10 Oversight and Regulation

What rules should govern artificial intelligence?

Unsurprisingly, it's a question that has attracted a wide range of answers. In recent years, however, a rough consensus seems to be emerging that there *should* be rules. The consensus doesn't just include the usual suspects, either—academics, activists, and politicians. In 2014, tech entrepreneur and celebrity Elon Musk surprised many by calling for "some regulatory oversight" of AI, "just to make sure that we don't do something very foolish."[1] Facebook CEO Mark Zuckerberg's calls are similar (though made for different reasons).[2]

But what form should this regulatory oversight take? What should be its target? Who should be doing the overseeing? And what sorts of values or standards should they be seeking to enforce?

It's hardly an original observation that emerging technologies present challenges for lawmakers and regulators. When we don't know for sure what form a technology will take, how it will be used, or what risks it will pose, it can be hard to tell whether existing rules will be adequate or whether a new regulatory scheme will be required instead. The challenge is especially daunting when the technology in question is something like AI, which is actually more like a family of diverse but related technologies with an almost unlimited range of uses in an almost unlimited range of situations. What sorts of rules or laws, we might wonder, are supposed to manage something like *that*? Rules, after all, are—pretty much by definition—relatively certain and predictable. A rule that tried to keep pace with a fast-evolving technology by constantly changing? That, you might think, would hardly qualify as a rule at all.

Luckily, AI is by no means the first emerging technology to have posed these challenges. Decisions about the right rules can be made against a

background of considerable experience—good and bad—in responding to predecessor technologies: information and communication technologies, for example, as well as genetic and reproductive technologies.

What Do We Mean by "Rules"?

So far, we have used the terms "rules," "regulation," and "laws" interchangeably and fairly loosely. But what exactly are we talking about? There's actually quite a wide range of options available to regulate AI. On one end of the spectrum, there are relatively "hard-edged" regulatory options. Predominantly, we're talking about regulation by *legal* rules.

When most people think about "legal rules," they probably think in terms of legislation, "the law on the statute books." They might also consider rulings by courts. Both of these are important sources of legal rules (the *most* important, really). But legal rules can come in a wider variety of forms. They can, for instance, be made by regulatory agencies with delegated responsibility. In the realm of assisted reproductive technologies, organizations like the UK's Human Fertilization and Embryology Authority and New Zealand's Advisory Committee on Assisted Reproductive Technology (ACART) are empowered by statute to set limits and conditions on their use, at least up to a point.

"Regulation" is an even wider concept than that. It's a much discussed and somewhat contested term in the academy, but it's still often thought to be a wider concept than just "law." For the purposes of this book, we've taken a fairly open-ended approach to what counts as "regulation." The questions raised by the emergence of AI are so many and so diverse that it would be rash to rule out any possible answers before we even start. For our purposes then, "regulation" means laws and court judgments, but it doesn't just mean that. Like AI itself, the rules we apply to it could come in a wide and varied array of forms.

Who Makes the Rules?

Regulation then, needn't be confined to legal prohibitions and orders, and neither is it limited to rules issued by legislatures, courts, or regulatory agencies. What other forms could it take? A commonly proposed alternative in the context of AI is that of *self-regulation*. Companies might set rules

for themselves, or industry bodies might elect to impose some rules upon their members.

Why would a company or a whole industry opt to do this, to tie its own hands in such a way? Cynically, we might see it as a self-interested alternative to having it imposed from outside. In other cases, other motives might play a part. It might be to offer reassurance to clients and customers, for example, or to establish and uphold the reputation of that industry as a safe and responsible one. Some critics of the adequacy of self-regulation are more overtly skeptical of corporate motivations. Cathy Cobey, writing in *Forbes* magazine, suggests that even when companies call for externally imposed regulations, the real reason stems from a concern that, "if the decisions on how to govern AI are left to them, the public or the court system may push back in the years to come, leaving corporations with an unknown future liability."[3]

