An Inquirer's Guide to Ethics in AI

Matthew S.W. Silk
and Ian J. MacDonald

broadview press

2.1 REFLECTIVE MORAL THINKING

In reflective moral thinking, to determine which option has greater moral worth, you need to make a value judgement. Value judgements involve consideration of two elements: ends and means. An **end** is a possible goal or settlement of your dilemma. In our previous example, we are faced with a choice between two ends: going on vacation or helping your cousin move. The **means** are the practical and logistical factors that are required to obtain an end. So, the means used to obtain the end of going on vacation might include the cab ride, the plane ticket, travellers' insurance, and so on, while your means of helping your cousin include possibly a truck and your own body. If an action is considered as an end, you need to consider whether you have the means to obtain it, and sometimes this can affect what you choose to value. If you had less money, for example, you might feel more hesitant about going on vacation in the first place, whereas if you had extra money you might choose to hire movers as means to obtain both of your ends: you go on vacation while your cousin gets the help they need.

We can now understand why the auto insurance example has a moral salience. The AI developer tasked with creating an algorithm to determine insurance rates uses credit scores as a means towards their end goal of creating a reliable insurance rate generator (Figure 1). However, the AI developer doesn't wish to negatively impact anyone by creating a system where people are discriminated against for arbitrary reasons. But the means they've chosen has now created a conflict between these two ends. Now the developer must either choose between one of those two ends or find a new means of satisfying both ends without creating further problems.

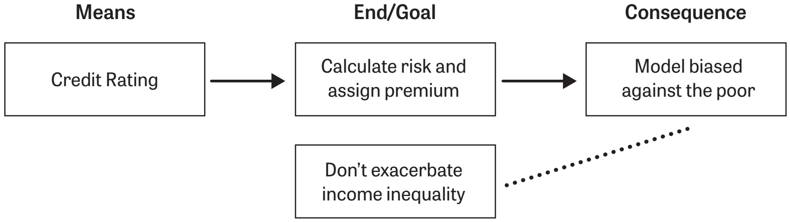
Means
End/Goal
Consequence
Credit Rating
Calculate risk and assign premium
Model biased against the poor
Don't exacerbate income inequality

FIGURE 1 • Using credit scores as a means helps bring about a reliable system for calculating insurance premiums, but it produces consequences that conflict with other goals.

This problem may be further compounded if the insurance company commissioning the algorithm insists on using credit scores, because they assume a higher-scoring applicant is less risky. It might also be easier to obtain credit scores compared to the alternative variables. These factors highlight the genuine moral tension that exists in this situation and that drives us to moral reflection and inquiry.

But how do we decide which ends are of greater moral worth? To appreciate the different perspectives, interests, goals, and logistics involved in AI development, we must inquire into situations and scenarios that call for our ethical attention. Thus, the aim of this text is to create a resource for moral inquirers to empower anyone to better study, understand, and prescribe reflective solutions to moral problems: an inquirer's toolbox. We've already learned that understanding the relationship between our end goals and the means we use to obtain those goals can help us understand how and why moral dilemmas come to be. We can add the first tool in this toolbox, which is the use of ends-means reasoning to help us evaluate the consequences of one proposed moral solution compared to another. Next, we will consider the potential use of moral principles in helping us choose among different ends.

- Analysis of ends-means relationships.

2.2 MORAL PRINCIPLES

A moral principle is an abstract general rule that can help determine which things are more morally important than other things. They tell us what we *ought* or *should* do. Sometimes these moral principles are general and might apply to anyone, such as "the golden rule" which asks us to treat others as we would prefer to be treated. Moral principles can also be aimed at specific kinds of cases. For example, a professional code of ethics might state that accountants must be honest with their client at all costs. The most well-known ethical principles come from prominent moral theories.

3. Solving Ethical Problems in AI: How Generalizable Is Ethics?

Moral principles are convenient because they provide definitive, systematic, and universal answers to moral questions. Relying on abstract universal moral principles would allow AI developers to efficiently anticipate and consistently respond to moral concerns that might arise from the use of their product because we can start to generalize over individual cases and pick up on common moral concerns.

We might imagine that reviewing ethical issues relating to artificial intelligence could be a streamlined process. In medical contexts, researchers developing new drugs or technology must submit their proposal to ethics review boards who can vet the proposal using ethical rules and principles and determine if it is ethically acceptable to proceed. Can we simply use ethics review boards to apply moral principles to AI proposals and determine if they are ethically acceptable, or does applying ethical principles require something more?

Approaches to practical ethics that focus on the application of moral theories are called “high theory.” Bioethics, the study of ethical issues emerging from biology and medicine, is strongly influenced by high

[8] Marino 2013, 739.

theory. However, people such as philosopher Stephen Toulmin argue against relying too heavily on moral principles. His 1981 paper, "The Tyranny of Principles," recounts his experience serving as a member of an ethics review board for biomedical research, where quite often board members were able to agree (sometimes unanimously) on recommendations but would disagree about what moral principle should be appealed to. Too often, Toulmin reports, ethical problems where solutions seemed obvious to most people quickly became less temperate, less discriminating, and less resolvable once the debate turned to "matters of principle." He explains that this gives rise to an "absoluteness of moral principles that is not balanced by a feeling for the complex problems of discrimination that arise when such principles are applied to particular real-life cases."[9]

Even if using a single theory, the answer may not be so cut and dried. For example, should we try to maximize utility by focussing on each act in its own situation (what is called **act utilitarianism**) or should we adopt rules that maximize utility overall (what is called **rule utilitarianism**)? The point is that so long as we believe that context should be a factor in deciding what is ethical, then ethics is going to require more than just thinking about principles.

The problem with focussing on high theory is that if we think that there may be morally relevant specific features in individual cases, a universal moral principle will gloss over those details, potentially making you neglect important moral factors. You may become so focussed on the principle that you believe is at stake, that you become less critically minded about how that principle should be applied and less attentive to other problems. As Toulmin notes, "Oversimplification is a temptation to which moral philosophers are not immune, despite all their admirable intellectual care and seriousness; and the abstract generalizations of theoretical ethics are … no substitute for a sound tradition in practical ethics."[10]

3.1 THE LIMITS OF MORAL PRINCIPLES AND AI

In his 2019 paper "Principles alone cannot guarantee ethical AI" Brent Mittelstadt considers the applicability of high theory to AI. He notes that there are already 84 initiatives by different groups articulating

[9] Toulmin 1981, 32.

