5 Control

Just about everybody who champions artificial intelligence agrees that there's a limit to how far AI should be allowed to usurp human jurisdiction.[1] That limit, understandably enough, is human destiny. Humanity should be able decide for itself what the ends of human life are. This is so, even though one of the most obvious things about ends is that humans rarely seem to agree on what they should be (beneath the level of platitudes, of course). The idea here is that AI should never *ultimately* replace human wishes, in some deep sense of "ultimate." So, even if people drawn from different demographics and geographically widely-dispersed cultures don't see eye-to-eye on the big questions, most of these same people would agree that it should in some sense be *up to people* to devise answers to these questions.

This chapter isn't about values or theories of value, human flourishing, and the "life well lived." But it does proceed on the reasonable assumption that, in an important sense, humans should always ultimately be in charge of what happens to them, that even if we decide to use labor-saving, creativity-liberating, and error-minimizing technology, we'll do so on *our* terms—that is, only so long as the technology conforms to our wishes. After all, that's what it means for a system to be under ultimate human control—for it to behave the way it should, the way we *want* it to behave, even if we aren't in moment-by-moment "operational" control of the system (and let's face it, why *would* we have operational control of an "autonomous" system?). The corollary of this setup, of course, is that if the system goes rogue, it's our prerogative to "switch it off."

This desire for humans to be in charge gets cashed out in various ways, but probably the most popular catchphrase at the moment—especially in discussions about the military deployment of lethal autonomous weapons—is the call for "meaningful human control." This is a more demanding

requirement than ultimate control. Having ultimate control doesn't mean that we can prevent mishaps or calamities in the event something goes horribly wrong with an AI. It doesn't even necessarily mean that we can mitigate the worst results in a crisis. All it really means is that we have the power to "switch off" a system that has gone rogue to prevent further damage. But by this stage, the worst might have already happened. By contrast, for an autonomous system to be under *meaningful* control, something stronger than ultimate control is required (otherwise what would make it "meaningful"?). We think meaningful control implies that the system is under *effective* control, so that operators *are* in a position to prevent the worst from happening and thereby mitigate or contain the potential fallout. So, whereas ultimate control allows us merely to reassert our hegemony over an AI to which we've voluntarily ceded operational control, *meaningful* control means that we can reassert this hegemony *effectively*, that is, in sufficient time to avert a catastrophe (for example). We agree that this is a standard worth aspiring to, and from here on in we're going to assume that humans shouldn't give up meaningful control over an autonomous system, and certainly never when the stakes are high.

This simple rule of thumb, however, isn't always easy to adhere to in practice, and the main obstacle is probably psychological. In this chapter, we're going to take a look at some of the problems identified by researchers in the field of industrial psychology known as "human factors." What human factors research reveals is that in some situations—essentially, when autonomous systems reach a certain threshold of reliability and trustworthiness—to relinquish operational control *is* to relinquish meaningful control. Put simply, once humans are accustomed to trust a system that's reliable *most* of the time (but not *all* of the time), humans *themselves* tend to "switch off," falling into a sort of "autopilot" mode where diffidence, complacency and overtrust set in. Humans with this mindset become far less critical of a system's outputs, and, as a result, far less able to spot system failures. Telling ourselves that we retain "ultimate" control in these circumstances because we can always just press the "stop" button if we want to is a little bit delusional (it must be said).

The challenge is vividly illustrated by lethal autonomous weapons systems (LAWS). In the LAWS literature, one finds a tentative distinction drawn between humans being "*in* the loop," "*on* the loop," and "*off* the loop." Being "*in* the loop" means a human gets to call the shots (literally) and say whether a target should be tracked and engaged. The buck stops

squarely with the human who decides to attack. A system with a human "*on* the loop," like a drone, would identify and track a target but not actually engage without human approval. On the other hand, if the drone handled *everything*, from identifying and tracking all the way through to engaging a target (without human intervention), the system would be fully autonomous. In that event, the human authority would be "*off* the loop."