To what extent can self-regulation adequately protect society from the worst risks of AI? For many commentators, this just doesn't offer enough by way of guarantees. James Arvanitakis has argued that "for tech companies, there may be a trade-off between treating data ethically and how much money they can make from that data." This has led him to conclude that, "in order to protect the public, external guidance and oversight is needed."[4] Jacob Turner has made a similar point, explaining that "considerations of doing good are often secondary to or at the very least in tension with the requirement to create value for shareholders."[5]

For Turner, another weakness with self-imposed rules is that they lack the binding nature of actual law. "If ethical standards are only voluntary," he suggests, "companies may decide which rules to obey, giving some organizations advantages over others."[6]

Recently, even some major industry players seem to be coming round to the view that, although self-regulation will have a major role to play, there is also a role for governments, legislatures, and "civil society" in setting the rules. In its 2019 "white paper," Google combined a fairly predictable preference for self- and co-regulatory approaches, which it claimed would be "the most effective practical way to address and prevent AI related problems in the vast majority of instances," with an acknowledgment that "there are some instances where additional rules would be of benefit."[7]

Interestingly, Google's position was that this benefit was "not because companies can't be trusted to be impartial and responsible but because to delegate such decisions to company uses would be undemocratic."[8] This relates to what

some emerging technology commentators refer to as "regulatory legitimacy." This has been explained as meaning that "as technology matures, regulators need to maintain a regulatory position that comports with a community's views of its acceptable use."[9] Simply leaving the rules up to industry to determine is unlikely to satisfy this demand, at least in areas where the technology is likely to have important impacts at an individual or societal level.

Google's preferred solution seems to involve a "mixed model," combining self-regulation with externally imposed legal rules. Precedents aplenty exist for such a hybrid. In many countries, the news media is governed by self-regulatory mechanisms, such as the UK's Press Complaints Commission, Australia's Press Council, or New Zealand's Media Council. But news media are also subject to "the law of the land" when it comes to matters like defamation law, privacy, and official secrets. The medical profession is another example of a mixed model, subject to "the law of the land" in many respects, but also to professional ethics and internal disciplinary structures. Doctors who break the rules of the profession can be suspended, restricted, or even "struck off," but doctors who break the laws of society can also be sued or occasionally even imprisoned.

Whether we think that would be a desirable model for AI might depend on how well we think it has functioned in other industries. Most suspicion probably arises when self-regulation functions as an *alternative* to legal regulation rather than as an adjunct to it. And there's obviously nothing to stop a company or a whole industry opting to set rules that go *further* than the existing law.

Flexibility

Regulation can also come in both more and less flexible forms. Very specific rules will allow little room for interpretation. To go back to the area of reproductive technologies, both UK and New Zealand laws contain a number of flat-out bans on certain practices—creating animal-human hybrids for instance or human cloning.

In the context of AI, there have been several calls for outright bans on certain kinds of applications. For example, in July 2018, an open letter was published by the Future of Life Institute, addressing the possibility of "[a]utonomous weapons [that] select and engage targets without human intervention." The open letter, which has now been signed by over 30,000

people, including 4,500 AI and robotics researchers, concluded that "starting a military AI arms race is a bad idea and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control."[10]

Other specific applications of AI technology have also been the subject of actual or suggested bans. Sometimes these are context-specific. For example, California has enacted a (temporary) ban on facial recognition on police body cams,[11] whereas San Francisco has gone further by banning the technology's use by all city agencies.[12]

These are examples of *negative* obligations, rules that tell those who are subject to them that they must avoid acting in certain ways. Rules can also impose *positive* obligations. California again provides a good example. The Bolstering Online Transparency Bill (conveniently abbreviating to "BOT"), which came into effect in July 2019, "requires all bots that attempt to influence California residents' voting or purchasing behaviors to conspicuously declare themselves."[13]

How effective any of these new laws will prove to be at achieving their objectives remains to be seen, but those objectives are quite clear and specific. Regulation needn't always be quite so clear-cut. It can also take the form of higher level guidelines or principles. These are still important, but they come with a degree of flexibility in that they have to be interpreted in light of particular circumstances.

