the probability that, for example, a black patient gets a kidney with the probability that a white patient gets a kidney, other features being equal.[6] Then we will be able to correct for that bias in an AI. It is much harder to correct for biases in humans. Nonetheless, an AI that excludes known biasing factors and also corrects for known indirect biases could still display some residual bias. Perfection is unattainable, but that should not stop us from trying to do better than we do now.

To become a proxy for ideal observers, an AI must also predict which moral judgments humans would make if they were informed and rational. The problem is that our survey participants are not fully informed or completely rational. The judgments they made in our surveys and experiments were probably often based on ignorance and confusion.

To solve this problem, recall the Marshall Hypothesis. Just as Marshall analyzed statistics to determine what people would say about capital punishment if they knew more about it, so we can use information from our surveys and experiments to determine how people would distribute kidneys if they were informed and rational. We can measure effects of ignorance by asking participants questions to reveal how much they know about the situation or by providing participants with different amounts of information: no, some, moderate, and much information. These manipulations will enable us to measure how much their moral judgments would have been affected if they had had that information when they originally took the survey. If their judgments change when they receive more information, their original judgment depended on ignorance of that information. And if that new information is relevant, and they see it as relevant, then they and we should agree that the more informed moral judgment is better or at least more in line with their real values.

Similar manipulations can be used to determine when moral judgments depend on confusion. In our kidney test cases, we can vary the number or complexity of features given for each potential recipient. We could present participants with pairs of patients who differ in three, six, or nine of their features. Or we could present many patients at once. If a participant places a great deal of weight on a feature when nine features of each participant are revealed or when

6. Data involving forbidden attributes (such as race, religion, gender, etc.) will be needed in order to detect the presence of bias and correct for it, so the system as a whole cannot be totally ignorant of these attributes, but it will use this information in a different way. See Kroll et al. (2016) for a discussion of this and related issues.

choosing among nine patients, but not when only three features of two patients are revealed, then this pattern provides some evidence that this participant's moral judgment results from the presence of the other features or patients. That might be because so many complex features of patients created confusion (though it might not be easy to distinguish confusion from context-dependence). If so, we can predict what the participant would say if she were not so confused.

Admittedly, these rough experimental suggestions are far from conclusive, and we have not tested all of them to see whether they would work. The point for now is just that nothing in principle stands in the way of gathering the information that an AI would need in order to predict the moral judgments that humans would make if they were not biased, ignorant, or confused. Such an AI could serve as a proxy for an ideal observer or at least evidence of how an ideal observer who is informed, rational, and impartial would make moral judgments and decisions in cases like these.

Goals

What does this accomplish? What does it not accomplish?

Our goal is *not* to create an AI to tell people what is *really* and *truly* moral or immoral. We do not assume, as some ideal observer theorists do, that moral judgments by idealized humans constitute moral wrongness or rightness. Ideal observers might be inaccurate in some cases, but they can still provide evidence of what is morally right or wrong. That is all we need in order for an AI that tells us how ideal observers would judge acts to be helpful to us in deciding what we should believe and do in complex moral situations.

Our goal is also *not* to replace doctors. We only want to help doctors like the kidney surgeon at 3:00 a.m. The doctor still has to decide. We would not want to hand the decision over to an AI completely (in the foreseeable future), but an AI can still help. Suppose the surgeon at 3:00 a.m. thinks that patient A should get the kidney but then runs an AI trained on the surgeon's own moral judgments in hundreds of conflicts, and the AI predicts what the surgeon would say in the absence of ignorance, confusion, or bias. If the AI agrees with the surgeon's moral judgment, then the surgeon is justified in being more confident than if the AI disagreed with her moral judgment. And if the AI disagrees with the surgeon's current moral judgment, then the surgeon would have reason to stop and reflect more and maybe seek help from others (such as a hospital ethics committee, if there is time). Doctors who must make difficult moral decisions under time pressure without adequate information should be grateful for such help.

Expanding the Scope: Artificial Improved Democracy (AID)

Doctors are not the only ones who need help. Such aids can be used in many different areas of morality, including law, military, business, personal life, and so on. In each area, we will need to:

(i) Ask the folk or experts to describe moral problems in that area,
(ii) Ask them which features are morally relevant,
(iii) Edit their features (for clarity, redundancy, and completeness),
(iv) Construct scenarios in which those edited features conflict,
(v) Ask which act is wrong in those conflicts,
(vi) Extract models for individuals, and
(vii) Learn or improve the model by applying it to new scenarios.

Then we could also correct those models for ignorance, confusion, and partiality in order to make them more ideal.

This method in general can be known as artificial improved democracy. It is *democracy* insofar as it rests on surveys of opinions from the general public. It *improves* on democracy by correcting for common errors to reveal what people really value and what they would judge if they were informed, rational, and impartial. And it is *artificial* because it is embodied in a computer program. Overall, it is AID (Artificial Improved Democracy), because its point is to *aid* people in making better moral judgments.