These alternatives represent a sliding scale of possibilities. The human factors issues we mentioned arise somewhere between humans being *on* and *off* the loop—that is, a point where human agents are still technically *on* the loop but so disengaged and inattentive that they might as well be considered *off* the loop. In the LAWS context, it's not hard to imagine what the loss of meaningful control could entail. An autonomous weapon that is ill-advisedly trusted to discriminate between enemy combatants and civilians or to apply only such force as is necessary in the circumstances (the principle of "proportionality" in the law of armed conflict) could cause unspeakable devastation. Think how easy it would be for an autonomous weapon to misclassify a child rushing toward a soldier with a stick as a hostile combatant. Remember the wolves and huskies problem from the prologue? Translated to the sphere of war, object classifier errors are no laughing matter. And without meaningful human control over such systems, who's to say these horrifying possibilities wouldn't materialize?

We'll leave discussion of LAWS behind now, though, because they raise a host of other issues too niche to be addressed in a general chapter on control (e.g., what if the possibility of "switching off" an autonomous weapon—the one protocol we should *always* be able to count on in an emergency—also fails us because an enemy has hacked our console? Should the decision on who gets to live or die ever be left to a machine? Etc.). Instead, let's consider some more humdrum possibilities of autonomous systems getting out of control. Take criminal justice, a forum in which machine learning systems have been taken up with real panache, assisting in everything from police patrol and investigations to prosecution decisions, bail, sentencing, and parole. As a recent French report into artificial intelligence notes, "it is far easier for a judge to follow the recommendations of an algorithm that presents a prisoner as a danger to society than to look at the details of the prisoner's record himself and ultimately decide to free him. It is easier for a police officer to follow a patrol route dictated by an algorithm than to object to it."[2] And as the AI Now Institute remarks in a recent report of its

own, "[w]hen [a] risk assessment [system] produces a high-risk score, that score changes the sentencing outcome and can remove probation from the menu of sentencing options the judge is willing to consider."[3] The Institute's report also offers a sobering glimpse into just how long such systems can go without being properly vetted. A system in Washington, D.C., first deployed in 2004 was in use for fourteen years before it was successfully challenged in court proceedings. The authors of the report attributed this to the "long-held assumption that the system had been rigorously validated."[4] In her book, *Automating Inequality*, Virginia Eubanks notes the complacency that high tech decision tools can induce in the social services sector. Pennsylvania's Allegheny County introduced child welfare protection software as part of its child abuse prevention strategy. The technology is supposed to assist caseworkers in deciding whether to follow up on calls placed with the County's child welfare hotline. In fact, however, Eubanks relates how caseworkers would be tempted to adjust their estimates of risk to align with the model's.[5]

When complaints have been made about these systems—and here we mean *formal* complaints, indeed in some of the highest courts of appeal—the remarks of judges often suggest that the full scale of the challenge hasn't really sunk in. Of all the uses of algorithmic tools in criminal justice, perhaps the most scrutinized and debated has been COMPAS (which we introduced in chapter 3). First developed in 1998 by the company Northpointe (now Equivant), COMPAS is used by criminal justice agencies across the USA.[6] In 2016, Eric Loomis launched an unsuccessful legal challenge to the use of COMPAS in the determination of his sentence. Without going into the details here, what's interesting is that, of all the concerns the appeal court expressed about the use of such tools, it obviously *didn't* think the control issue was a major one—judging by how casually the court's provisos on the future use of COMPAS were framed.

The court noted that sentencing judges "must explain the factors in addition to a COMPAS risk assessment that independently support the sentence imposed. A COMPAS risk assessment is only one of many factors that may be considered and weighed at sentencing."[7] The court also required that sentencing judges be given a list of warnings about COMPAS as a condition of relying on its predictions.[8] These warnings draw attention to controversies surrounding use of the tool and the original motivations behind its development—it was primarily devised as an aid to post-sentencing decision-making (like parole) rather than sentencing per se.