There have been many examples of this sort of "soft law" for AI. The EU Commission's recent AI principles are of this kind, but we've also seen guidelines issued by the OECD,[14] the Beijing Academy of Artificial Intelligence,[15] the UK's House of Lords Select Committee on AI,[16] the Japanese Society for Artificial Intelligence,[17] and the Asilomar Principles from the Future of Life Institute.[18] Google and Microsoft have also issued statements of principle.[19] One thing we can safely say is that, if anything goes badly wrong with AI, it probably won't be for want of guidelines or principles!

Examining these various documents and declarations reveals a fair degree of agreement. Almost all of them identify principles like fairness and transparency as being important, and most insist that AI be used to promote something like human well-being or flourishing. This might seem like a promising sign for international consensus. On the other hand, we could say that the principles and the agreement are at a very high level of generality. Would anyone *disagree* that AI should be used for human good or argue that it should be unfair? For lawyers and regulators, the devil will

be in the detail, in turning these commendable but very general aspirations into concrete rules and applying them to real life decisions. They'll also want to know what should happen when those principles conflict with one another, when transparency clashes with privacy maybe, or fairness with safety.

Sometimes the firmer and more flexible rules work together. As well as clear and specific bans on things like human cloning, New Zealand's Human Assisted Reproductive Technology Act contains general guiding principles and purposes that the regulators must keep in mind when making decisions. These include things like "the human health, safety, and dignity of present and future generations should be preserved and promoted" and "the different ethical, spiritual, and cultural perspectives in society should be considered and treated with respect." These principles must inform how the regulators act, but they are open to a range of interpretations in the context of particular decisions.

Rules can be binding. That's probably true of most of the rules that come readily to mind. The laws against drunk driving, burglary, and tax fraud aren't discretionary. But rules can also be advisory, perhaps setting out "best practice" standards. The binding vs. nonbinding classification is different from the specific vs. generic one. A rule can be specific (say, about how to use a particular type of software), but merely advisory. Programmers working on that specific software could choose to ignore the recommendation without penalty. On the other hand, rules could be quite general or vague (like the principles we just mentioned) and yet mandatory. Programmers would *have* to follow them (even if there's wiggle room on interpretation).

We can see then that "regulation" is a fairly wide concept. It can come from a range of sources—from governments, legislatures, and courts, certainly, but also from agencies delegated to regulate particular areas and even from industries or companies themselves. Its form can range from the very specific (bans on particular uses or practices) to more high-level principles. Finally, its effect can range from the binding, with legal penalties for noncompliance, to the advisory and merely voluntary.

Any of these approaches—or a combination of them—could be used to regulate AI. The next question concerns what exactly it is we're trying to regulate in the first place.

The Definition Problem: What Are We Even Talking About?

The first task for anyone proposing regulation probably involves defining what exactly it is they want to regulate. That's no big challenge when the target is something like "Lime scooters" or "laser pointers," but when the target is something as broad as "AI," the task can be daunting.

This absence of a "widely accepted definition of artificial intelligence" has been a recurrent theme in the literature,[20] and at first glance, it does look like a major stumbling block. Courts, after all, need to know how to apply new rules, regulators need to know the boundaries of their remit, and stakeholders need to be able to predict how and to what the rules will apply.

Is it actually necessary to have a definition of "AI" before we can begin to assess the need for regulation? That might depend on a couple of factors. For one thing, the need for a precise definition is likely to depend a lot on whether we're attempting to create AI-specific rules or regulatory structures. If laws of more general application would be adequate in this context, then the definitional challenge becomes less pressing because whether or not something qualifies as "AI" will have little bearing on its legal status. But if we think we need AI-specific laws and regulations, then we certainly would need a definition of what they apply to.

Another possible response to the definitional challenge is to forswear an all-purpose definition and adopt a definition suited to a particular risk or problem. Rules could be introduced for predictive algorithms, facial recognition, or driverless cars without having to worry unduly about a general definition of "AI."

Regulatory Phase: Upstream or Downstream?