This method could potentially help a wide range of people avoid the most common sources of error in human moral judgment in a wide range of areas. Drivers use GPS to tell them where they ought to turn, because they know that they are ignorant and confused, and they often do and should trust the GPS even when their instincts point them in a different direction. This analogy makes us hopeful that people might be willing to use AI to help them avoid the common kinds of mistakes to which we are all prone. An AI that tells them which moral judgments they would make if they were more ideal could serve almost (though not exactly) like a conscience. They could correct their moral mistakes in light what they learned from the AI. The AI would thereby reduce the incidence of human moral mistakes both in judgment and in action.

Who does not want that? Some (Cave et al., 2019) fear that depending on a moral AI will erode humans' skills at making their own moral judgments and that moral mistakes by the AI will be hard for humans to detect. However, we think that these potential dangers are overblown, can be minimized, and, hence, are overridden by potential benefits of a moral AI in reducing moral mistakes by humans.

Differences

An AI that creates a model for each participant might also explain *differences* among individuals and groups. We could compare individual models to understand moral judgments by old friends whose moral judgments still remain mysterious. We could also learn *how many* and *which* other people would disagree with us. A doctor might like to know how many others in the hospital or the public would make a different moral judgment on a case. And if the models are interpretable, they could also tell us *why* we disagree with other individuals, because they will cite combinations of features that humans see as morally relevant.

In addition, we could aggregate models for individuals into models for *groups*. There are several competing ways to aggregate, and it is not clear which is best for these purposes (Brandt, Conitzer, & Endriss, 2016). Still, some group models might be able to help us understand, for example, why people in the US and the UK make different moral judgments about some issues but not others. The differences between their aggregate models can explain why they output different moral judgments when they do.

Applications in Medicine

Having laid out our general method and our test case with kidney exchanges, we want to conclude by suggesting that AID could also be applied to many other issues in bioethics. In the first place, it should be easy to see how our proposed methods could be extended to transplantations of other organs, such as heart, lung, liver, pancreas, etc.

We think AID could also be applied to other cases involving the allocation of scarce medical resources. The method could be adapted for issues such as life-sustaining treatments, experimental therapies, emergency medicine, end-of-life issues, or critical care. In all of these cases, questions about how to distribute equipment, treatments, funding, drugs, clinicians' time, etc., could be subjected to hybrid methods that combine some top-down principles (such as optimizing for the overall number of patients seen or caring for the sickest patients first) with some bottom-up principles learned from representative surveys of many different stakeholders. These findings could then be used to design systems to reflect both shared values and ideal decision procedures.

The application of our methods, however, need not be limited to cases involving scarce resources. In our discussion above, we also raised the prospect of correcting for various kinds of biases. This aspect is especially promising, given increasing scholarly interest in how various kinds of cognitive biases challenge foundational concepts in bioethics (e.g. Blumenthal-Barby, 2016).

To see this, consider informed consent, which is "currently treated as the core of bioethics" (Eyal, 2019). Very roughly, informed consent is crucial because disclosing necessary information to a patient or research participant (so that they can understand the information and voluntarily decide whether or not to undergo treatment or take part in research) is one of the best ways to ensure the protection of the patient's interests, health, and well-being, while also respecting them as autonomous individuals.

Various kinds of biases can undermine this foundational concept in many ways. We will focus here on just one kind, optimism biases, which have been extensively explored by Lynn Jansen and colleagues (e.g. Jansen, 2014; Jansen et al., 2018). Jansen (2014) asks us to consider the difficult decision of a cancer patient who is given the option to participate in an early-phase research trial. In cases like this, standard therapies have probably been ineffective for the patient. The probability of significant improvement from participating in an early-phase research trial is, almost by definition, very low. But when patients who decide to take part in such trials are asked to explain their decision, "they often reveal unrealistically high expectations for therapeutic benefit from participation" (Jansen 2014, 26). In other words, patients often overestimate the benefits they are likely to receive and underestimate the risks. The worry is that, to the extent that patients are mistaken about the likely outcomes, their consent to participate in the research is not truly *informed*.

Just as Marshall raised the question of whether people would still support the death penalty if they were not ignorant of the relevant evidence, we can raise the question here of whether patients would still agree to participate in early-phase research trials if they were not ignorant of the relevant probabilities of therapeutic benefit. Perhaps some patients would want to participate in the trial, no matter how low the probability. Perhaps other patients would change their mind about participating if they thought more carefully and accurately about the likely outcomes. Per our method, we treat this as an open empirical question which should make use of all of the tools at our disposal, including AI methods that might be able to shed light on difficult questions about the extent to which individual judgments relevant to informed consent are driven by various kinds of cognitive bias.

