During the Middle Ages, animals could be tried for criminal offenses. There are documented stories of cases brought against chickens, rats, field mice, bees, gnats, and pigs.[1] Back then, people apparently thought animals capable of knowing right from wrong and behaving accordingly, in a way that we don't today. They believed that animals had what's called *moral agency*.

A widely accepted characterization of moral agents is that they must be capable of two things. They must be able to perceive the morally pertinent consequences of their actions, and they must be able to choose between the relevant courses of action.

Curiously, neither of these two requirements relies on any subjective, innate sense of right or wrong. It simply says that agents have to be able to control their own actions and evaluate the effects of their actions against some putative moral standard. Whether that standard is self-generated, whether they understand the theory underlying that standard, whether they agree with it or not, whether they can "feel" the difference between righteousness and sin—all that is irrelevant.

Consider the predicament of the psychopath. He or she has little or no ability to feel empathy or remorse for his or her actions. However, many if not most psychopaths are quite intelligent, certainly capable of both understanding moral concepts and controlling their own behavior accordingly—they just don't experience an emotional reaction to moral questions. Psychologists estimate that over 1 percent of the U.S. population are psychopaths.[2] And yet, we don't see one out of a hundred people running around committing crimes willy-nilly. Psychopaths may privately wonder what the big deal is, but they understand how they are supposed to behave, and most somehow manage to suck it up and get along with the rest of us.

Today we may find the medieval notion that animals can commit crimes laughable, but the modern interpretation of moral agency is hardly confined to humans.

In 2010, the oil rig Deepwater Horizon in the Gulf of Mexico suffered an underwater blowout. Eleven workers were killed, and large quantities of oil fouled the water and beaches. The federal government filed *criminal*—in addition to civil—charges against BP, the oil company that owned the rig. The company settled the charges for $4 billion. That's on top of large civil penalties and fines.

The criminal case against BP illustrates that you don't need to be conscious or sentient to have moral agency. In our legal system, a corporation is considered to have moral agency and can be held criminally liable. That is, BP was supposed to know better and be capable of doing the right things to ensure the accident didn't happen, but in this case, it failed to do so. The corporation itself, as distinct from its employees, had a duty to put sufficient controls in place to avoid incidents like this one.

So modern legal theory accepts the notion that both people and corporations can be moral agents, and therefore can be charged with crimes. How about a synthetic intellect? Can it meet the requirements for moral responsibility as well?

Yes, it can. If it is sufficiently capable of sensing the morally relevant aspects of its environment, and it has a choice of actions, it qualifies as a moral agent. These systems don't have to be very sophisticated to cross this seemingly anthropological boundary. A robotic lawnmower may be able to see that it's about to run over a child's leg, as opposed to a stick, and it may be capable of selecting whether to stop or continue. The question, of course, is how it is supposed to "know" it should stop in one case but should proceed in the other. Without some sort of guidance, we wouldn't expect it, a priori, to make a good decision.

This problem is far from theoretical. An active intellectual debate is quietly taking place sub rosa regarding how to program autonomous vehicles. It's easy to construct ethically challenging scenarios for such products that are virtually certain to occur, no matter how much we try to avoid them. Your self-driving car can run over a dog to save your life: pretty clear what you would want it to do. But what if it has to choose between running over an elderly couple or a bunch of kids crossing the street? How about a Sophie's Choice of which of your own children to kill, the one in the front seat or the one in the back? We can ignore such questions because they are so painful to consider, but that itself would be an immoral act.

Okay, so we'll grit our teeth and program in a moral code. Sounds like an engineering problem, but it's not that simple. Despite considerable attention to this topic, there's no accepted consensus among experts as to what such a moral code might look like. Over the centuries, philosophers have developed a rich panoply of ethical theories, and arguments over which is best—or even viable—that continue unabated to this day.

Even if we could reach some consensus on this difficult question, there's no reason to believe that the result could easily be reduced to practice and implemented programmatically. Some researchers in the emerging field of computational ethics, which seeks to create "artificial moral agents," have tried using a "top-down" approach. They select and implement moral principles a priori, then build systems that attempt to respect those principles (duty-based normative ethics). Others pursue a "bottom-up" strategy, relying on machine learning algorithms presented with a large collection of relevant examples. But this approach has a significant drawback. Like humans, machines are far from guaranteed to acquire and implement acceptable moral principles, much less be able to articulate them. Other approaches include "case-based reasoning," essentially, resolving moral challenges by relating them to a catalog of (hopefully similar) known cases. One challenge dogging this nascent field is that at least some of our own moral sense seems to be rooted in our human ability to feel sympathy and compassion—we instinctively reason that if something hurts us, it's probably not right for us to inflict it on others. This shortcut to ethical behavior is presumably unavailable to machines. In short, we're a long way from developing a curriculum to teach engineers the practice of moral programming.[3]

Quite aside from the issue of machine moral agency is the question of who is responsible when it makes a bad decision. To answer this, it's useful to understand the legal theory behind the relationship between "principals" and their "agents." To explore this, let's return to the BP case.