If we can agree on *what* and *how* they want to regulate, regulators then have to decide on the question of *when*. In some cases, the advantages of very early regulation—before the technology has been brought to market or indeed even *exists*—can be obvious. If a technology is seen as particularly dangerous or morally unacceptable, it may make sense to signal as much while it's still at a hypothetical stage. In less extreme cases, early intervention might also be beneficial, shaping the direction of research and investment. Gregory Mandel has argued that intervention at an early stage might meet less stakeholder resistance, as there will predictably be fewer vested

interests and sunk costs, "and industry and the public are less wed to a status quo."[21]

Alternatively, there may be times when it's better to wait and see, dealing with problems if and when they arise rather than trying to anticipate them. The internet is often held up as an example of a technology that benefitted from such an approach. Cass Sunstein has made the case for caution before rushing to regulate, pointing out that, at an early, "upstream" stage, it's harder to make accurate predictions and projections about costs and benefits. "If we will be able to do so more accurately later on, then there is a (bounded) value to putting the decision off to a later date."[22]

There's almost no straightforward answer to whether it's better to regulate early or late. Everything depends on the particulars of the technology. Luckily, we are not faced with a binary choice. The decision is not "everything now" or "nothing until later." It's entirely possible to take regulatory steps to address some risks early—maybe those that are particularly serious, immediate, or obvious—and to defer decisions about those with are more speculative or distant until later when we are better informed.

Regulatory Tilt and Erring on the Safe Side

In many cases, however, there is just no way to defer all decision-making until later. When a technology exists in the here and now and is either being used or the subject of an application for use, the regulatory option of putting off the decision just isn't on the table.

What sort of approach should regulators adopt when making decisions in the face of uncertainty? The idea of "regulatory tilt" describes the starting point or default setting that they should adopt. If they can't be sure of getting it right (and they will often be in doubt), in which direction should they err?

One obvious approach to this question is that regulators faced with uncertainty should err on the side of caution or safety. This is sometimes expressed in terms of the precautionary principle. When the EU Parliament made its recommendations on robotics to the EU Commission in 2017, it proposed that research activities should be conducted in accordance with the precautionary principle, anticipating potential safety impacts and taking due precautions.[23]

When an emerging technology presents an uncertain risk profile, there does seem to be something appealing about the application of a precautionary

approach. Waiting for conclusive evidence about the dangers it presents may result in many of those risks materializing, causing preventable harm or loss. In some cases, the loss may be of a nature that can't easily be put right, not only in terms of individual people harmed, but in the sense that it may be impossible to put the genie back in the bottle. Think of a genetically modified bacterium or runaway nanobot released into the wild. Some of the concerns about runaway superintelligent AI are of that nature. If the concerns are at all credible, then they are of the stable door variety; there's not much to be done once the proverbial horse has bolted.

Nick Bostrom is probably the best-known voice of caution about such risks. "Our approach to existential risks," he warns, "cannot be one of trial-and-error. There is no opportunity to learn from errors."[24] Bostrom has proposed what he calls a "maxipok" approach to such risks, which means that we act so as to "maximize the probability of an okay outcome, where an 'okay outcome' is any outcome that avoids existential disaster."[25]

When the risks are of an existential nature—as some commentators really seem to believe they are where AI is concerned—the case for precaution looks pretty compelling. The challenge, of course, is to determine when such risks have any credible basis in reality. Our legal and regulatory responses, presumably, should not be responses to the direst dystopian visions of futurists and science fiction writers. But how are regulators meant to tell the real from the fanciful? History—indeed the present!—is littered with horror stories of major risks that were overlooked, ignored, or disguised, from asbestos to thalidomide to anthropogenic climate change and global pandemics. Less well known, perhaps, is that history is also replete with examples of worries about new technologies that amounted to nothing—that to modern eyes, look quite ridiculous. Cultural anthropologist Genevieve Bell has recounted that, in the early days of railway, critics feared "that women's bodies were not designed to go at fifty miles an hour," so "[female passengers'] uteruses would fly out of [their] bodies as they were accelerated to that speed"![26]

Some cases are genuinely very hard for regulators because the science involved is both demanding and contested. Shortly before the Large Hadron Collider (LHC) was turned on at CERN, an attempt was made by some very concerned scientists to obtain a court order preventing its initiation.[27] The basis of their claim was a concern that the LHC could create a black hole, posing an existential threat to the planet. It is hard to envy the judge, trying to weigh up competing claims from theoretical physicists, most of whom

thought the operation safe, but a minority of whom believed it could spell the end the world.

