Prologue: What's All the Fuss About?

There's always something about astonishing technical feats that gives us pause. In a way they're both humbling *and* ennobling. They remind us how powerless we are without them and yet how powerful we must have been to produce them.

Sometimes what initially look like promising gadgets fizzle out and are soon forgotten (who uses a Blackberry anymore?). At other times the opposite seems to happen: an ingenious invention seems lackluster but later proves its mettle. In the twentieth century, neither the futurist H. G. Wells nor the British Royal Navy thought that submarines would amount to much.[1] Today, stealth submarines are an indispensable part of naval operations. Barely a century earlier, Charles Babbage's designs for an "Analytical Engine" were dismissed as crackpot fantasies, even though history would eventually, and spectacularly, vindicate his ambition and foresight. The Analytical Engine was essentially the world's first programmable general-purpose computer. Remarkably, Babbage's designs also anticipated the von Neumann architecture of every standard desktop in use today, complete with separate memory, central processor, looping, and conditional branching.

Every so often, however, an invention is unveiled to much fanfare, and *deservedly* so—when for once the hype is actually justified. Hype is the default setting in tech circles because new technology by its very nature tends to generate hype. The trick is to peruse the catalogue calmly and pick out only those items that justify their hype.

Of course, if we knew how to do this, we'd be in a better position to answer more profound questions about human history and destiny. For instance, at what moment in history is "a new reality" finally allowed to sink in? When does it become obvious that a Rubicon has been crossed and that things will never quite be the same again? To some extent your answer will depend on which theory of history you subscribe to. Is the history of the past hundred years, for example, a smooth, continuous narrative in which, at any point along the way, the next three to ten years might have been predicted in general outline at least? Or is the history of the last century a history of interruptions, of fits and starts, of winding roads and unforeseen eventualities? If history is smooth—perhaps not entirely predictable, but unsurprising let's say—it's going to be much harder to pinpoint significant moments because they're likely to fall beneath our radar. Maybe the election of a new leader looked so like every previous election that there was no way of telling at the time what a significant occasion it was. History ticks over more or less smoothly, on this theory, so truly epoch-making events are likely to be disguised in the ebb and flow of the familiar and mundane. On the other hand, if you subscribe to the "bumpy" theory of history, it should be easier to spot a watershed moment. On the bumpy theory, watershed moments are taking place all the time. History is forever taking unexpected courses and loves flank moves.

The truth is probably somewhere in between. History—infuriatingly? thankfully?—is both full of surprises *and* mundane. It has taken turns that nobody could have seen coming, and yet fortunes are still occasionally made on the stock market. Nothing that has happened in the past has ever quite been unthinkable. As the old saying goes, "there is nothing new under the sun." This means that, whether it's a relationship, the stock market, or the history of the entire planet that is concerned, no one is really in any better position than anyone else to state with confidence *this time things are different—from now on things are not going to be the same again*. Knowing what is momentous and what is banal, what will prevail and what will fade, is really anybody's guess. You might read Marx and Engel's *Communist Manifesto*, or Jules Verne's *Twenty Thousand Leagues Under the Sea*, or Alvin Toffler's *Future Shock* and marvel at the uncanny resemblance between what they "foresaw" and certain aspects of the modern world. But they were as often as not mistaken about the future too. And if anything, their mispredictions can be amusing. The 1982 film *Blade Runner* depicted a 2019 boasting flying cars, extraterrestrial colonization, and (… wait for it …) *desk fans*!

And yet such questions about the future exercise us whenever there looks to be a rupture in the otherwise smooth, predictable, orderly flow of time. Precisely because technology tends to generate hype, every fresh advance allures and beguiles anew, forcing us to contemplate its tantalizing possibilities. In effect, every major example of new technology insinuates a question—could *this* be the next big thing, the game-changer?

That's what we'd like to ponder in this book. Artificial intelligence, or AI for short, has generated a staggering amount of hype in the past seven years. Is it the game-changer it's been cracked up to be? In what ways is it changing the game? Is it changing the game in a *good* way? With AI these questions can seem especially difficult. On the one hand, as Jamie Susskind points out, "we still don't have robots we would trust to cut our hair."[2] On the other hand, Richard and Daniel Susskind describe a team of US surgeons who, while still in the United States, remotely excised the gall bladder of a woman in France![3]

No less important than these questions, however, are those that affect us as *citizens*. What do we, as citizens, need to know about this technology? How is it likely to affect us as customers, tenants, aspiring home-owners, students, educators, patients, clients, prison inmates, members of ethnic and sexual minorities, voters in liberal democracies?