You might wonder how a corporation can commit a criminal act, as opposed to its employees committing criminal acts. Eleven people died on the Deepwater Horizon, but that doesn't mean that any particular individual was negligent or engaged in criminal activity. On the contrary, every employee may simply have carried out his or her assigned duties, and none of those duties were to kill eleven people.

The employees were the *means* by which the corporation committed the crime. By the same theory, when you hold up a bank, your legs are the *means* by which you walk into the bank. Your legs, of course, aren't criminally liable. But there's a big difference between a means of getting something done, like your legs carrying you into the bank, and the managers on the Deepwater Horizon failing to detect or correct a potentially dangerous situation. The managers are considered to be "agents" of the corporation, and so potentially shoulder some of the liability.

An agent is an independent party who is authorized, by mutual agreement, to act on behalf of a principal. Now your legs are neither an independent party nor are they in a position to knowingly enter into a mutual agreement to act on your behalf. On the other hand, an employee of BP is an independent party who can knowingly act on BP's behalf.

When acting on your behalf, your agent has what's called a fiduciary responsibility to carry out your intent and protect your interests—but only within certain limits. For instance, if your agent knowingly commits a crime on your behalf, that doesn't get him or her off the hook. If I hire you to kill my romantic rival, you share responsibility for the murder because you are presumed to understand that you are part of a conspiracy to break the law.

But what if an agent commits a crime and doesn't know he or she is doing it? I say, "Here, press this button," you comply, and a bomb goes off at the Super Bowl. You acted as my agent, but you are responsible only if you reasonably should have known the consequences.

Now let's turn this around. Suppose the agent commits a crime in the service of a principal without the principal's knowledge. I tell you to go get me $100 from the bank. You go down to the bank with a gun and hand the teller a note ordering him or her to put the money in unmarked bills into a paper bag. You return and give me the bag. Am I responsible for your theft? Under most circumstances, the answer is no. (I'm oversimplifying a bit, because if the supposedly innocent party benefited from the crime, they can also be held legally responsible even if they were unaware of it.)

There's a long history of legal principles and precedents addressing who is responsible for what in a principal-agency relationship—or, more accurately, apportioning liability between the parties when their relative responsibility is unclear.

In the BP case, the government concluded that the actions of the individual employees didn't themselves constitute criminal acts, but these acts taken in aggregate did. So it indicted BP itself, as a principal with sufficiently broad responsibilities.

So modern legal theory accepts the notion that both people and corporations can be principals and agents, and can independently be charged with crimes. How about an intelligent machine? Who is responsible when a synthetic intellect acts on your behalf? You might think the obvious answer is you, and today that's probably right. But this isn't entirely fair, and it's likely to change in the future, for good reasons.

Consider the following scenario. Imagine that you purchase a personal home robot that is capable of taking the elevator down from your tenth-floor Greenwich Village apartment, crossing the street, and purchasing a caramel flan Frappuccino for you from Starbucks. (This isn't entirely science fiction. A prototype of just such a

MORAL CODE + LIABILITY

robot was recently demonstrated at Stanford.)[4] In addition to being preprogrammed with a variety of general behavioral principles, the robot is able to hone its navigational and social skills by watching the behavior of the people it encounters. After all, customs and practices vary from place to place. It might be appropriate to shake hands with females you meet in New York, but it is forbidden in Iran unless you are related. Unbeknownst to you, your robot recently witnessed a rare event, a Good Samaritan subduing a purse snatcher until the police arrived, earning the approval and admiration of a burgeoning crowd of spectators.

On the way to fetch your coffee, your robot witnesses a man grappling with a woman, then taking her purse, over her apparent objections. It infers that a crime is taking place and, consistent with its general programming and its specific experience, it wrestles the man to the ground and detains him while calling 911.

When the police arrive, the man explains that he and his wife were merely having an animated tussle over the car keys to determine who was going to drive. His wife confirms the story. Oops! They turn their attention to your well-intentioned but hapless robot, which dutifully explains that it was merely acting on your instructions to fetch a drink. Incensed, the two insist that the police arrest you for assault.

