1

Here’s How You Should Think about Ethics

This book gives you advice on how to build, procure, and deploy AI in an ethically (and thus reputationally, regulatory, and legally) safe way, and to do it at scale. We’re not here to tackle existential and metaphysical questions like, “How does AI affect what we should think about what it is to be human?” or “What does AI teach us about the nature of consciousness?” That said, we cannot get clear direction without clear conceptual foundations. This chapter lays those foundations.

The senior executive who described AI ethics as “squishy” wasn’t some rube. He was a person with a long and successful career in risk and compliance. And the others who describe it as “fuzzy” and “subjective” likewise are smart, accomplished people.

As a former professor of philosophy, I heard these sentiments for nearly twenty years. I now see them again with my clients and whenever people talk about AI ethics in almost any context. And when I do hear it, I’m quick to point out that such a mindset stymies progress.

When people say ethics is squishy—or as I’ll say from now on, “subjective”—they’re effectively saying that they’re not quite sure how to think about it, and they usually give up trying.

It’s particularly difficult for senior leaders trying to effect change within their organization. Those leaders are trying to create a comprehensive AI ethical risk program, which requires buy-in at every level of the organization. They often have the experience of walking into a room of engineers, telling them AI ethics is really important, and then facing the inevitable question, “But isn’t ethics subjective?”

This is the kiss of death. Engineers tend to like things that are concrete, quantifiable, empirically verifiable. Whatever is not that doesn't deserve their intellectual attention or care. That senior leader talking about AI ethics? That's just PR. That's political correctness getting in the way of technological progress. It's touchy-feely stuff that has no place in a conversation among serious people, and certainly not in business.

The leader who doesn't know how to respond to such protestations is in trouble. If they say, "Well, yes, it's subjective, but …" they've already lost. So senior leaders need to get this right if they are to get the organizational buy-in they'll need to drive—sorry, nestle—AI ethics throughout operations.

That's just the beginning, though. You'll also need people to think about ethics in a way that isn't met with a shrug whenever they need to do some ethical thinking, as required by the AI ethics program you'll put in place; in performing ethical-risk due diligence during product development, for instance, or in the deliberations by an ethics committee.

So, getting your people to think about AI ethics as something other than subjective is imperative both for organizational buy-in and for the sake of effective ethical risk analysis during product development, procurement, and deployment. As it happens, there is a very good reason for you and your team to stop thinking of ethics as subjective and to start thinking of it in a way that lends itself to fruitful discussions and, ultimately, risk identification and mitigation. Put slightly differently, if you're prone to talk about "responsible AI," then you'll need to think about ethics in a way that lends itself to responsible inquiry into the ethical risks of AI. And put differently one more time for those of you in the back seats: AI ethics is about two things—AI and ethics. In the previous chapter we got clarity on what AI or ML is and how it works. Now it's time to get clear on ethics.

Don't worry: I'm not going to write a philosophical treatise here. Turns out, all you need to think about ethics effectively is to get clarity on *a question*, *a confusion*, and *three notoriously bad but ubiquitous reasons for thinking ethics is subjective*. Once we do that, off we go to putting ethics into practice.

The Question

One question I get a lot is “What is ethics?” The inquirer is standardly looking for a “definition” of ethics. They might even ask, “How do you define ‘ethics’?” or “What’s your definition of ‘ethics’?”

But my view of how to get a grip on what ethics is *about*—and this is really what the inquirer is after—is to think about some of the core questions that we naturally characterize as *ethical* questions, like these:

- What is a good life?
- Do we have any obligations to each other? What are they?
- Is compassion a virtue? Courage? Generosity?
- Is abortion ethically permissible? Capital punishment? Euthanasia?
- What is privacy and do people have a right to it?
- What is discrimination and what makes it bad?
- Do people have equal moral worth?
- Do individuals have an obligation to engage in self-improvement?
- Is it ever ethically permissible to lie?
- Do corporations have obligations to their employees? To society at large?
- Is Facebook unreasonably incentivizing or manipulating its users into clicking on ads?
- Is it ethically permissible to use black-box algorithms to diagnose illnesses?

And so on. What is ethics? Well, don’t worry about a definition of the term—if you really want a definition, just look it up in a dictionary. If you want to know what ethics is about, think about these kinds of questions, and those in the neighborhood of these questions. If you understand this, there’s no reason to get worked up over definitions.