Human vs. Machine

To start off, it's important to get a sense of what AI is, both as a field and as a kind of technology. Chapter 1 will tackle this in more detail, but it's helpful at the start to say a little about what AI hopes to achieve and how well it's going.

AI comes in many stripes. The kind that has generated most of the hype in recent years can be described in broad terms as "machine learning." One of the most prevalent uses of machine learning algorithms is in prediction. Whenever someone's future movements or behavior have to be estimated to any degree of precision—such as when a judge has to predict whether a convicted criminal will re-offend or a bank manager has to determine the likelihood that a loan applicant will repay a loan—there's a good chance a computer algorithm will be lurking somewhere nearby. Indeed, many coercive state powers in liberal democratic societies actually *require* an assessment of risk before those powers can be exercised.[4] This is true not just in criminal justice, but in areas like public health, education, and welfare. Now how best to go about this task? One way—and for a long time the *only* way—has been to rely on what's called "professional" or "clinical" judgment. This involves getting someone with lots of experience in an area (like a judge, psychologist, or surgeon) to venture their best bet on what's likely to happen in the future. (Will this criminal re-offend? What are the odds that this patient will relapse? And so on.) Professional judgment is basically trained gut instinct. But another way to approach this guessing game is to use a more formal, structured, and disciplined method, usually informed by some sort of statistical knowledge. A very simple statistical approach in the context of welfare decisions might just be to take a survey of previous recipients of unemployment benefits and ask them how long it took them before they found work. Decision makers could then use these survey results to hone their estimates of average welfare dependency and tailor their welfare decisions accordingly. But here's the rub. Although some people's intuitions are no doubt very good and highly attuned by years of clinical practice, research indicates that *on the whole* gut instinct comes a clear second to statistical (or statistically informed) methods of prediction.[5] Enter the new-fangled next-generation of machine learning algorithms. Mind you, this application of machine learning isn't exactly new—some of the predictive risk technology attracting media focus today was already being used in the late 1990s and in fact has precursors dating back decades and even centuries, as we'll discuss. But the speed and power of computing as well as the availability of data have increased enormously over the past twenty years (that's why it's called "big data"!).

There's a lot more that can be said about the relative performance of algorithms and statistics on the one hand and people and their experience-honed intuitions on the other. Suffice to say that sometimes, in specific cases, human intuition has proved to be better than algorithms.[6] Our point for now is that, on the whole, algorithms *can* make many routine predictive exercises more precise, less prone to idiosyncratic distortion—and therefore fairer—than unaided human intuition. That at any rate is the promise of AI and advanced machine learning.

Now let's consider another application of this same technology. Object classification is just what its name implies: assigning an instance of something to a particular class of things. This is what you do when you recognize a four-legged furry creature in the distance as either a cat or a dog; you are assigning the instance of the cat in front of you to the class CAT and not to the class DOG or the class CAR or the class HELICOPTER. This is something we humans do extremely quickly, effortlessly, and precisely, at least under normal viewing conditions. The other thing to say about how humans fare as object classifiers is that we tend to do so *gracefully*. A property of human object classification is what cognitive and computer scientists sometimes call "graceful degradation," meaning that when we make object classification errors, we tend to *near-miss* our target, and *almost-but not-quite* recognize the object. Even when the visibility is poor, for instance, it's easy to imagine ourselves mistaking a cow for a horse or a tractor for a buggy. But has any sane person ever mistaken a horse for an airplane or a buggy for a person? Unlikely. Our guesses are generally plausible.

What about machine learning object classifiers? Here there's a mixed bag of results. It's interesting to compare the way humans fail with the way machine learning systems fail. If we mistakenly classify a dog (say a husky) as a wolf—an easy mistake to make if you're none too familiar with huskies—what would have led you to that misclassification? Presumably you'd have focused your attention on things like the eyes, ears, and snout, and concluded, "Yep, that's a wolf." It might surprise you to learn that this eminently reasonable route to misclassification isn't the one an AI system would be obliged to take in reaching the same conclusion. An AI could just as easily focus on the shape of the *background* image—the residual image left after subtracting the husky's face from the photo. If there's a lot of snow in the background, an AI might conclude it's looking at a wolf. If there's no snow, it might decide otherwise.[7] You may think this is an odd way to go about classifying canines, but this odd approach is almost certainly the one an AI would take if most of the many thousands of images of wolves it was trained on contained snow in the background and if most of the images of huskies didn't.

If the same technology used for predicting whether someone is eligible for unemployment benefits or at risk of reoffending lies behind the classifier that concentrates on the absence or presence of snow to determine whether a canine is a husky or a wolf, we have a problem. Actually, there are several problems here.