Your defense attorney's argument is simple: you didn't do it, the robot did. You purchased the robot in good-faith reliance on its design and were using it in accordance with its intended purpose, so the company that sold you the robot should be held responsible for the incident.

But that company also has lawyers, and they successfully argue that they have met all reasonable standards of product liability and acted with due diligence and care. They point out that in millions of hours of use, this is the first event of its kind. From their perspective, this was simply a regrettable though unpredictable freak accident no different from an autonomous vehicle driving into a sinkhole that suddenly appears.

Perplexed at this liability gap, the judge looks for precedents. He finds one in the antebellum "Slave Codes" (as they were called) of the seventeenth and eighteenth centuries.[5] Prior to the Civil War, various states and jurisdictions maintained a separate (and very unequal) body of laws to govern the treatment, legal status, and responsibilities of slaves. For the most part, these codes characterized slaves as property having limited rights and protections, particularly from their owners. While we certainly believe today that southern plantation slaves were conscious human beings, deserving of the same basic human rights as all others, it's worth noting that not everyone at that time agreed with this assessment.[6] Regardless, these codes inevitably held the slaves, not the owners, legally culpable for their crimes and subjected them to punishment.

The judge in this case sees a parallel between the status of a slave—who is legal "property" but is also capable of making his or her own independent decisions—and your robot. He decides that the appropriate punishment in this case is that the robot's memory will be erased, to expunge its purse-snatching experience, and, as reparation for the crime, the robot will be consigned to the injured party's custody for a period of twelve months.[7]

The victim of the crime feels this is an acceptable resolution and will be happy to have a free, obedient servant for the next year. You are unhappy that you will temporarily lose the use of your robot and then have to retrain it, but it beats going to prison for assault.

And thus begins a new trail of precedents and body of law.

To recap, there's no requirement in our laws that a moral agent be human or conscious, as the BP Deepwater Horizon case demonstrates. The relevant entity must merely be capable of recognizing the moral consequences of its actions and be able to act independently. Recall that synthetic intellects are commonly equipped with machine learning programs that develop unique internal representations based on the examples in the training set. I use this pile of jargon to avoid the danger inherent in using anthropomorphic language, but only because we don't yet have the common words to describe these concepts any other way. Otherwise, I would simply say that synthetic intellects think and act based on their own experience, which in this case your robot clearly did. It just happened to be wrong. It may have been acting as your legal agent, but since you didn't know what it was doing, even as its principal you aren't responsible—it is.

There's only one problem. If you accept that a synthetic intellect can commit a crime, how on earth do you discipline it? The judge in this case effectively punished the robot's owner and compensated the victim, but did he mete out justice to the robot?

For guidance, consider how corporations are treated. Obviously, you can't punish a corporation the same way you can a human. You can't sentence a corporation to ten years in prison or take away its right to vote. In the words of Edward Thurlow, lord chancellor of England at the turn of the nineteenth century, "Did you ever expect a corporation to have a conscience, when it has no soul to be damned, and no body to be kicked?"[8]

The key here is that humans, corporations, and synthetic intellects all have one thing in common: a purpose or goal. (At least within the context of the crime.) A human may commit a crime for a variety of reasons, such as for material gain, to stay out of prison (paradoxically), or to eliminate a romantic competitor. And the punishments we mete out relate to those goals. We may deprive the perpetrator of life (capital punishment), liberty (incarceration), or the ability to pursue happiness (a restraining order, for instance).

When corporations commit crimes, we don't lock them away. Instead, we levy fines. Because the goal of a corporation is to make money, at least most of the time, this is a significant deterrent to bad behavior. We can also void contracts, exclude it from markets, or make its actions subject to external oversight, as is sometimes the case in antitrust litigation. In the extreme, we can deprive it of life (that is, close it down).

So we've already accepted the concept that not all perpetrators should suffer the same consequences. Not only should the punishment fit the crime, the punishment should fit the criminal. Punishing a synthetic intellect requires interfering with its ability to achieve its goals. This may not have an emotional impact as it might on a human, but it does serve important purposes of our legal system—deterrence and rehabilitation. A synthetic intellect, rationally programmed to pursue its goals, will alter its behavior to achieve its objectives when it encounters obstacles. This may be as simple as seeing examples of other instances of itself held to account for mistakes.

Note that, in contrast to most mass-produced artifacts, instances of synthetic intellects need not be equivalent, for the same reason that identical twins are not the same person. Each may learn from its own unique experiences and draw its own idiosyncratic conclusions, as our fictional robot did in the assault case.