The Confusion

A significant source of confusion for many people who think of ethics as subjective is failing to distinguish between people's *beliefs* about ethics—what they believe to be ethically right or wrong, good or bad, and so on—and ethics itself. And in running these two things together, they make misguided claims about the subjectivity of ethics when they're really making claims about the variance of people's beliefs. To see this, let's take a step back.

There's our belief about whether the earth is flat or round, on the one hand, and there's the actual shape of the earth, on the other. There's our belief about the chemical composition of water being H_2O or H_3O, on the one hand, and there's the actual chemical composition of water, on the other. There's our belief about whether the 2020 election was stolen or legitimate, on the one hand, and there's the actual legitimacy of the election, on the other.

We generally distinguish between our beliefs about X and what X is actually like, and sometimes our beliefs are true, and sometimes they are false. If we didn't make this distinction between our beliefs about X, on the one hand, and what X is actually like, on the other, then we'd have to think that believing X makes it so, but no one thinks that believing the earth is spherical or flat, that water is H_3O or H_2O, or that the election was stolen or legitimate makes the earth spherical, or water composed of H_2O, or the election legitimate.

Of course, people's beliefs about these can change or evolve over time. At one point, most people believed that the earth is flat, they didn't believe that water is H_2O (in their defense, they didn't know anything about chemistry), and some people changed from believing the election was stolen to believing it was legitimate. So, our beliefs change about these things, but the things they had (or didn't have) beliefs about were what they were all along. It's not as though the earth changed from being flat to spherical.

Let's keep going with this distinction: there's our belief about the ethical permissibility or impermissibility of slavery, on the one hand, and there's whether slavery is ethically permissible, on the other. If anything is ethically *impermissible*, it's slavery.

At one point, most people—particularly those who benefited from slavery—believed that slavery was ethically permissible. But people's beliefs changed or evolved over time, and now all believe slavery is wrong. The wrongness of slavery didn't change; it was always wrong. (Quick note: there's a separate issue about the extent to which those who thought it was ethically permissible are deserving of *blame*, given that everyone around them also thought it was permissible, but we won't discuss that here.)

In a way, all this is fairly obvious. *Of course* there's a difference between what people believe about X and what X is actually like. But things tend to get very weird when people talk about ethics; the distinction goes right out the window. People will say things like, "Your ethics is different from my ethics" or "Ethics is subjective because ethics or morality varies across cultures and individuals," or "Ethics has evolved over time; people once thought slavery was ethically permissible and now they think it's not."

But now we can see that "your ethics is different from my ethics" can mean either "what's ethically right for you is ethically wrong for me" or "what you believe is ethically right is something I believe is ethically wrong." And we've already seen that while its clear ethical beliefs change or evolve over time, that doesn't mean that what is right or wrong changes over time. What's weird is that, when people say these things, they are often thinking of ethical beliefs as *the same thing as* ethical right and wrong, and that's just confusion.

The question about whether ethics is subjective is, to be clear, not a question about whether people's ethical beliefs vary over time and across individuals and cultures. Of course they do! The question about whether ethics is subjective is about whether what's right or wrong, or good or bad, varies across time, individuals, and cultures. Now that we understand that, we can look at common reasons for thinking that ethics is subjective.

Three Really Bad Reasons for Thinking Ethics Is Subjective

To say ethics is subjective is to say that there are no facts about what is ethically right, wrong, good, bad, permissible, impermissible, and so on. If ethics is subjective, then not only do ethical *beliefs* vary by individual and culture, but *ethics itself* varies by individual and culture. If ethics is subjective, then there's no such thing as *responsible* ethical inquiry because no one can possibly be incorrect in their conclusions (and so much for responsible AI ethics, or "responsible AI"). If ethics is subjective, then it's touchy-feely, squishy, fuzzy, and not a subject for serious people and certainly not for serious people in a business context.

Now we know what ethics is about. And we know to distinguish between ethical beliefs about what is right or wrong and what *is* right or wrong. But even people who know these things can still think ethics is subjective. And in those nearly twenty years of teaching philosophy, I've noticed three primary reasons for the belief that ethics is subjective, each of which is flatly misguided. I'll lay out the reasons and then explain what's wrong with them. And to be clear: this is not just my view that these are bad reasons. Philosophers don't agree about a lot, but there's a consensus that even if ethics is subjective, it's not for any of these reasons.