But that's it! The scale of the human factors challenge apparently hadn't dawned on anyone. (And why would it? They're judges, not psychologists!) A sentencing judge might well fish around for, and believe themselves to be influenced by, other factors "that independently support the sentence imposed." But if the control problem is taken seriously—in particular, the problem of "automation complacency" and "automation bias," which we'll describe in a moment—this strategy offers only false reassurance. The warnings themselves were mild. And even if they were more pointed, the truth is that we simply don't know if this sort of guidance is enough to knock a judge out of their complacency. Some research suggests that even explicit briefings about the risks associated with the use of a particular tool won't mitigate the strength of automation bias.[9]

Our worries aren't academic. Warning a judge to be skeptical of an automated system's recommendations doesn't tell them *how* to discount those recommendations. It's all very well being told that a system's recommendations are not foolproof and must be taken with a pinch of salt, but if you don't know *how* the system functions and *where* and *why* it goes astray, what exactly are you supposed to do? At what point *and in what way* is a judge meant to give effect to their skepticism? Research in cognitive psychology and behavioral economics (some of it discussed in chapters 2 and 3) also points to the effects of "anchoring" in decision-making. Even weak evidence can exert an unwholesome influence on a decision maker who is trying to be objective and fair. A judge that is given a high-risk score generated by a machine with fancy credentials and technical specifications may lean toward a higher sentence unwittingly under the force of anchoring. We just don't know if warnings and a duty to take other factors into account are powerful enough, *even in combination*, to negate such anchoring effects.

The Control Problem Up Close

What we're calling "the control problem" arises from the tendency of the human agent within a human-machine control loop to become complacent, over-reliant, or overtrusting when faced with the outputs of a reliable autonomous system. Now at first blush this doesn't seem like an insurmountable problem, but it's harder to manage than it seems.

The problem was first recognized in the 1970s,[10] but it didn't receive a definitive formulation until a little paper came along in 1983 with the

succinctly telling title: "Ironies of Automation." The author was Lisanne Bainbridge, and the chief irony she grappled with was this, "that the more advanced a control system is, so the more crucial may be the contribution of the human operator."[11] Although writing at a time before deep learning had anything to do with algorithmically automated decision tasks, what she had to say about the role of the human agent in a human-machine system still rings true today.

> If the decisions can be fully specified, then a computer can make them more quickly, taking into account more dimensions and using more accurately specified criteria than a human operator can. There is therefore no way in which the human operator can check in real-time that the computer is following its rules correctly. *One can therefore only expect the operator to monitor the computer's decisions at some meta-level, to decide whether the computer's decisions are "acceptable."*[12]

As we see things, this residual monitoring function of the human operator generates at least four kinds of difficulties that should be treated separately (see box 5.1). The first relates to the cognitive limits of human processing power (the "capacity problem"). As Bainbridge put it, "If the computer is being used to make the decisions because human judgment and intuitive reasoning are not adequate in this context, then which of the decisions is to be accepted? The human monitor has been given an impossible task."[13]

Humans are often at a severe cognitive disadvantage vis-à-vis the systems they are tasked with supervising. This can be seen very clearly in the case of high-frequency financial trading. It's impossible for a monitor to keep abreast of what's happening in real time because the trades occur at speeds that vastly exceed the abilities of human monitors to keep track. As Gordon Baxter and colleagues point out, "[i]n the time it takes to diagnose and repair [a] failure … many more trades may have been executed, and possibly have exploited that failure."[14] Similar issues arise from the use of autopilot systems in aviation that are becoming "so sophisticated that they only fail in complex 'edge cases' that are impossible for the designers to foresee."[15]

The second difficulty relates to the *attentional* limits of human performance (the "attentional problem").

> We know from many "vigilance" studies … that it is impossible for even a highly motivated human being to maintain effective visual attention toward a source of information on which very little happens, for more than about half an hour. This means that it is humanly impossible to carry out the basic function of monitoring for unlikely abnormalities. …[16]

Box 5.1
Breaking Down the Control Problem

The control problem breaks down into four more basic problems:

1. The Capacity Problem

Humans aren't able to keep track of the systems they're tasked with supervising because the systems are too advanced and operate at incredible speeds.