One is what statisticians call "selection" or "sampling" bias. The classifier above had a bias in that the sample of images it was trained on had too many images of wolves in the snow. A better training set would've featured greater lupine diversity! For the same reason, a face recognition system trained on white men will have a hard time recognizing the faces of black or Asian women.[8] We'll come back to the subject of bias in chapter 3.

Another problem here is that it's often very hard to say why an AI "decides" the way it does. Figuring out that a classifier has fixed its attention on the snow around an object rather than the object itself can be a tricky business. We'll talk about algorithmic "transparency" and "explainability" in chapter 2.

Another issue is responsibility, and potentially legal *liability*, for harm. What if one of these classifiers were to be installed in an autonomous vehicle? If one day the vehicle ran into a child, thinking it was a tree, who'd be to blame? The vehicle's machine learning system may have learned to classify such objects autonomously, without being programmed *exactly* how to do it by anyone—not even its developers. Certainly the developers shouldn't get off scot-free. They would have curated the machine's training data, and (we hope!) made sure to include numerous examples of important categories like children and trees. To that extent the developers will have *steered* the machine learning system toward particular behavior. But should the developer *always* be responsible for misclassifications? Does there ever come a point when it makes sense to hold an *algorithm* responsible, or legally liable, for the harm it causes? We'll discuss these questions in chapter 4.

Then there's the issue of control. What happens when a time-pressed judge in a back-logged court system needs to get through a long list of bail applications, and a machine learning tool can save them a lot of time if its "objective" and statistically "precise" recommendations are simply accepted at face value? Is there a danger that a judge in this situation will start uncritically deferring to the device, ignoring their own reasonable misgivings about what the tool recommends? After all, even if these systems are frequently better than chance, and perhaps even better than humans, they are still far from perfect. And when they do make mistakes—as we saw they can—they often make them in strange ways and for odd reasons. We'll discuss the control problem in chapter 5.

Another (huge!) issue is data privacy. Where did all that training data come from? Whose data was it to begin with, and did they consent to their data being used for training private algorithms? If anything, the 2020 COVID-19 global pandemic raises the stakes on such questions considerably. As this book goes to press, governments the world over are contemplating (or implementing) various biosurveillance measures that would enable them to track people's movements and trace their contacts using mobile phone data. This is all very well for crisis management, but what about when the crisis is over? Experience shows that security and surveillance measures can be scaled-up fairly quickly and efficiently if needed (as they were after the 9/11 attacks). But as the phrase "surveillance creep" suggests, their reversal is not approached with anything like the same alacrity. Chapter 6 delves into these and similar sorts of data protection issues.

A final question, and perhaps the most important one of all, relates to the effects that the use of such systems will have on human autonomy and agency in the longer term. We'll investigate these effects in chapter 7, but it's worth saying just a little about this topic now because there's a lot to ponder and it'll be useful clearing up a few things from the get-go.

One of the topics that's exercised AI folk in recent years has been the effect of increasingly sophisticated AI on human dignity. The phrase "human dignity" isn't easy to define, but it seems to be referring to the *worth* or *value* of human life. The question is whether advanced AI systems diminish the worth of human life in some way. The possibility that machines could one day reproduce and even exceed the most distinctive products of human ingenuity inspires the thought that human life is no longer special. What are we to make of such concerns?

As the object classifier example illustrates, machine learning tools, even quite sophisticated ones, aren't "thinking" in anything like the way humans think. So if object classifiers are anything to go by, it doesn't look like machines will compete with us on the basis of *how* we do things, even if they can do *what* we do in different ways. For all we know, this state of affairs might change in the future. But judging from today's technological vantage point, that future seems a long way off indeed. This ought to offer some reassurance to those concerned for human dignity.

But is that enough? You might think that if a machine could come to do the kinds of things we pride ourselves on being able to do as human beings, what does it matter if a machine can perform these very same feats differently? After all, it's not *how* a mousetrap works but *that* it works that's important, isn't it? Maybe airplanes don't fly as elegantly or gracefully as eagles—but then why should that matter? Would an aircraft need to fly so like a pigeon that it would fool other pigeons into thinking it was a pigeon before we could say it could "fly"?[9] Hardly. So, to repeat, if AI can match or surpass the proudest achievements of humankind, albeit through alternative means (the way planes can rival avian flight without making use of feathers), is human dignity any the worse for that?

We don't think so. Human calculating abilities have been miles behind humble desktop calculators for many decades, and yet no one would seriously question the value of human life as a result. Even if machine learning systems start classifying objects more reliably than humans (and perhaps from vast distances, too), why should this diminish the worth of a human life? Airplanes fly, and we don't think any less of birds.