[10] Toulmin 1981, 31.

ethical principles for AI who seek to translate these principles into more specific governance frameworks and professional ethics codes.[11] Most of these initiatives, such as the European Union's High-Level Expert Group on AI, have converged on principles such as fairness, prevention of harm, explicability, and respect for human autonomy.[12] These principles are supposed to embed normative consideration in technology design and governance and function like high-level principles in bioethics. Nevertheless, Mittelstadt notes important differences between bioethics and AI that make the use of principles more complicated.

Unlike in medicine, where there is a common recognized aim (to promote the health and well-being of the patient), AI development is largely driven by private sector aims such as cutting costs or increasing profit. The fundamental aims of developers, users, and affected parties are not the same. Because there is no patient whose interests would have the highest priority, there is also no relationship of trust. A doctor acting on a patient's behalf for the patient's best interest is involved in a relationship of trust—that is, has certain **fiduciary responsibilities** to that patient. AI developers, by contrast, have no recognized fiduciary responsibilities to the public or to public interests over private ones. This makes it more difficult to develop commonly understood and agreed upon sets of principles to govern the field.

In addition, AI development lacks a long history of professional development or a singular professional culture. There is no sense of what makes a "good" AI developer when compared to a profession like medicine, which has a wealth of standards for what would constitute a "good" doctor. Such professional cultures, "provide a historically sensitive account of the obligations of the profession against which negligent content and practices can be identified."[13] They make it easier for high theory to work because the principles provide a common language to deal with practices where ethical necessity is recognized.

AI development has no common professional culture. Creating ethical frameworks that cover all ethical concerns is challenging when "the impact of decisions taken in designing, training, and configuring AI systems for different uses may never become apparent to developers."[14]

[11] Mittelstadt 2019, 501.

[12] European Commission 2019.

[13] Mittelstadt 2019, 502.

[14] Mittelstadt 2019, 503.

Often in AI development no single person has a full understanding of the system's functions and being unable to understand the impact of ethical decisions makes it difficult to create ethical norms for the profession.

Broad principles like "fairness" or "respect for human dignity" are often too abstract to provide meaningful guidance. What one developer might consider "fair" another might consider grossly "unfair." As Mittelstadt notes, "statements reliant on vague normative concepts hide points of political and ethical conflict. 'Fairness', 'dignity' and other such abstract concepts … have as many possible conflicting meanings requiring contextual interpretation through one's background political and philosophical beliefs."[15]

As Mittelstadt explains, "Professional societies and boards, ethics review committees, accreditation and licensing schemes, peer self-governance, codes of conduct, and other mechanisms supported by strong institutions help determine the ethical acceptability of day-to-day practice by assessing difficult cases, identifying negligent behaviour, and sanctioning bad actors."[16] These institutions help translate principles into practice by studying, testing, and revising recommendations and norms over time by considering different cases. The field of AI development lacks similar institutions to translate principles into practice. Translation involves "the specification of high-level principles into mid-level norms and low-level requirements," although "norms and requirements cannot be deduced directly from mid-level principles without accounting for specific elements of the technology, application, context of use, or relevant local norms."[17] As we will discuss in future chapters, for example, it can be difficult to articulate the meaning of privacy as a high-level concept and then translate that into practical guidelines for AI development, and even more difficult to evaluate those guidelines given the potential risks and trade-offs when they are put into practical use. So, to return to our earlier question, clearly all the ethical problems of AI cannot be solved by simply relying on ethics review boards that apply broad ethical principles.

[15] Mittelstadt 2019, 503.
[16] Mittelstadt 2019, 503.
[17] Mittelstadt 2019, 503.

5. Negligence and Recklessness

Bridgman's argument supports the notion that applied science and pure science are distinct and that when engaged in pure science, a scientist should keep value judgements out of their research. But in 1953 philosopher Richard Rudner published an article titled "The Scientist Qua Scientist Makes Value Judgments." He argues that scientists make value judgements when it comes to deciding whether to accept a hypothesis. He explains: "no scientific hypothesis is ever completely verified"; thus, "in accepting a hypothesis the scientist must make the decision that the evidence is sufficiently strong or that the probability is sufficiently high to warrant the acceptance of the hypothesis."[47]

Imagine that a scientist tests a drug for safety and knows there's a chance that their test could be wrong. Even if the test shows evidence that the drug is safe, there's still a chance that the test is wrong. How sure should you be before declaring that the drug is safe? As Rudner explains, "our decision regarding the evidence and respecting how strong is 'strong enough', is going to be a function of the importance, in the typically ethical sense, of making a mistake in accepting or rejecting a hypothesis … How sure we need to be before we accept a hypothesis will depend on how serious a mistake would be."[48] This is an example of **inductive risk**, or the risk of error in accepting or rejecting a hypothesis. Since it is the scientist who makes this decision, and since this decision is ethical in scope, then clearly the scientist must be held responsible for these value judgements.

But why should scientists care about inductive risk? Can't they simply ignore such concerns? According to philosopher of science Heather Douglas, the answer is no. Inductive risk considerations are questions about the consequences of error and they stem from considerations about recklessness and negligence. According to Douglas, individuals should be held ethically accountable for not being reckless or negligent

[47] Rudner 1953, 2
[48] Rudner 1953, 2.

unless there is a reason for a special moral exemption. She argues that scientists do not deserve moral exemptions from being reckless or negligent and thus are as responsible as the rest of us for considering the consequences of their errors.

Moral responsibility extends to things we intend to do, but also certain actions we don't intend. If a driver doesn't regularly inspect their vehicle and its wheel comes off and causes an accident, we say that the driver is morally responsible even if they did not intend that result. **Negligence** occurs when a person takes an action not knowing that they are risking harm, but they should be aware of the risk. Alternatively, **recklessness** occurs when someone knows the risks of harm that their actions could cause but chooses to do them anyway.[49]

In the early twentieth century, mechanical engineer Thomas Midgley Jr. developed a solution to a problem that plagued many early automobiles called "knocking" by adding a lead-based compound called tetraethyllead to gasoline. Midgley worked for General Motors and knew that his discovery would make a lot of money. The problem is that lead was well known, even as early as 1920, for its toxic effects, causing hallucinations, insanity, and death.[50] Despite this, and although workers at the factories producing the fuel (and even Midgley himself) became sick with lead poisoning, they proceeded to sell it and would insist on the safety of leaded gasoline for decades.