Conclusion

The foregoing should suffice to show the initial promise of the AID method for bioethics, ranging from issues as diverse as organ transplants to informed consent. While very much work remains to determine whether this initial promise will be fulfilled, it seems likely that, as the practice of medicine becomes

increasingly intertwined with Machine Learning and Artificial Intelligence, so too will the practices and methods of bioethics become similarly intertwined with these powerful technologies.

References

Anderson, M., and Anderson, S.L. (2011). A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists. In *Machine Ethics*, edited by M. Anderson and S.L. Anderson, pp. 476–492. Cambridge: Cambridge University Press.

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: CRC Press.

Asimov, I. (1950). Runaround. Reprinted in *I, Robot*. New York, Doubleday.

Beauchamp, T. L. and Childress, J. F. (2019). *Principles of Biomedical Ethics*, Eighth Edition. New York: Oxford University Press.

Blumenthal-Barby, J. S. (2016). Biases and Heuristics in Decision Making and Their Impact on Autonomy. *The American Journal of Bioethics*, 16 (5): 5–15.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brandt, F., Conitzer, V., and Endriss, U. (2016). Computational Social Choice. In *Multiagent Systems*, edited by G. Weiss, pp. 213–284. Cambridge: MIT Press.

Cave, S., Rune, N., Vold, K., and Weller, A. (2019). Motivations and Risks of Machine Ethics. *Proceedings of the IEEE*, 107 (3): 562–574.

Doyle, K., Schaich Borg, J., Sinnott-Armstrong, W., and Conitzer, V. (In progress). Survey of Morally Relevant Features of Kidney Recipients.

Driver, J. (2012). *Consequentialism*. Abingdon: Routledge.

Eyal, N. (2019). Informed Consent, *The Stanford Encyclopedia of Philosophy*. E.N. Zalta (ed.), https:// plato.stanford.edu/archives/spr2019/entries/informed-consent/ [accessed 24 November 2020].

Firth, R. (1951). Ethical Absolutism and the Ideal Observer. *Philosophy and Phenomenological Research* 12 (3): 317–45.

Gert, B. (2004). *Common Morality: Deciding What to Do*. New York: Oxford University Press.

Gert, B., Culver, C., and Clouser, K. D. (1997). *Bioethics: A Return to Fundamentals*. New York: Oxford University Press.

Gigerenzer, G. (2015). *Risk Savvy: How to Make Good Decisions*. London: Penguin.

Gips, J. (1995). "Towards the Ethical Robot", Second International Workshop on Human & Machine Cognition, Perdido Key, Florida, 1991. Published in *Android Epistemology*, K. Ford, C. Glymour, P. Hayes (eds.), MIT Press, 1995.

Guarini, M. (2006). Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems Magazine*, 21 (4): 22–28.

——— (2011). Computational Neural Modeling and the Philosophy of Ethics. In *Machine Ethics*, edited by M. Anderson and S.L. Anderson, pp.316–334. Cambridge: Cambridge University Press.

Horty, J. F. (2014). *Reasons as Defaults*. New York: Oxford University Press.

Jansen, L.A., Appelbaum, P.S., Klein, W., Weinstein, N. Mori, M., Degnin, C., & Sulmasy, D.P. (2018). Perceptions of Control and Unrealistic Optimism in Early Phase Cancer Trials. *Journal of Medical Ethics*, 44: 121–127.

Jansen, L. A. (2014). Mindsets, Informed Consent and Research. *The Hastings Center Report*, 44: 25–32.

Kant. I. (1797). "On a Supposed Right to Lie from Altruistic Motives." Prussian Academy Volume VIII; translation by Lewis White Beck in *Immanuel Kant: Critique of Practical Reason and Other Writings in Moral Philosophy*. Chicago: University of Chicago Press, 1949; reprint: New York: Garland Publishing Company, 1976.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable Algorithms. *University of Pennsylvania Law Review*, 165: 633.

Marshall, T. (1976). Dissenting Opinion in *Gregg v. Georgia*, 1976.

Mikhail, J. (2007). Universal Moral Grammar: Theory, Evidence, and the Future. *Trends in Cognitive Science,* 11 (4): 143–152.

Rawls, J. (1952). Outline of a Decision Procedure for Ethics. *Philosophical Review,* 60 (2): 177–197.

——— (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Ross, W. D. (1930). *The Right and the Good*. Oxford: Clarendon Press.

Roth, A. E., Sönmez, T., & Ünver, M. U. (2004). Kidney Exchange. *The Quarterly Journal of Economics*, 119 (2): 457–488.

Sarat, A., & Vidmar, N. (1976). Public Opinions, the Death Penalty, and the Eighth Amendment: Testing the Marshall Hypothesis. *Wisconsin Law Review*, 171–206.

Snedegar, J. (2017). *Contrastive Reasons*. New York: Oxford University Press.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2): 841–887.

Wallach, W., and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.