For a more contemporary example, consider a credit card fraud detection program that uses machine learning algorithms. It may inadvertently run afoul of antidiscrimination laws by taking the race of the cardholder into account, or it may have independently discovered some other variable that is closely correlated with race. Unscrambling the digital omelet in which this knowledge is embedded may be entirely impractical, so the penalty might be to delete the entire database.

That may sound innocuous, but it's not. It could have substantial economic consequences for the bank or owner of that program, which has relied on billions of real-time transactions collected over many years to fine-tune its performance. You can bet the owner would fight hard to avoid this outcome.

But forced amnesia is not the only way to interfere with a synthetic intellect's goals. It may be possible to revoke its authority to act. In fact, the licensing of synthetic intellects to permit their use and holding them responsible for their own behavior go hand in hand.

For instance, it's likely that the government or insurance companies will review and approve each model of autonomous vehicle, pretty much as they do for all vehicles now. The same is true for computer programs that operate medical equipment, which fall under the definition of medical devices. In the future, we may revoke authority by recalling the medallion of an autonomous taxi, requiring a legal program to retake the bar exam, or deleting the account credentials from an automated trading program.

So synthetic intellects will be accorded rights (for example, in the form of licenses) and will have responsibilities (for example, to refrain from damaging the property of others), just like other entities that can sense, act, and make choices. The legal framework for this is called *personhood.*

Late-night comedians delight in making fun of the well-established legal principle that corporations are people, for instance, in the aftermath of *Citizens United v. Federal Election Commission* (2010), in which the U.S. Supreme Court affirmed that corporations are entitled to the free-speech protection of the First Amendment to the Constitution. Of course, this doesn't mean what the comedians pretend it means, that judges foolishly equate corporations with humans. It merely means that corporations have selected rights and responsibilities, and the legal shorthand for this is personhood.[9]

The functional parallels between corporations and synthetic intellects are so strong that courts will likely establish the principle that synthetic intellects can be *artificial persons* in an attempt to make sense of a patchwork of precedents like the robotic assault case described earlier. The attendant rights and responsibilities will evolve over time.

The most important of these are the right to enter into contracts and own assets. Arguably, we already permit computer-based systems to enter into contracts when they trade stocks, or when you make an online purchase. It's just that their owners are the legal entities bound by those contracts.

There will also be strong pressure to permit artificial persons to own assets because such assets can be subject to seizure or fines independent of the artificial person's owners. In the robotic assault example, the judge effectively condemned the robot to a year of servitude precisely because its own labor was the only asset it had. There was no way to order the robot to pay a fine, and presumably the judge thought this sentence better than asking the owner to pay. But if the robot had its own burgeoning bank account, it would be a very tempting target.

Owners of synthetic intellects will also favor granting contractual and property rights to artificial persons because this will have the side effect of insulating their own assets from liability—the most common motivation for forming a corporation today.

Unlike most predictions, this isn't fanciful speculation about one possible future among many. On the contrary, it will be hard to prevent, because the effect can be simulated today by wrapping each synthetic intellect in its own legal corporation, just as your lawyer or doctor might be a "professional corporation" or LLC. If I were the owner and operator of a fleet of autonomous taxis, I would seriously consider incorporating each vehicle as an asset of its own legal entity for precisely this reason; I wouldn't want a single catastrophic mistake to bankrupt my entire enterprise. Other than that, I would leave my roving minions to mint money as best they could, squirreling away their profits for me to collect like honey from a beehive.

Which leads us back to the essential problem with intelligent machines as agents. They will ruthlessly pursue the goals we assign them, outcompeting humans, and may be under our control only nominally—at least until we develop the ethical and legal framework for integrating them as productive partners into human society. As they enrich our lives, enhance our prosperity, and increase our leisure, the irresistible and undeniable benefits of all this technology will obscure a disquieting truth: synthetic intellects and forged laborers will be running around as independent agents, performing work and making money on behalf of their owners, without regard to the consequences to others or to society in general. Instead, as in the case of the HFT programs, they are likely to be skimming off the lion's share of the enormous wealth they create for the benefit of a few lucky individuals.

As you might expect, this scenario has already started. Superhuman omniscient systems observe our individual and group behavior, then guide us to what we purchase, listen to, watch, and read—while the profits quietly pile up elsewhere. You don't have to look very far to find an example of how this affects you—there's no waiting on checkout 1 in the Amazon cloud!

6 America Land of the Free Shipping