Eventually some workers began to die. In 1924 at a press conference Midgley attempted to demonstrate how "safe" the gasoline was by inhaling its vapors for a whole minute.[51] Midgley, who knew the risks, always advised the public that it was safe and recklessly insisted on the product. For decades lead was burned in gasoline and released into the atmosphere, until leaded gasoline was phased out in the 1970s. However, as a result of the amount of lead released into the atmosphere, everyone was exposed to higher concentrations of lead, resulting in higher crime rates, deaths, and a general decline in human IQ levels.[52]

Midgley's other major contribution was the development of Freon gas, which is used as a refrigerant. Prior refrigerants were flammable or toxic, but Freon is nontoxic and was eventually used in refrigerators and

[49] Douglas 2010, 68–72.

[50] Markowitz and Rosner 2013, 43.

[51] Markowitz and Rosner 2013, 22.

[52] McFarland, Hauer, and Reuben 2022.

in aerosol products. The problem is that Freon is a kind of chlorofluorocarbon (CFC) and it was later discovered that this gas goes into the atmosphere, where it remains and occasionally releases chlorine atoms which react with ozone, breaking it apart. Midgley's invention and other CFCs were discovered to be destroying Earth's protective ozone layer. CFCs were eventually phased out as well. In this case we might say that a reasonable person should have been aware of the major side effects of such a widely used product, and that because Midgley did not test this first, he was negligent in the development of Freon. For these two reasons, environmental historian J.R. McNeill has said that Midgley "had the most adverse impact on the atmosphere than any other single organism in Earth's history."[53]

These two cases demonstrate how a scientist can be reckless or negligent and what can be the consequences. One benefit of understanding moral responsibility in terms of recklessness or negligence is that these concepts do not rely on the notion of perfect foreknowledge of the future. No one expects the negligent driver to predict the future, but certain consequences are reasonably foreseeable, and in these cases we expect the person to do due diligence and consider the consequences that we could reasonably predict.

Bridgman might reply that ultimately the choice to accept a hypothesis is a matter of applied science, a choice for industrialists or policy makers. Scientists could advise whether something is likely or not likely to be the case, but the decision to proceed is not the scientists'; hence they are not responsible for what happens. And to maintain the integrity of pure research, scientists should be exempted from such ethical considerations.

Douglas believes this argument doesn't work. In some cases, it is acceptable to exempt certain people from certain kinds of moral responsibilities because of the role they are fulfilling. For example, a defence lawyer who is aware of their client's criminal wrongdoing is not obligated to report it because of lawyer-client confidentiality.[54] Douglas argues that scientists shouldn't be permitted similar exemptions.

[53] McNeill 2001, 421.

[54] Douglas 2010, 73.

1.2.1 Modelling Assumptions

To better understand the issue of modelling and its ethical implications, imagine you have been asked to design an algorithm that can predict if someone has clinical depression. Machine learning requires lots of training data, so what metrics would you use as data to teach it to do this? According to the American Psychiatric Association, symptoms of depression might include severe feelings of sadness, a loss of interest or pleasure in activities once enjoyed, changes in appetite, trouble sleeping, loss of energy, feelings of worthlessness, difficulty thinking or concentrating, and so on. Despite the lengthy list of symptoms, there is no single unifying biological cause. For some patients there might be decreased activity in a neurotransmitter, while for others the symptoms might be caused by overactivity in the hormonal system.[15]

In other words, while we might use known symptoms of clinical depression to help determine if someone has clinical depression, it isn't something that we can directly measure on a biological level. If we want

[12] Kearns and Roth 2020, 9.

[13] O'Neil 2016, 18.

[14] O'Neil 2016, 18.

[15] Nemeroff 1988, 44–48.

to measure if someone has clinical depression, we must do so indirectly based on the empirically measurable symptoms that we think their depression is causing. If we could use these symptoms as metrics, we might be able to create a model that could predict if someone had clinical depression.

The process of defining the empirical measure of an abstract concept that is not directly observable is called **operationalization**. The electron, for example, is not something that we can directly measure or see, so how can it be empirically detected? The electron was discovered when positively and negatively charged plates were applied to a cathode ray. From the observable deflection of the rays, it was inferred that the particle had carried unobservable negative particles. Today, we can use wire chambers that detect subatomic particles and provide information on their trajectory by tracking the trails of gaseous ionisation. Such practices thus partially define how such unobservable particles can be operationalized and measured.

How should a sociologist operationalize a concept like human happiness? We can't directly measure everyone's happiness levels, so we might define empirical metrics that we can use. We could conduct a survey where we ask questions about happiness and compile those responses into a happiness index. But what if we can't reasonably expect to send out and get responses to surveys from that many people? In statistics, a researcher might try to substitute one measurement we would prefer (because they are direct or at least more direct) for a **proxy**. For example, to measure happiness we might create a statistical index based on well-known and publicly available data such as economic growth, poverty rates, or crime.

All of this raises an important question. What justifies us operationalizing a concept in a certain way or using certain proxies as stand-ins for the thing we are trying to measure? What justifies the connections we make between certain concepts and the phenomena that we take as evidence for those concepts for the purposes of making an inference? To answer this and gain insight into the ethical issues at stake, we can consider philosopher Helen Longino's account of evidential reasoning.

According to Longino, "states of affairs … do not carry labels indicating that for which they are evidence or for which they can be taken as evidence."[16] If a child presents with red spots on their stomach we might

[16] Longino 1990, 40.

take that as evidence that the child has chicken pox, but it is also possible that it is evidence of measles. The spots by themselves don't reveal which. According to Longino, "states of affairs are taken as evidence in light of regularities discovered, believed, or assumed to hold … What explains why I come to believe [the child] has the measles rather than that, say, the moon is blue, is some belief that I have about the relationship between having a red-spotted stomach and having the measles."[17] These beliefs are what Longino calls background beliefs or **background assumptions**.