Really Bad Reason #1: Ethics is subjective because people disagree about what's right and wrong. People engage in ethical disputes; they disagree about whether abortion and capital punishment are morally permissible, whether you should lie to the police to protect your friend, and whether collecting people's data without their knowledge in exchange for the otherwise free use of your services is ethically permissible. And since there is so much disagreement—so many different moral and ethical beliefs—ethics is subjective; there's *no truth* to the matter.

Really Bad Reason #2: Science delivers us truth. Ethics isn't science, so it doesn't deliver us truth. Science, and more specifically, the scientific method, is *the only* way we discover truths about the world. Empirical observations ("seeing is believing") and investigations (scientific experiments, for instance) deliver facts about the world. Everything else is interpretation, which is to say, subjective. Again, ethics is subjective because empirical observations have a monopoly on truth; ethics and ethical inquiry, because it is not empirical inquiry, concerns the realm of nontruth. In short: *only scientifically verifiable claims are true*.

Really Bad Reason #3: Ethics requires an authority figure to say what's right and wrong; otherwise, it's subjective. You have your beliefs and I have mine and that other person has theirs. And it’s not like we have scientific evidence that one view is right and another is wrong, so who’s to say what’s right and what’s wrong? It’s all subjective. Or in short: *if there are ethical truths, then there must be an authority figure who makes this right and that wrong.*

What’s So Bad about the Bad Reasons

Why Really Bad Reason #1 is really bad. The first reason for thinking ethics is subjective is that people disagree about what is right or wrong, and if people disagree about that stuff, then there’s no truth to the matter. Is this a good argument? You guessed it … it’s really bad! And you can see how bad it is when you consider the following principle:

If people disagree about X, then there’s
no truth to the matter about X.

Now that principle is obviously false. People disagree about all sorts of things about which there’s a truth to the matter. People disagree about whether humans are the product of evolution, whether self-driving cars will replace human-driven cars within a decade, whether there’s something at the center of a black hole, and even whether the earth is flat or spherical. But no one thinks, “Well, guess there’s no truth to the matter about the shape of the earth!”

The fact that people disagree about X doesn’t show that there’s no truth to the matter about X.

And so, too, with ethics. The fact that people disagree about whether lying to protect your friend from the police, whether people should own their data, whether Facebook engages in ethically unacceptable manipulation of its users, and so on, doesn’t show that there’s no truth to the matter about those issues.

“But,” people retort, “it’s different with ethics. That’s an exception to the principle.”

But why should we think it’s different with ethics? Why should we think it’s exempt from the lesson we just learned about disagreement and truth?

The answer is the same 99 percent of the time: “Because in those other cases of disagreement, they can be scientifically settled. With ethics, there’s no way to scientifically settle it.”

That’s a fine reply, in a way. But it’s really an *abandonment* of the first reason for thinking ethics is subjective and a retreat to Really Bad Reason #2 for thinking ethics is subjective. The reply just says, “Only scientifically verifiable claims are true.” So let’s investigate that.

Why Really Bad Reason #2 is really bad. This one is surprisingly easy to refute. It says only scientifically verifiable claims are true. Actually, let’s really throw this into high relief:

Claim: Only scientifically verifiable claims are true.

If you’re particularly astute, you just asked yourself a question: “If this is a claim, and the claim says that only scientifically verifiable claims are true, what about this claim?”

This question reveals the problem with the position: it’s self-undermining. After all, how would you scientifically verify this claim? Give it to the chemist, or the biologist, or the physicist, or the geologist, and say, “Please perform an experiment to verify this claim.” What could they possibly do? Write it on a piece of paper and measure how much weight the paper has gained? Attach it to a seismic reader? Put some cells on it? There’s just nothing for them to do, and that’s because the claim is not itself scientifically verifiable. So, anyone who believes the claim would, if they are to be consistent, have to stop believing it. And for those who never believed it in the first place, they’re fine. So, whatever you do, don’t believe the claim. It’s false.

Why Really Bad Reason #3 is really bad. OK, almost there. You might think that for ethics not to be subjective, there have to be ethical facts, and for there to be ethical facts, there will have to be an authority figure to say what’s right and wrong.