In the event, the court opted to trust the majority view, the supercollider was switched on, and (as far as we know) nothing catastrophic occurred. However, an approach that prioritized the minimization of existential risk above all else would presumably have erred in the other direction.

Even where we have some good reason to take seriously a major warning, it won't always be the case that the "moral math" produces an obvious answer about how to respond. Often, decisions to avoid some risks will mean foregoing certain kinds of benefits. In the case of some forms of AI, those benefits could be very considerable. A precautionary approach to driverless cars, for instance, would certainly eliminate the risk of deaths by driverless cars. But an estimated 1.2 million people currently die on the roads every year. If driverless cars had the potential to reduce that number quite dramatically, passing up that chance would amount to quite a risk in itself.

It's easy to think of other potential applications of AI that might present the same kind of trade-off. For example, in medical diagnostics the possibility exists that AI could do considerably better than humans and that better performance could be translated into lives prolonged or improved.

What about the sort of superintelligent AI that seems to worry Bostrom and Musk, though? Maybe there's a case for a precautionary ban even on researching that sort of thing. But again the moral math dictates that we consider what we may be passing up in so doing. What avenues of research would need to be closed off in order to prevent someone even accidentally taking us into the realms of superintelligence? Might that mean that we had to pass up potential AI solutions for hitherto intractable problems, maybe even something as serious as climate change?[28] Is the threat of a runaway superintelligence really more existentially pressing than that?

A simple rule that says "always avoid the worst outcome" might seem appealing at first glance, but it isn't clear how to err on the side of safety when there may be existential threats in all directions or when the alternative seems less bad but more likely. Looked at like that, it seems that there really is no simple heuristic that will guarantee we avoid the worst scenario.

An FDA for AI?

Laws enacted by legislatures are certainly an important source of law. They offer certain advantages. They have democratic legitimacy, being enacted

by the elected representatives of that society. They are likely to be highly transparent (it's hard to pass a law without anyone noticing), which means that, ideally, their effects should be predictable.

But on the deficit side, legislation is notoriously slow to adapt. Getting a new law through the legislature can take a long time.* And since trying to anticipate every possible situation in a few pages is notoriously difficult, much of the new law's effect will only be decided when courts come to apply it in particular situations.

An alternative—or supplementary—option could be a specialized regulatory agency. Matthew Scherer has argued that regulatory agencies have several advantages when dealing with an emerging technology. For one thing, agencies "can be tailor-made for the regulation of a specific industry or for the resolution of a particular social problem."[29] Although legislatures will be populated by generalists, members of agencies can be appointed on the basis of their specialist knowledge in a particular area. For example, New Zealand's assisted reproduction regulator, ACART, is required by statute to have at least one member with expertise in assisted reproductive procedures and at least one expert in human reproductive research.

Regulatory agencies, Scherer argues, also have an advantage over court-made rules. Judges are significantly limited in their remit, being restricted to making decisions about the cases that actually appear before them. Regulatory agencies aren't similarly constrained.

Regulatory agencies come in a wide variety of forms and have a wide array of remits and responsibilities. As with legislation, they could be specific to a particular technology or family of technologies—ACART and the UK's HFEA, for instance, are focused on reproductive technologies. Or they could be fashioned with a particular policy objective or value in mind—the UK's Information Commissioner's Office and New Zealand's Office of the Privacy Commissioner were established to focus specifically on privacy and data protection.

---

*Often, but not always. On March 15, 2019, a mass shooting in the New Zealand city of Christchurch was live-streamed on Facebook. The Australian Government introduced legislation that required social media platforms to "ensure the expeditious removal" of "abhorrent violent material" from their content service. The Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019 came into effect on April 6, a mere three weeks after the event to which it was a response. The Government has been criticized for its failure to consult on the content or likely effects of the Act, a regular criticism of particularly fast-track lawmaking.