Background assumptions represent a regularity that allows us to take something as evidence for something else. If there was a large plume of smoke in the distance, you might infer the presence of a fire. Why? The background assumption "where there's smoke, there's fire" is likely operating as part of your inference. Longino points out that background assumptions can have all sorts of effects on the ways that we might take something as evidence of something else. Different background assumptions might enable us to reach the same conclusion, but for different reasons. One background assumption might have us pay attention to colour, while another might focus on location yet reach the same conclusion.[18] Alternatively, different background assumptions might lead us to reach different conclusions about the same state of affairs. Different aspects of the situation can be taken as evidence for competing conclusions. We thus need to consider what justification each side might have for their competing background assumptions.

2. Am I defining this problem too broadly or too narrowly?
7. How would I test any assumptions I have regarding the nature of the problem or a hypothetical solution?

Background assumptions connect a concept to the phenomena taken as evidence for it, and this means that the way we operationalize a concept or the kinds of proxies that we use as evidence for a concept we are measuring will be justified given whatever background assumptions we hold to be true or relevant. Sometimes these background assumptions can be directly confirmed.[19] For example, Kepler's laws of planetary motion hold that planets orbit in an ellipse. This generalization by Johannes Kepler

[17] Longino 1990, 41.
[18] Longino 1990, 42.
[19] Longino 1990, 49.

was based on direct observations of planetary orbits made by Tycho Brahe in the sixteenth century.[20] However, as we have discussed, direct confirmation of a background belief isn't always possible.

If we can't directly confirm every background assumption that we might make use of, then we might believe them for bad reasons. As Longino notes, unless every background belief is directly confirmed then there seems to be no clear way to shield our reasoning about evidence from social and individual values and subjective preferences.[21] While we might be tempted to limit the scope of science and statistical reasoning to what we can directly confirm to bolster our certainty in the results, this would severely limit scientific investigation. It would mean, for example, that physicists would not be able to talk about unobservable particles. Statisticians would be limited in their ability to use proxies.[22]

We must be conscious of how science is done, rather than how it might ideally be done. Longino explains: "When a theory is being developed, the criterion for inclusion of specific hypotheses or principles is not that they are directly confirmed … but that they are relevant to the explanation of the phenomena comprehended by the theory."[23] Notice that this attitude is similar to the account of inquiry presented in Chapter 1, which describes how thinking about information not only involves judgements about accuracy but also relevance to the problem at hand.

If we cannot directly empirically confirm every background assumption that we might use, then values and subjective preferences will play a role in determining what kinds of background assumptions you use. This will determine what kinds of evidence that you take as making a convincing case for something. Background assumptions may also govern how you understand the problem you are creating the algorithm for; they will affect how that evidence needs to be framed, how you will test your ideas and identify "success," and how you will deploy your results.[24] Thus, if an inquirer's background assumptions reflect biased attitudes, the models they create will reflect those biases as well. Therefore, considerations about background assumption are also ethical considerations.

[20] Duhem 1954, 191.
[21] Longino 1990, 48.
[22] Longino 1990, 49–51.
[23] Longino 1990, 51.
[24] Biddle 2020, 4.

1.2.2 Values and Assumptions

Reconsider our attempt to operationalize human happiness. A statistician, unable to directly measure human happiness, might choose proxies such as economic growth, the amount of poverty, and the level of crime to create an index. What background assumptions are at play? One might be that "If an economy is growing, people will be happier" or that "If crime increases, people will be less happy." Notice two things about these assumptions. The first is that these assumptions might reflect capitalist values. For example, one assumes that people will be happier if they are wealthier or that economic growth translates into happiness. Secondly, these assumptions may not always be true; greater misery can sometimes follow economic growth. If we conceive of and measure happiness in this way, we may miss out on other aspects of happiness that we aren't measuring. This can lead to conclusions that reflect biases implicit in the background assumptions we have used.

We may sometimes want our values to influence our thinking. To understand this, reconsider making a model to detect clinical depression. We've discussed the fact that there is no single biological measure of depression, and that depression is measured according to a series of symptoms. One potentially relevant background assumption is whether we should understand it as a brain disorder. However, some researchers have argued that mental disorders should not be understood purely as brain disorders and that attempting to find correlations between symptoms and brain states to explain these symptoms is problematic. Instead, they argue that we should understand mental illnesses as a network of symptoms that interact with each other without assuming common biological causes.[25]

The central point is that if we assumed that we could operationalize clinical depression by picking a list of symptoms that correlate with brain states we think are associated with depression, this would only be justified given background assumptions about the relationship between mental health and biological events which are not directly verified. But what if we give up the assumption that there is a biological cause? This introduces yet another question. As philosopher Kristen Intemann asks, "if clinical depression is defined in terms of its symptoms, what criteria

[25] Borsboom, Cramer, and Kalis 2019, 3–4.

should be used to distinguish clinical depression from the common blues?"[26]

Another way to understand Intemann's question is this: Without a single biological cause and given that patients suffer from a wide range of symptoms, how do we even know we are talking about a single illness? Intemann's answer is that what is common to all the symptoms of clinical depression is that they impair the functions most essential to human well-being such as eating, sleeping, and engaging in relationships. Depression thus impairs functions that are vital to a good life for humans. But what counts as a good life for humans is the value judgement that eating, sleeping, and engaging in relationships are central to a good human life. As she argues, "this value judgement (if justified) also justifies us in grouping together symptoms and cases of clinical depression. If we did not rely on this sort of value judgement, then cases of clinical depression would look naturalistically gerrymandered and we would be less justified in treating them as one disease."[27]

In other words, to operationalize a concept like clinical depression, we would do so given background assumptions that are themselves partially justified by value judgements. Such value judgements impact what we take as evidence for the concept, and thus whether we will identify the concept in the world and how. For example, to identify symptoms, a doctor might have to evaluate how much sleep is "excessive" or what counts as "abnormal" indecisiveness, or "appropriate" forms of guilt.[28] It should be noted that when values play this role in evidential thinking, the results will not always be ethically bad. In the case of clinical depression, for example, we might agree with Intemann's position that such a value judgement is justified because we do think that depression is harmful to human flourishing. Nevertheless, a background assumption could be invoked carelessly given values that reflect bias and prejudice.