But this is to ignore some basic ways we think about facts. No one says that if there are going to be facts about the shape of the earth, or the evolutionary history of humans, or the chemical composition of water, then there must be an authority figure that makes these things facts. Instead, there are facts about these things and there are people (scientists, of course) who give us the *evidence* for the claims that the earth is spherical, humans are the product of biological evolution, and water is composed of H_2O. It's the evidence, the arguments they offer us for those conclusions, that wins the day, and it's certainly not the *discoverers* of that evidence that make the earth spherical or water composed of H_2O.

If there are moral or ethical facts, then we should expect them to act the same way as other facts. There is no need for an authority figure to make them true. Instead, there are people (philosophers and theologians, for instance) who give evidence or arguments for the ethical claims they make. Think of the many arguments, counterarguments, and counters to the counterarguments in discussions of the moral permissibility of abortion. None of these people say, "I think it's wrong, so it's wrong." If they did, we'd pay them no attention. When we're at our best, we pay attention to their arguments and investigate whether they're sound, just as we do with scientific arguments.

Why This Matters

It's important we get rid of these confusions and bad arguments. In ethics, including in areas related to artificial intelligence, we face very real ethical problems. Those problems, if not properly solved both ethically and technically, can lead to disastrous consequences. But if we give up on ethics as being something objective, as something we can reason about and give arguments for and reasonably change our minds about, then we give up ethical inquiry being a tool at our disposal to solve these real problems.

Let me get a little more specific. I've witnessed hundreds if not thousands of discussions on issues of ethical import. And they all go the same way, so long as people feel comfortable to air their views. You get some people arguing for one side, others arguing for others, and the ones who are not sure where they stand. This is difficult stuff. And then someone says, "I mean, what does it matter? This is all subjective anyway." And then everyone looks at each other, blinks, and shrugs. End of discussion. Every. Time.

Until I ask, "Why do you think ethics is subjective?" Without fail, I get the Really Bad Reasons. Once we've dismantled them, people take up the issue again, this time invulnerable to a comment that would lead them completely off the rails.

You do not need to start your AI ethics program with talking about the nature of ethics. But if you do not involve it at some point, I promise you—I *promise you*—people are going to bring up the Really Bad Reasons. And then you'll have a bunch of people who resent political correctness or touchy-feely stuff standing in the way of technological greatness. You'll also get reduced compliance with your AI ethics program and increased risk. What people think about AI ethics is going to play a role in whether you have an effective AI ethics program.

C'mon … Ethics? … Objective?

I'm not *really* trying to convince you that ethics is not subjective. I'm not trying to convince you that there are ethical facts. I am trying to convince you that the standard reasons for thinking ethics is subjective are really bad reasons, and failure to see this can lead to a lot of trouble. But the fact that there are three really bad reasons for thinking ethics is subjective doesn't mean that there can't be a fourth reason for thinking ethics is subjective that is really good.

We're not going to go there in this book. The point of defusing those really bad reasons is that they continually put a stop to fruitful discussions and are an impediment to genuine organizational buy-in, from the top to the bottom. Ninety-nine percent of people who see why they're really bad reasons are ready to accept that ethics isn't subjective, and so the practical aim of discussing this is accomplished.

For those of you who are not convinced, however, you should at least think that thinking that there are ethical facts is not an unreasonable position. It's certainly not *crazy*. And so, *for the purposes of creating and participating in an AI ethical risk program*, I hereby invite you to join those people in practice. It will enable you to think about what an effective system for identifying and mitigating those risks would look like. It will enable you to have reason- or evidence-based conversations with colleagues about what is the right thing to do. And it will ensure you don't unhelpfully shut down important conversations that lead to the protection of your organization from ethical, reputational, regulatory, and legal risks.

Why Not Just Talk about Consumer Perception Instead of Ethics?

You might be wondering why we need to talk about ethics at all. Why not just talk about consumer ethical beliefs or consumer perception more generally? Then you can do your standard market research and simply make those your ethical standards internally. AI ethics is really just brand values built into AI product development and deployment, so let's put all this ethics talk to the side. In fact, let's drop this "AI Ethics" label and call it what it is: "AI Consumer Perceptions."