Regulatory agencies can also be conferred a wide array of powers. Some have compulsory inspectorate functions; others are able to hand out penalties and sanctions. Some can create rules, whereas others exist to enforce or monitor compliance with rules created by others. Often, they will operate at a "softer" level, for example, in issuing best practice guidelines or codes of practice or simply giving advice when requested. Which of these models would be best suited for the context of AI is obviously going to be an important consideration.

An example of an existing regulatory setup that has recently attracted some interest as a possible model for AI and algorithms is in the pharmaceutical industry. Although specifics vary between jurisdictions, most countries have some sort of regulatory agency in place. Perhaps the best known is the United States' Food and Drug Administration (FDA). Several writers have proposed the idea of a sort of FDA for AI.[30]

As Olaf Groth and his colleagues note, not every application of AI technology would need to be scrutinized by the regulatory agency. Instead, the agency "would need distinct trigger points on when to review and at what level of scrutiny, similar to the ways the FDA's powers stretch or recede for pharmaceuticals versus nutritional supplements."[31]

But Andrea Coravos and her colleagues are skeptical of the notion of a single AI regulator that could operate across all disciplines and use cases. Instead, they suggest that "oversight can and should be tailored to each field of application."[32] The field of healthcare, they claim, would be one field "already well positioned to regulate the algorithms within its field," whereas "other industries with regulatory bodies, such as education and finance, could also be responsible for articulating best practices through guidance or even formal regulation."[33] Coravos and her colleagues may have a point, but it's not clear how well their modular institutional setup could handle some of the novel conundrums thrown up by a radically innovative tech sector. We have talked, for example, about the risk that "bots" or recommender algorithms in social media could influence the outcomes of elections, or potentially lead people to increasingly extreme political positions. These aren't areas of life that have so far been subject to much in the way of regulatory scrutiny.*

*There are, of course, rules about influencing politics, but these are typically about matters like donations to campaigns or candidates—and even those are fairly minimal in some jurisdictions (Citizens United). But individually targeted campaign

In 2019, researchers at New Zealand's Otago University (including some of the authors of this book) proposed the creation of a specialist regulator to address the use of predictive algorithms in government agencies.[34] How would such an agency function? In an article that also made the case for an FDA-based model, Andrew Tutt considered a range of possible functions. These occupy a range of places on a scale from "light-touch" to "hard-edged." The agency could, for example

- act as a standards-setting body,
- require that technical details be disclosed in the name of public safety, and
- require that some receive approval from the agency before deployment.[35]

The last of these suggestions would, on Tutt's analysis, be restricted for the most "opaque, complex, and dangerous" uses. It could "provide an opportunity for the agency to require that companies substantiate the safety performance of their algorithms." Tutt also suggests that premarket approval could be subject to usage restrictions; as with drugs, distribution without permission or "off-label" use of AIs could attract legal sanctions.

Pre-deployment approval is worth taking seriously. Regulation pitched at the level of particular AI algorithms or programs seems likely to overlook the fact that these are in many cases highly flexible tools. An AI approved for an innocuous use could be repurposed for a much more sensitive or dangerous one.

Rules *for* AI?

The rules we have discussed so far are rules *about* AI, but they are rules that will be applied to the human beings that manufacture, sell, or use AI products. What about rules programmed *into* AIs? Is there a case for requiring certain things to be required or prohibited?

Again, most of the attention to date has focused on embodied AIs, like driverless cars and robots of various sorts, for which the risks of harm are most obvious. In the context of driverless cars, much attention was garnered by the apparent acknowledgment by Mercedes-Benz that they would program their cars to prioritize the lives of the occupants over any other

---

adverts and chatbots purporting to be human contributors to online discussions have thus far avoided much in the way of regulatory scrutiny—San Francisco's new BOT law being a very rare example.

at-risk parties.[36] The logic, expressed in an interview by Christoph von Hugo, manager of driver assistance systems and active safety, was ostensibly based on probability. "If you know you can save at least one person, at least save that one," he said. "Save the one in the car."[37]

There may be something initially attractive about this bird-in-the-hand logic, but the moral math may not so readily support it. What if the cost of saving "the one in the car" is the likely deaths of the ten on the pavement? Or the thirty on the school bus? How much more confident would the AI have to be of saving the occupant for such a judgment to be justified?