Unfortunately, many harmful biases go undetected. The problem with background assumptions is that we aren't always conscious that we are using them. For example, if you check your speedometer in your car, you might take the dial as evidence that your car is moving at a certain speed. You may not consciously realize it, but such an inference only

[26] Intemann 2001, S508.
[27] Intemann 2001, S509.
[28] Intemann 2001, S509.

makes sense if we assume various tenets of electromagnetism are true. The problem with commonly shared assumptions that sit in the background of our inferences is that they become invisible.

As Longino explains, when background assumptions "are shared by all members of a community, they acquire an invisibility that renders them unavailable for criticism. They do not become visible until individuals who do not share the community's assumptions can provide alternative explanations of the phenomena without those assumptions."[29] This is true in computer science; as Timnit Gebru explains, "The predominant thought that scientists are 'objective' clouds them from being self-critical and analyzing what predominant discriminatory view of the day they could be encoding, or what goal they are helping advance."[30]

1.2.3 Assumptions and Recidivism

This has been a long explanation of how values can become infused in a model and how it can lead to bias. A model requires that we understand some relationship between phenomena that we are using as evidence or data, and predicted outputs that we take to be examples of the thing we are measuring (Figure 8). Sometimes background assumptions are adopted for empirical reasons based on observed regularities, and sometimes they are based on practical reasons, such as an inability to directly measure something or a lack of resources permitting a better measurement. In machine learning, for example, operationalizing any concept means that you need large amounts of data that you can access for training purposes, and so you might be inclined to use proxies that are more public and accessible over other proxies that might be more relevant but less accessible.

As philosopher Justin Biddle explains, ML systems "are value laden in ways similar to human decision making because the development and design of ML systems requires human decisions that involve tradeoffs that reflect values."[31]

Let's return to the topic of predictive recidivism algorithms and why they generate bias. Recidivism is the tendency of a convicted criminal to reoffend. So, how would we operationalize the concept of recidivism

[29] Longino 1990, 80.
[30] Gebru 2020, 253.
[31] Biddle 2022, 2.

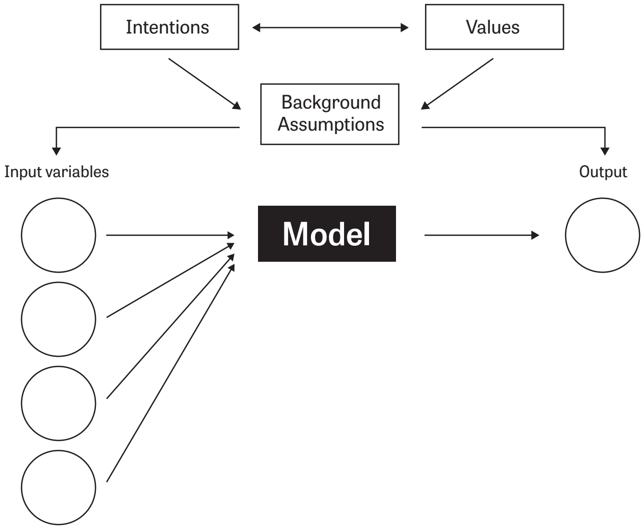

FIGURE 8 • The coherence of the model depends on background assumptions that justify the relationship between variables used as input data and what seems like an appropriate output from the model. The intentions of the developer and their values will influence what background assumptions they adopt. A background assumption may justify using certain variables as proxies for other properties or explain the meaning of an output value.

and how do we know if this will create a bias? You might think the most obvious predictor of recidivism is past criminal history as a safe background assumption. If someone has an extensive criminal history of repeat offences, they are more likely to offend again. But a single piece of information like this is hardly definitive. Many people have a past criminal history and don't reoffend, so if we rely on this intuition alone, we will label people as recidivists when we shouldn't. Some models, such as the "Level of Service Inventory-Revised" model, include lengthy questionnaires for the prisoner to fill out, determining attributes taken to be relevant to predicting if someone will reoffend.

If you were creating such a questionnaire, what questions would you ask? What things would you look for? Would you ask about the role that drugs and alcohol played in the crime? What background assumption are you relying on? As Cathy O'Neil points out, "Ask a criminal who grew up in comfortable suburbs about 'the first time you were ever involved with the police,' and he might not have a single incident report … Young black males, by contrast, are likely to have been stopped by police dozens of times, even when they've done nothing wrong."[32] People of colour are far more likely, for example, to have police stop-and-frisk them than white people. If early 'involvement' with the police signals recidivism, "poor people and racial minorities look far riskier."[33]

The questionnaire also asks if friends have criminal records. If you are from a middle-class neighbourhood, the answer is likely going to be very different than if you are from a poorer neighbourhood. What this means is that if you think that having prior involvement with the police or having friends with criminal records (independent of context) is indicative of recidivism, then any algorithm based on those assumptions will naturally generate biased results. An algorithm fed with historical crime data "will pick out the patterns associated with crime. But those patterns are statistical correlations—nowhere near the same as causations. If an algorithm found, for example, that low income was correlated with high recidivism, it would leave you none the wiser about whether low income actually caused crime."[34]

Remember Longino's claim that states of affairs do not carry labels. The fact that you have friends who have a criminal record might be evidence that you are a recidivist, but it might also be evidence that you are poor. The fact that you have had prior contacts with police might mean that you will commit more crime, but it also might mean that police like to target people who look like you. If you are a prisoner, the algorithm is not evaluating whether you are likely going to be a recidivist, it is evaluating whether people it thinks are like you are likely going to be a recidivist according to whatever assumptions the creators had about recidivism. This makes it easier to understand how any stereotypes we have about recidivists can be transferred to an algorithm that will find those same stereotypical correlations.