The philosopher Luciano Floridi helpfully reminds us of several other developments in history that threatened to diminish human dignity, or so people feared. He mentions how Nicolaus Copernicus, Charles Darwin, Sigmund Freud, and Alan Turing each in their own way dethroned humanity from an imagined universal centrality, and thus destabilized the prevailing and long-held "anthropocentric" view of nature.[10] Copernicus showed that we are not at the center of the universe; Darwin showed that we are not at the center of the biosphere; Freud showed that we are not at the center of the psychosphere or "space of reason" (i.e., we can act from unknown and introspectively opaque motivations); and Turing showed that we are not at the center of the infosphere. These cumulative blows weren't easy to take, and the first two in particular were violently resisted (indeed the second still so). But what these revelations did *not* do—for all the tremors they sent out—was demonstrate the idea of human dignity itself to be incoherent. Of course, if human dignity means "central," you certainly could say human dignity was imperiled by these events. But dignity and centrality aren't the same. An object doesn't have to be at the center of a painting for it to capture our attention or elicit admiration.

There's another reason why, for the foreseeable future, human dignity is likely to be unaffected by the mere fact that machines can beat us at our own game, so to speak. Every major AI in existence today is *domain-specific*. Whether it's a system for beating chess champions or coordinating transactions on the stock exchange, every stupendous achievement that has been celebrated in recent years has occurred within a very narrow domain of endeavor (chess moves, stock trades, and so on). Human intelligence isn't like this—it's *domain-general*. Our ability to play chess doesn't preclude our ability to play tennis, or write a song, or bake a pie. Most of us can do any of these things if we want to. Indeed many researchers would regard the holy grail of AI research as being able to crack this domain-general code. What is it about human minds and bodies that make them able to adapt so well, so fluidly, to such divergent task demands? Computers do seem to find easy what we find hard (try calculating 5,749,987 × 9,220,866 in a hurry). But they also seem to experience staggering difficulty with what most of us find really easy (opening a door handle, pouring cereal from a box, etc.). Even being able to hold a conversation that isn't completely stilted and stereotyped is very difficult for a machine. When we ask our roommate to pick up some milk on the way home, we know this means stop by the corner store, purchase some milk, and bring it home. A computer needs to be programmed so as *not* to interpret this instruction in any number of far more literal—and so algorithmically simple—ways, such as find milk somewhere between your present location and home; upon finding it, elevate its position relative to the ground; then restore the milk to its original position. The first (and obviously sane) interpretation requires a subtle integration of linguistic and contextual cues that we manage effortlessly most of the time (linguists call this aspect of communication "pragmatics"). This simply isn't the case for machines. Computers can do syntax well, but the pragmatic aspects of communication are still mostly beyond them (though pragmatics *is* an active area of research among computational linguists).

The same can be said of consciousness. No existing AI is presently conscious, and no one's got any real clue just how or why conscious experiences arise from vacant material processes. How do you get the felt sense of an inner life, of an internal "point of view," from a group of proteins, salt, and water? Why does anything touchy-feely have to accompany matter at all? Why can't we just be dead on the inside—like zombies? There are many theories, but the precise character of consciousness remains elusive. For all the science-fiction films that play on our worst fears of robots becoming conscious, nobody's any closer to achieving Nathan Bateman's success with "Ava" in the 2015 film *Ex Machina*.

Thus for so long as the advent of sophisticated AI is limited to domain-specific and unconscious systems, we needn't be too worried about human dignity. This isn't to say that domain-specific systems are harmless and pose no threat to human dignity for *other* reasons—reasons unconnected with an AI's ability to trounce us in a game of chess. Lethal autonomous weapons systems are obviously attended by the gravest of dangers. New data collection and surveillance software also poses significant challenges to privacy and human rights. Nor are we saying that domain-specific technologies cannot lead to revisions in how humans perceive themselves. Clearly, the very existence of a system like Google DeepMind's AlphaGo—which thrashed the human Go world-champion in a clean sweep—must change the way humans think of themselves on some level. As we noted at the start of this prologue, these systems can reasonably evoke both pride and humility. Our only point here is that human dignity isn't likely to be compromised by the mere fact that our achievements can be matched or outdone.

Is AI Rational?