More cynically, we may wonder how many driverless cars Mercedes or any other manufacturer would sell if they promised anything else. A study published in *Science* a few months earlier had demonstrated that, although a significant number of people recognized the morality of sacrificing one occupant for ten pedestrians, far fewer wanted this for their own car.[38]

You may not find this outcome entirely surprising. Prioritizing our own well-being and that of those close to us is a common trait with obvious evolutionary benefits. The authors certainly were not surprised. "This is the classic signature of a social dilemma," they wrote, "in which everyone has a temptation to free-ride instead of adopting the behavior that would lead to the best global outcome."[39] A typical solution, they suggested, "is for regulators to enforce the behavior leading to the best global outcome." Should we have rules to prevent free riding in cases like this?

The idea of sharing the road with vehicles programmed to prioritize the lives of their own customers over everyone else seems to strike at the core of several of our moral values—an egalitarian concern for the equal value of lives, as well as a utilitarian concern with minimizing harm. It may also violate strong ethical (and legal) rules against active endangerment. The car that swerves to hit a cyclist or a pedestrian to save its occupant wouldn't just be failing to make a noble sacrifice; it would be actively endangering someone who may not otherwise have been in danger. Although criminal law has typically been slow to blame human drivers who flinch from altruistic self-sacrifice in emergencies, it may well take a different view of someone who makes such choices in the calm surroundings of a computer lab or a car showroom.

Other survival priorities, though less obviously self-serving, could be equally challenging to what we might consider our shared values. A 2018 ("moral machine") study found an intriguing range of regional and cultural differences to driverless vehicle dilemmas across a range of countries.[40]

Some common preferences (for example, sparing young people over old) are intelligible, if controversial. Some (the preference for prioritizing female lives over male) might seem archaic. And some (sparing those of higher over lower social status) are likely to strike many people as simply obnoxious.

These studies, though somewhat artificial, might suggest a need for rules to ensure that, when choices have to be made, they are not made in ways that reflect dubious preferences. Indeed, Germany has already taken steps in this direction. In 2017, the Ethics Commission of its Federal Ministry of Transport and Infrastructure published a report on automated and connected driving. The report specifically addressed dilemma cases and the sorts of rules that could be programmed for such eventualities—"In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical, or mental constitution) is strictly prohibited."

The report also said that, although "[g]eneral programming to reduce the number of personal injuries may be justifiable," it would not be permissible to sacrifice parties not "involved in the generation of mobility risks" over those who are so involved.[41] In other words, innocent pedestrians, bystanders, cyclists, and so on are not to be sacrificed to save the occupants of autonomous vehicles.

Autonomous vehicles might be the most obvious current example of this kind of challenge, but they are unlikely to be the last. Decisions that until now have been too rare or too instinctive to merit much by way of a legal response are now going to be made in circumstances in which rules can meaningfully be brought to bear. But as the "moral machine" study shows, regulators might struggle to find an ethical consensus about what rules we should agree upon, and the research to date suggests that leaving these decisions to market forces won't guarantee an outcome that's satisfactory to most of us.

Summing Up

We need new rules for AI, but they won't always need to be AI-specific. As AI enters almost every area of our lives, it will come into contact with the more general rules that apply to commerce, transport, employment, healthcare, and everything else.

Nor will we always need *new* rules. Some existing laws and regulations will apply pretty straightforwardly to AI (though they might need a bit of

tweaking here and there). Before we rush to scratch-build a new regulatory regime, we need to take stock of what we already have.

AI will, of course, necessitate new AI-specific rules. Some of the gaps in our existing laws will be easy to fill. Others will force us to revisit the values and assumptions behind those laws. A key challenge in fashioning any new regulatory response to technology is to ensure that it serves the needs of all members of society, not just those of tech entrepreneurs and their clients, or indeed those of any other influential cohort of society.