The idea that a recidivism prediction algorithm must be "fair" also requires background assumptions that reflect certain values. What does it mean to be fair and how can we capture this concept in a statistically measurable way? At least two forms (there are many more) of "fairness" are represented by the concepts of "predictive parity" or "equalized odds." An algorithm will have predictive parity if it generates the same rate of true predictions as a fraction of all positive predictions regardless

[32] O'Neil 2016, 25.
[33] O'Neil 2016, 26.
[34] Hao 2019.

of race. Equalized odds on the other hand holds that an algorithm is fair if it isn't more likely to generate false predictions for one group rather than another. Unfortunately, in some cases it is mathematically impossible to have an algorithm that satisfies both equalized odds and predictive parity.[35] Thus, it will be left to machine learning designers to judge which is the more appropriate form of "fairness."

Debates about competing definitions of fairness are important for how we evaluate real world use cases. For example, in 2016 an investigative report from ProPublica on the use of COMPAS revealed that the algorithm is biased against Black people. They found that Black defendants were 45% more likely to be assigned higher risk scores than white defendants after controlling for age, gender, and prior crimes.[36] In response, the corporation that owns COMPAS (now Equivant) argued that their algorithm is fair because it satisfies the criterion of predictive parity even though the ProPublica investigation used an equalized odds standard.[37]

Values and background assumptions can also affect how we understand the results. For example, the COMPAS algorithm generates probability scores of 1 to 10 where scores of 1–4 are considered "low," 5–7 are considered "medium," and 8–10 are "high." Declaring that what it means to have a "low" or a "high" score also reflects values about where we believe cut offs are for risk categories and how we think people with different scores should be treated. One of the background assumptions at work here is the idea that if someone has a higher score, they are likely to be a recidivist, and thus it is best that they remain behind bars longer.

If we reconsider what seemed like an obvious indicator of recidivism, a defendant's criminal history, in greater detail, we can understand how this can reflect biased values as well. If we use prior arrests to determine recidivism, then any biases that exist in pre-existing arresting procedures will bias the data as well. This is because "using arrests as a proxy for recommitting a crime means the algorithm is codifying biases in arrests, such as police officer bias to arrest more people of color or to patrol more heavily in poor neighborhoods."[38]

[35] Chouldechova 2017.

[36] Larson et al. 2016.

[37] Biddle 2020, 14.

[38] Ito 2019.

1.2 THE ETHICAL SIGNIFICANCE OF OPACITY

Understanding the ethical problems of opacity is a matter of understanding the ways in which AI can be said to be opaque, with respect to both the consequences that AI will produce and in the motivations that governed its development and usage. Opacity can take on different ethical meanings depending on what is lacking in transparency and on whether we are on the receiving end of the consequences of AI or participating in its development.

We can think of machine learning as being part of a system that includes various agents interacting with the model, the input data, and those who are affected by the output choices. This is called a **machine learning ecosystem**, and it includes stakeholders such as *creators and developers* who create machine learning models, *operators* who provide input and receive outputs from the model, *executors* who carry out decisions informed by the model, *decision-subjects* (or end-users) who are affected by decisions made by executors, *data-subjects* whose personal data is used to train the model, and *regulators* who are responsible for investigating and regulating the machine learning ecosystem.[16]

According to Carlos Zednik, "opacity," "transparency," and "explanation" will mean different things to different stakeholders in the machine learning ecosystem depending on what they consider to be epistemically relevant.[17] For example, a decision-subject will likely be interested in explanations that cite reasons for a model's output. A regulator may be more interested in understanding the environmental regularities that are being tracked by a model and whether it is processing data fairly or is violating privacy rights. A creator might be more interested in explanations that allow for intervention with the model to produce specific kinds of outputs.

We can use a variation of our diagram from Chapter 3 to represent the morally salient features of the machine learning ecosystem, such as the possibility that the model might get the wrong answer, and to understand which features are more important to different stakeholders (Figure 15). While creators will be responsible for the intentions and background assumptions that go into a model, decision-subjects will be the ones experiencing the morally salient consequences of the model's output.

[16] Tomsett et al. 2018, 9.

[17] Zednik 2021, 269.

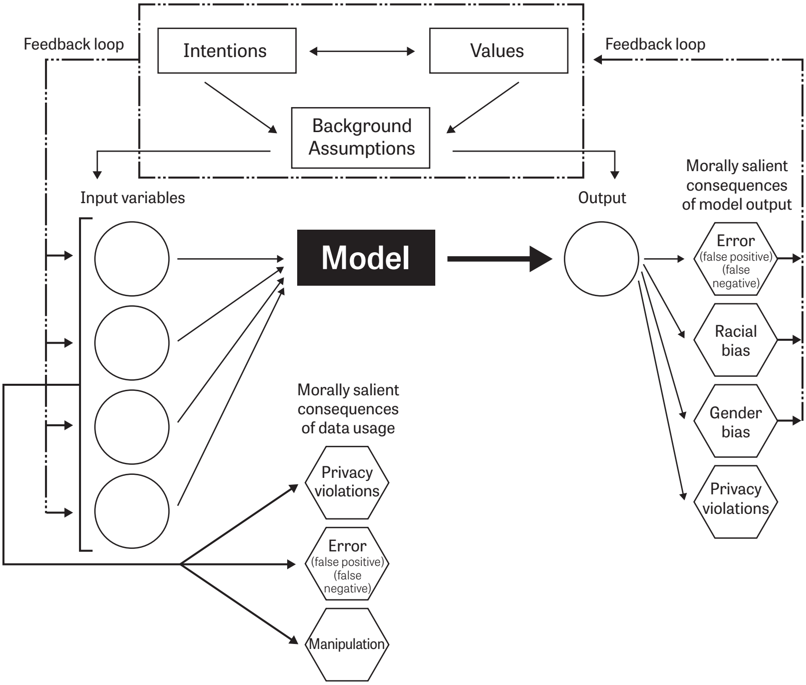
Feedback loop
Intentions
Values
Feedback loop
Background Assumptions
Input variables
Output
Morally salient consequences of model output
Model
Error (false positive) (false negative)
Racial bias
Morally salient consequences of data usage
Gender bias
Privacy violations
Privacy violations
Error (false positive) (false negative)
Manipulation

FIGURE 15 • This diagram represents the morally salient features of a machine learning ecosystem, including the morally salient consequences that follow from the collection and use of data, the assumptions that justify the usage of a model, and some of the morally salient consequences of the output of that model. While data-subjects may have to deal with the morally salient consequences of their data being collected, and decision-subjects or end-users must deal with the consequences of a model's output, creators and developers must concern themselves with the intentions and background assumptions that justify the model, and to an extent the other remaining morally salient features and consequences of the ecosystem.