This is an entirely reasonable question. I do not think it rests on confusion or misunderstanding or naivete or an ethical character flaw. It's not a *crazy* thing to do. That said, it's unwise. Here are three reasons why I advise against it.

Lucy, you have some operationalizing to do … Adopting your consumers' or clients' ethical views still gives you an ethical view that you now have to operationalize. The problem is that the relatively coarse-grained analyses of customer perception are not well suited to answer the fine-grained decisions you'll need to make. For instance, everyone of minimal moral decency—including your consumers, of course—opposes discriminating against people on the basis of race. But how you should think about discrimination in the context of, say, determining which metrics to use for evaluating whether your model outputs are discriminatory is not something consumers can help you with. Here the issue is actually twofold: first, your customers' ethical perceptions are too coarse-grained to easily transfer to your fine-grained problems, and second, your problems are ones that your customers haven't even thought about yet.

Facebook, for instance, is under a lot of scrutiny for the way its algorithms determine what content to serve up in people's news feeds. A popular documentary, *The Social Dilemma*, has been made about the issue. But when Facebook engineers and product developers were writing those algorithms, they couldn't even begin to ask their customers what they thought because their customers don't understand anything about algorithms, the potential pitfalls, what data is being collected and what's being done with it, and so on. So Facebook—and any company engaged in producing innovative technologies—needs to anticipate the ways various ethical risks might be realized in advance of its customers knowing that these technologies exist and thus in advance of them having any ethical beliefs or perceptions to detect in the first place.

Trust requires ethical leadership. Companies require consumers to trust them if they are to attract and maintain those consumers, and nothing violates trust quite as mercilessly as an ethics breach at scale. Think #DeleteFacebook, #DeleteUber, #BoycottStarbucks. Each of those companies violated consumer trust—trust that they would protect your data, trust that they would treat their workers respectfully, and trust that they would ensure a nondiscriminatory environment for African Americans, respectively—that led to going viral for all the wrong reasons. The vaccine is ethical leadership.

I don't mean that corporations need to be Gandhi. I mean they need to articulate their ethical values clearly and explain how they live up to them. If a company can't do that, but instead simply mouths fealty to an ethical code and then runs to market research on consumer sentiments of the day and makes decisions solely on the basis of that, it can expect its ethics messaging to justifiably ring hollow and thereby lose the trust of its consumers. Just think about how much time you want to spend with the person who espouses various values and then does whatever the cool kids tell them to do.

Organizational buy-in. I've stressed that creating an AI ethics program requires buy-in from the top to the bottom, and an anathema to that buy-in is for employees to see AI ethics as PR or bowing to political correctness. If your general AI ethics strategy just reduces to a strategy of deferring to consumer survey results, you should expect that you won't only fail to get buy-in from a great variety of internal stakeholders, but also that you'll alienate the growing number of employees who are passionate about AI ethics and will accuse you of ethics washing. If you want buy-in, you're going to have to take ethics, not just ethical perceptions of your consumers, very seriously.

Where We Go Next

The foundations of ethics, now concrete, are settled. It's time to build on that foundation by understanding bias, explainability, and privacy concerns in a way that reveals what Structure we need. We'll start in the next chapter with the most contentious issue of all: discriminatory AI.

Recap

- Putting an AI ethics program into practice requires getting organizational buy-in. Getting that buy-in requires that people understand what AI ethics is. Understanding what AI ethics is requires understanding something about AI and something about ethics.

- Major stumbling blocks for understanding crucial aspects of what ethics is include confusing ethical beliefs with ethical facts and the Really Bad Reasons that lead people to think and talk about ethics as being "squishy" or "subjective." Thinking about ethics in this way is to think that there are no ethical facts to be discovered, and this ultimately leads people to shrug their shoulders when it comes to thinking carefully about identifying and mitigating ethical risks.

- AI ethics should not be reduced to reactions to market research on consumer perception or sentiment, for at least three reasons: (1) an AI ethics program involves operationalizing a set of values, which consumer perception reports are silent on, (2) consumers are looking for ethical leadership, and a mere appeal to the sentiment of the day does not meet that bar, and (3) that approach will alienate both those who are not particularly concerned about the ethical risks of AI within your organization and those who are, leading to a lack of compliance and turnover, respectively.

- All of this means you'll need to provide your people with education about both AI and ethics.