2. The Attentional Problem

Humans get very bored very quickly if all they have to do is monitor a display of largely static information.

3. The Currency Problem

Use it or lose it. Skills that aren't regularly maintained will decay over time.

4. The Attitudinal Problem

Humans have a tendency to overtrust systems that perform reliably *most* of the time (even if they are not reliable *all* of the time).

Automation has a significant impact on situation awareness.[17] For example, we know that drivers of autonomous vehicles are less able to anticipate takeover requests and are often ill prepared to resume control in an emergency.[18]

The third difficulty relates to the *currency* of human skills (the "currency problem"). Here is Bainbridge, again: "Unfortunately, physical skills deteriorate when they are not used. … This means that a formerly experienced operator who has been monitoring an automated process may now be an inexperienced one."[19]

The fourth and final difficulty, and the one we've chosen to focus on in this chapter, relates to the *attitudes* of human operators in the face of sophisticated technology (the "attitudinal problem"). Except for a few brief remarks,[20] this problem wasn't really addressed in Bainbridge's paper.[21] It has, however, been the subject of active research in the years since.[22] Here the problem is that as the quality of automation improves and the human operator's role becomes progressively less demanding, the operator "starts to assume that the system is infallible, and so will no longer actively monitor what is happening, meaning they have become complacent."[23] Automation complacency often co-occurs with automation *bias*, when human

operators "trust the automated system so much that they ignore other sources of information, including their own senses."[24] Both complacency and bias stem from *overtrust* in automation.[25]

What makes each of these problems especially intriguing is that each gets worse as automation *improves*. The better a system gets, the more adept at handling complex information and at ever greater speeds, the more difficult it will be for a human supervisor to maintain an adequate level of engagement with the technology to ensure safe resumption of manual control should the system fail. When it comes to the current ("SAE Level 2") fleet of autonomous vehicles* that allow the driver to be hands- and feet-free (but not *mind*-free, because the driver still has to watch the road), legendary automotive human factors expert Neville Stanton expressed the conundrum wryly: "Even the most observant human driver's attention will begin to wane; it will be akin to watching paint dry."[26] And as far as complacency and bias go, there is evidence that operator trust is directly related to the scale and complexity of an autonomous system. For instance, in low-level partially automated systems, such as SAE Level 1 autonomous vehicles, there is "a clear partition in task allocation between the driver and vehicle subsystems."[27] But as the level of automation increases, this allocation gets blurred to the point that drivers find it difficult to form accurate assessments of the vehicle's capabilities, and on the whole are inclined to overestimate them.[28]

These results hold in the opposite direction too. *Decreases* in automation reliability generally seem to *increase* the detection rate of system failures.[29] Starkly put, automation is "most dangerous when it behaves in a consistent and reliable manner for most of the time."[30] Carried all the way, then, it seems the only safe bet is to use dud systems that don't inspire overtrust, or, on the contrary, to use systems that are provably better-than-human at particular tasks. The latter option is feasible because once a system is provably better (meaning less error prone) than a human agent at performing a particular task, an attentive human supervisor over that system will be superfluous. Whether there's a human keeping watch over the system or

*The Society of Automotive Engineers (SAE) framework, running from Level 0 (no automation) to Level 5 (full automation) classifies vehicles in accordance with the degree of system functions that have been carved out for automation (SAE J3016 2016). Tesla Autopilot and Mercedes Distronic Plus (Level 2) require the driver to monitor what is going on throughout the whole journey, whereas Google's self-driving car does everything except turn itself on and off.

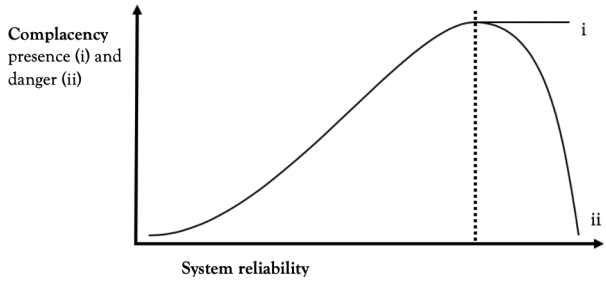

Figure 5.1
Presence (i) and danger of complacency (ii) as a function of system reliability. The dashed line represents near perfect (better-than-human) reliability.

not—and if there is, whether the human succumbs to automation complacency or not—it won't really matter, because the system will have fewer error rates than the poor human struggling to keep up.