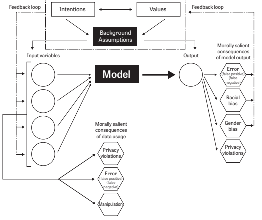
Feedback loop
Intentions
Values
Feedback loop
Background Assumptions
Input variables
Output
Morally salient consequences of model output
Model
Error (false positive) (false negative)
Racial bias
Morally salient consequences of data usage
Gender bias
Privacy violations
Privacy violations
Error (false positive) (false negative)
Manipulation

FIGURE 16 • From the perspective of creators and developers, the black box characteristics of the model make it opaque to them. Commonly held background assumptions will also not be transparent to them.

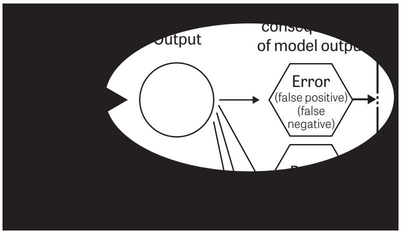

FIGURE 17 • From the perspective of decision-subjects, the only aspect of the ecosystem they will be aware of is the model's output and the morally salient consequences they may have to experience as a result. To them, almost the entirety of the ecosystem is a black box.

First, consider how the development and use of AI could be lacking in transparency. The black box problem makes the decisions of algorithms difficult to explain. However, recall from the previous chapter that a model developed from AI is built on background assumptions informed by our values and intentions. When background assumptions are shared by all members of a community, they acquire an invisibility that shields them from criticism. A background assumption may be required to justify a model developed by AI, yet it may not be apparent to the developer. Thus, commonly accepted background assumptions can represent another form of opacity (Figure 16).

Another form of opacity exists if we don't understand what data is being used for training and which variables are factored into the model. For example, if you are applying for a loan, you might not know that the algorithm is using your address as part of an e-score to gauge your reliability. The intentions of a developer might also be opaque. For example, a corporation developing a facial recognition algorithm might not disclose that its purpose is to target protestors. A teaching evaluation algorithm might be designed with the intention of busting teachers' unions. A hiring algorithm might not be trying to identify the best candidate, but to exclude as many applications as cheaply as possible.

Decision-subjects and the developers don't share the same understandings, yet the fact that the decision-subject presumably faces the brunt of the consequences of AI outputs means that opacity will affect them differently. They simply do not understand the algorithm, yet it affects them more (Figure 17). As Zednik notes, "Practically, end users are less likely to trust and cede control to machines whose workings they do not understand."[18]

The level of trust that we can reasonably expect the end-user to have will be different from that of the developers and this will produce ethical consequences. For example, if you don't know where the data used to train the AI comes from, if you don't know how representative it is, then it makes it difficult for the end-user to assess how generalizable its conclusions are. If you don't understand what the algorithm is really trying to measure, you can't even say for sure if it is really making an error. Because of this difference and because the end-user knows even less, their moral responsibilities, when it comes to responding to and

[18] Zednik 2021, 266.

meaningless patterns and correlations that appear to generalize well, yet because of the black box nature of the model, the developer wouldn't be aware of this.

3.4 INDUCTIVE RISK AND THE BLACK BOX

Let's recap. A machine learning developer can follow a process whereby they have a good understanding of the algorithm that was used to find the model, but not the model itself. We understand the inputs and the outputs of the model, but we don't understand what patterns the model is relying on to yield accurate predictions. Despite this, the process that finds these models might pick up on correlations that appear to yield accurate predictions yet mean nothing. The developer can appraise the model in terms of its predictive accuracy using training data and later test data, but they don't know which patterns the model is specifically looking for. This raises a distinctive ethical problem concerning inductive risk in the face of opacity.

As Kearns and Roth note, "Algorithms—especially models derived directly from data via machine learning—are different. They are different both because we allow them a significant amount of agency to make decisions without human intervention and because they are often so complex and opaque that even their designers cannot anticipate how they will behave in many situations."[53] This is especially the case when the input data is complex and the space of possible models is very large.

This brings us to the central problem. A machine-learning-derived model might rely on correlations that seem statistically significant to derive seemingly accurate predictions, yet these correlations are meaningless. Despite this, we are relatively unaware of what correlations the model considers significant, or how significant each correlation is for the model in terms of explaining its prediction. We are unaware of these correlations and their relative importance in comparison to other correlations for finding the answer we are looking for, yet these findings could all be subject to errors that might be ethically significant consequences. Normally, we would expect a developer to be responsible for inductive risk concerns such as these, yet due to black box opacity, they cannot manage these inductive risks.

[53] Kearns and Roth 2020, 7–10.

This contributes to what Ryan Felder calls an "accountability gap" in cases where decision-subjects are concerned. An accountability gap exists when there is a vacuum created when a task is taken out of human hands and put into the hands of a machine.[54] The opacity of the model means that a developer or executor will be unaware of the statistical relationships that the model is looking for and hence unaware of whether there is good evidence for relying on them and their relative importance for producing a predicted output. Thus, not only is there a lack of accountability for someone like an operator who might be expected to explain the answer of a model to a decision-subject even though they don't understand the model, but there is a lack of accountability for the inductive risks that go into the model construction itself.

Let's consider a peculiar thought experiment. Imagine that we created an algorithm that could detect certain forms of cancer from a medical imaging device. The hope is that we could create a data set of image scans from this device and use machine learning to create a model that can determine which scans are cancerous and which ones are not. Ideally, our neural network will create a model that will generalize such that when we expose it to new scans it was not previously trained with, it will detect cancer accurately.

Now let's imagine that unbeknownst to us, the process involved in producing these medical image scans produces an artifact on the image that coincidentally happens to be present every time there is cancer and absent when there isn't cancer, despite the reason for the artifact's creation being unrelated to cancer. It might be so small and insignificant (like the size of a pixel) that we may not even realize it is there. The ANN producing the model picks up on this correlation and uses it to predict cancer; however, due to the opaque nature of the model, we are unaware of this.