There are several other issues around AI we haven't mentioned and won't be exploring in this book. We've picked out the ones we thought would be of most interest. One other issue we'd like to briefly mention concerns what we might call the *rationality* of machine learning. The fact is, many machine learning techniques aim at the discovery of hidden relationships in data that are generally too difficult for humans to detect by themselves. That is, their whole modus operandi is to look for *correlations* in the data. Correlations are a legitimate source of knowledge, to be sure, but as most of us learn in high school: *correlation is not causation*. Two clocks might always strike twelve at the same time each day, but in no sense does one clock's striking cause the other clock's striking. Still, as long as a correlation is reliable—the technical term is "statistically significant"—it can provide actionable insights into the world despite not being causal. For example, it might be discovered that people who prioritize resistance training tend to make particular nutrition choices, eating on the whole more meat and dairy than those who prioritize aerobic forms of exercise. A fitness club could reasonably go off such insights when deciding what sorts of recipes to include in its monthly health magazine—even if, strictly speaking, no causal link between the type of exercise people do and the type of food they eat can be firmly established. Perhaps weight-trainers would personally prefer plant-based proteins, but as a demographic, it's simply easier for them to source animal-based proteins.

But at this point you might wonder: What if the correlations are reliable but also utterly bizarre? What if an algorithm discovers that people with small shoe sizes eat a particular type of breakfast cereal or that people with a certain hair color are more prone to violence than others? Such relationships could, of course, be merely incidental—flukes arising from poor quality or insufficiently large training data. But let's put that concern to one side. It's true that a machine learning system that finds bizarre correlations is likely to have been mistrained in ways familiar to statisticians and data scientists. But the whole point of a "statistically significant" correlation is that it accounts for—and, in theory, rules out—this possibility.

These aren't altogether fanciful illustrations, mind you. One of the benefits of unsupervised machine learning (see chapter 1) is that it can detect the sorts of correlations that we ourselves would never think to unearth. But in English law, if a public official were to decide a case on the basis of correlations as apparently spurious as these, the decision would be quashed. In a passage well-known to English and Commonwealth lawyers, Lord Greene once said that an official

> must exclude from his consideration matters which are irrelevant to what he has to consider. … Similarly there may be something so absurd that no sensible person could ever dream that it lay within the powers of the authority. Warrington LJ … gave the example of the red-haired teacher, dismissed because she had red hair. … It is so unreasonable that it might almost be described as being done in bad faith.[11]

But, we suspect, *just these sorts of correlations* are likely to proliferate with the steady march of machine learning in all areas of public and private decision making: correlations that would seem untenable, illogical, and even drawn in bad faith if a human had been behind them. To be clear, we're not saying that such correlations will defy all attempts at explanation—we're willing to bet that a proper explanation of them will, in many cases, be forthcoming. But what are we to do in the meantime, or if they turn out *not* to be intuitively explicable after all? Let's imagine that an algorithm discovers that people who like fennel are more likely to default on their loans. Would we be justified in withholding credit from people who like fennel? Last time we checked, "liking fennel" wasn't a protected attribute under antidiscrimination law, but *should* it be? Denying someone a loan because they reveal a penchant for fennel certainly *looks* like a kind of discrimination: "liking fennel" could well be a genetically determined trait, and the trait certainly *seems* irrelevant to debt recovery. But what if it's not irrelevant? What if there's actually something to it?

In these and other ways, machine learning is posing fresh challenges to settled ways of thinking.

Citizen vs. Power

So if the biggest challenge posed by AI this century isn't the rise of a conscious race of robots, the brief for us as authors is to produce a book whose sweep is necessarily political rather than technical. We'll cover only so much of the technical background as is required to make the political issues come to life (in chapter 1).

A book with a political bent, of course, can't claim to be very useful if it doesn't offer at least a few suggestions about what might be done about the problems it diagnoses. Our suggestions will be liberally sprinkled throughout, but it's not until we hit chapter 10 that we'll discuss regulatory possibilities in more detail. As long ago as 1970, Alvin Toffler, the American futurist, wrote of the need for a technology ombudsman, "a public agency charged with receiving, investigating, and acting on complaints having to do with the irresponsible application of technology."[12] In effect he was calling for "a machinery for screening machines."[13] Well, that was 1970, a time when only the faintest rumblings of disruption could be heard and then only by the most acute and perceptive observer (as Toffler himself was). Today, the need for regulatory responses in some form or another—indeed probably taking a variety of forms—is not just urgent; it's almost too late. The fractiousness and toxicity of public discourse, incubated in an unregulated social media environment, have already contributed to the malaise of our times—reactionary politics, the amplification of disreputable and sectarian voices, and atavistic nihilism. So today we need to think much more boldly than we might have done in 1970. Indeed, the COVID-19 pandemic has made one thing brutally apparent: some challenges may require interventions so drastic that they could be seen to herald a new understanding of the relationship between a state and its citizens.