Figure 5.1 depicts both the presence and danger of complacency as a function of system reliability. Notice that at a certain point of reliability (represented by the dashed line), the presence of complacency no longer matters because it will no longer pose a danger.

Using "Better-Than-Human" Systems: Dynamic Complementarity

If the question is interpreted literally, the answer to "Can the control problem be solved?" appears to be straightforwardly negative. The control problem can't *literally* be solved. There is nothing we can do, as far as we know, that *directly* targets, much less directly curbs, the human tendency to fall into automation complacency and bias once an autonomous system operates reliably most of the time, and when the only role left for the operator is to monitor its largely seamless transactions. However, by accepting this tendency as an obstinate feature of human-machine systems, we might be able to work around it without pretending we can alter constraints imposed by millions of years of evolution.

The insights of human factors research are instructive here. One important human factors recommendation is to foster mutual accommodation between human and computer competencies through a *dynamic* and *complementary* allocation of functions. Humans should stick to what they do best, such as communication, abstract reasoning, conceptualization, empathy, and

intuition; computers can do the rest.[31] At the same time, the allocation should be flexible enough to support *dynamic* interaction whenever it would contribute to optimal performance (or is otherwise necessary), with hand-over and hand-back for some tasks, for instance, when a driver disengages cruise control and thereby resumes control of acceleration. This assumes that some decisions can be safely handled by both humans and computers and that humans and computers have shared competencies within particular subdomains. Hand-over and hand-back may also go some way toward alleviating the currency problem, as operators are thereby afforded an opportunity to practice and maintain their manual control skills.

The obvious assumption behind this approach is that decision tasks can be cut more or less finely. We can assume, for instance, that *border control* (i.e., whether to admit, or not admit, persons moving between state boundaries) is one big decision involving customs clearance, passport verification, drug detection, and so on. We can also assume that either: (a) the *entire* border control decision is handled by one large, distributed border control software package, or (b) that only some automatable subcomponents of the overall decision have been carved out for discrete automation, the rest being left to human controllers. Currently, of course, border control decisions are only partially automated. SmartGate allows for fully automated electronic passport control checks, but customs officials still litter most immigration checkpoints. And that's the point—their job is to handle only those parts of the overall decision that can't be efficiently automated. Maybe one day the whole decision chain *will* be automated, but we're not there yet.

Under a regime of dynamic and complementary allocation, obviously some autonomous systems will replace human agents and be left to operate without supervision. Human-machine systems that contain automated subcomponents work best when the human operator is allowed to concentrate their energies on chunks of the task better suited to human rather than autonomous execution. But obviously—as we've already intimated—this setup only avoids the control problem if the automated subroutines are handled by systems approaching near-perfect (better-than-human) dependability. Otherwise the autonomous parts might work very well most of the time but still require a human monitor to track for occasional failures—and it's clear where *this* path leads.

It's reasonable to ask, though: how many autonomous systems actually reach this threshold? Truth is, it's difficult to say. SAE Level 2 (and higher)

autonomous vehicles certainly don't yet approach this kind of reliability.[32] But many subcomponents within standard (nonautonomous, SAE Level 0) vehicles clearly do, such as automatic transmission, automatic light control, and first-generation cruise control.[33]

In more typical decision support settings, arguably diagnostic and case prediction software are approaching this better-than-human standard. For instance, there are AI systems that can distinguish between lung cancers and give prognoses more accurately than human pathologists armed with the same information, and systems that can spot Alzheimer's with 80 percent accuracy up to a decade before the first appearance of symptoms, a feat surely outperforming the best human pathologists.[34] In the legal sphere, advances in natural language processing and machine learning have facilitated the development of case prediction software that can predict, with an average 79 percent accuracy, the outcomes of cases before the European Court of Human Rights when fed the facts of the cases alone.[35] Most impressively, a similar system had better luck predicting the rulings of the US Supreme Court than a group of eighty-three legal experts, almost half of whom had previously served as the justices' law clerks (60 percent vs. 75 percent accuracy).[36] Beyond these reasonably clear-cut cases we can only speculate. One advantage of dynamic complementarity is precisely that, by carving up a big decision into smaller and smaller chunks, the more likely we'll be able to find a better-than-human system to take up the baton.