So long as the model is exposed to scans produced by the same process, it would not only not be caught as a training error, but if the test data contains the same artifacts, it will not be caught as a test error either. If that unknown feature is present in the scan, the model will generate an accurate prediction. On this basis, the model is accepted and put into use in medical practices. Now imagine that we eventually change the process that produces those image scans and the artifact is no longer present. We wouldn't realize that there is a change, but without

[54] Felder 2021, 40.

that key feature, the model will start producing errors and we wouldn't even know it because we trust the model but don't know it was looking for an irrelevant feature in the first place.

A thought experiment like this reveals the ethical risks involved when machine learning can pick up on correlational relationships that can seem to yield accurate results, yet in reality be meaningless. Moreover, the developer might not know about this and it wouldn't be revealed as training error or test error. If we accept the model and put it into widespread use and then later these features change, we may not realize that the model might start producing false negatives. The morally salient consequences of these errors might not be realized until a long pattern of errors is detected (which may not be believed if we trust the model). This makes it incredibly difficult to manage the ethical risks of error.

This thought experiment is a bit far-fetched since it assumes a purely coincidental artifact will affect the model construction in a very specific way. Is there evidence that real algorithms rely on useless information to seemingly provide accurate predictions in this way? According to Daphne Koller of Stanford University, such cases can happen. She explains:

> Imagine that you're trying to predict features from X-ray images in data from multiple hospitals. If you're not careful, the algorithm will learn to recognize which hospital generated the image. Some X-ray machines have different characteristics in the image they produce than other machines, and some hospitals have a much larger percentage of fractures than others. And so, you could actually learn to predict fractures pretty well on the data set that you were given simply by recognizing which hospital did the scan, without even actually looking at the bone. The algorithm is doing something that appears to be good but is actually doing it for the wrong reasons.[55]

Of course, the problem is that we may not realize that things like this are even happening owing to the opaque nature of the model.

Let's consider what happens when we can understand some of these models. In 2019, a team at the University of Tübingen created an image classification algorithm generated using a combination of a deep neural

[55] Smith 2019.

network to recognize image parts combined with a transparent process that uses the number of detected features in an image to classify it. They trained this algorithm using images from a dataset known as ImageNet and then were able to have the algorithm point out what features in the image were most important for its decisions. Some of the results were surprising. For example, when it was asked what the most important features were for identifying a tench (a kind of fish), the result was that it was looking for human fingers against a green background.[56] All the images of a tench in the dataset included humans holding up the fish to a camera with their hands as a trophy, thus making it a predictive feature. An image-generating algorithm trained on the same dataset was asked to produce an image of a tench, and it produced photos of humans holding a fish with an emphasis on human fingers.[57]

In another example, a research team created an image classification algorithm that was trained to recognize different kinds of animals, including distinguishing between huskies and wolves. Using a process to try to detect which patch of pixels the model is using to classify images, they discovered that it distinguished between huskies and wolves not based on any features of the animal themselves but on whether there was snow in the background. If there was snow, the image was classified as a wolf, and without snow, it was classified as a husky (Figure 23). Why did this happen? All the images of wolves in the wild had snow in the background. As the team noted, "Often artifacts of data collection can induce undesirable correlations that the classifiers pick up during training."[58]

As Longino says, the universe does not come with labels. The way that a neural network might find patterns in the data to produce an answer means that they may find patterns that make little sense to us or would be difficult to explain. These models not only detect statistical relationships within the data, but they determine how significant those relationships are for producing an output. Yet, due to the opaque nature of the model we may not know which patterns the model is looking for or how important any of those patterns might be for producing a result.

This makes it very difficult to consider what counts as sufficient evidence for relying on these patterns. How important should the statistical connection be between one variable and another before you would use it

[56] Brendel and Bethge 2019, 5.

[57] Shane 2020.

[58] Ribeiro, Singh, and Guestrin 2016, 1142.

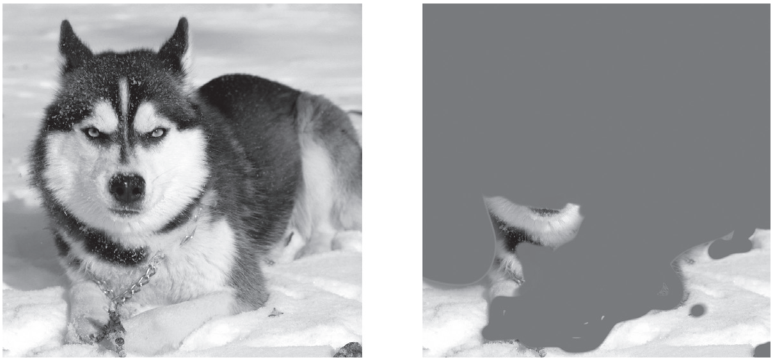

FIGURE 23 • Using techniques such as LIME, we can explain why a network classified an image a particular way. The model would have falsely concluded that this husky on the left is a wolf based on the snow in the background rather than any features of the dog. If we aren't careful with our training data, a neural network can rely on meaningless correlations to generate predictions that might appear accurate.

in the world? Also, input data itself is also subject to error. Credit reports (summarizing a person's record at payback of loans, etc.) are a commonly used piece of input data in machine learning, but studies show that more than a third of credit reports contain errors.[59] If the reliability of a given piece of data is in question and on top of that you don't know how important that particular variable is in terms of producing an answer, how can you manage inductive risks associated with the use of that data?

If we can explain why a model produced a particular answer given specific inputs, it is **interpretable**.[60] In order to manage inductive risks in the face of opacity, we might seek to make AI-generated models more interpretable. Is there a way to try to peek inside the black box and figure out what the model is really doing? If we can gain some understanding of these models, how can this insight inform solutions for our ethical problems concerning the risks of error?

INQUIRER'S TOOLBOX

7. Am I considering all the information relevant to a solution? Is that information reliable?
8. When I consider how chosen ends might function as means to future ethical situations, are there major ethical concerns to consider?

[59] Gill 2021.

[60] Ribeiro, Singh, and Guestrin 2016, 1136.