And what if the sort of better-than-human accuracy we have in mind here can't be assured? The upshot of our discussion is that a decision tool shouldn't replace a human agent, at least in a high-stakes/safety-critical setting, unless the tool reaches a certain crucial threshold of reliability. But what if this standard can't be met? Can less-than-reliable systems be deployed? The short answer is yes. As we've noted, the control problem doesn't arise from the use of patently suboptimal automation, only from *generally* dependable automation. So, depending on the circumstances, a *less*-than-reliable system might safely replace a human agent charged with deciding some matter within a larger decision structure (e.g., passport verification within the larger border control decision structure). The problem with a tool like COMPAS (among other problems) is that it straddles the line between reliability in particular settings and overall optimality. It's not reliable enough to meet the better-than-human standard, but it's still useful in some ways. In other words, it's exactly the kind of tool liable to induce

automation complacency and bias, and it does so in a high-stakes setting (bail, sentencing, and parole decisions).

Are There Other Ways to Address the Control Problem?

There is some evidence that increasing accountability mechanisms can have a positive effect on human operators whose primary responsibility is monitoring an autonomous system. An important study found that "making participants accountable for either their overall performance or their decision accuracy led to lower rates of automation bias."[37] This seems to imply that if the threat of random checks and audits were held over monitors, the tendency to distrust one's own senses might be attenuated. What effects these checks could have on other aspects of human performance and job satisfaction is a separate question, as is the question of how accountability mechanisms affect automation *complacency* (as opposed to *bias*). More creative accountability measures, such as "catch-trials," in which system errors are deliberately generated to keep human invigilators on their toes, could also be useful in counteracting automation bias. Catch-trials are quite popular in aviation training courses. The aviation industry is actually a fine example of how to manage automation bias, since automation bias is a life-threatening problem in this arena, and taken very seriously. But in any case, much like other touted solutions to the control problem, they don't offer a literal, *direct*, solution. Rather, they render systems that are mostly dependable (but not better than human) *less reliable by stealth* (as it were), capitalizing on the premise that less reliable systems don't induce the same complacency and bias that attend more reliable systems.

What about teamwork? Might having a *group* of humans in the loop, working together and keeping watch on one another, alleviate automation bias? Apparently not.

> Sharing monitoring and decision-making tasks with an automated aid may lead to the same psychological effects that occur when humans share tasks with other humans, whereby "social loafing" can occur—reflected in the tendency of humans to reduce their own effort when working redundantly within a group than when they work individually on a given task. … Similar effects occur when two operators share the responsibility for a monitoring task with automation.[38]

Finally, guidelines recommending that decision makers exercise their own judgment *before* consulting an algorithm could assist in offsetting some of

the effects of automation complacency and bias. In these cases, the algorithm would serve merely as a check on a decision maker's intuitions. Note that this approach comes close to telling decision makers *not* to use the algorithm. So again, it's not so much a *solution* to the problem—a way of directly targeting and curbing an ingrained psychological bent—as it is a way of managing, negotiating, and (in this case) *avoiding* the problem.

The Take-Home Message

Automation introduces more than just automated parts; it can transform the nature of the interaction between human and machine in profound ways. One of its most alarming effects is to induce a sense of complacency in its human controllers. So among the factors that should be considered in the decision to automate any part of an administrative or business decision is the tendency of human operators to hand over meaningful control to an algorithm just because it works well in most instances. It's this problem, not machines taking over per se, that we really have to watch out for.