Minds and Machines
https://doi.org/10.1007/s11023-019-09513-7

Algorithmic Decision-Making and the Control Problem

John Zerilli[1] · **Alistair Knott**[2] · **James Maclaurin**[3] · **Colin Gavaghan**[4]

Received: 23 January 2019 / Accepted: 3 December 2019

Abstract
The danger of human operators devolving responsibility to machines and failing to detect cases where they fail has been recognised for many years by industrial psychologists and engineers studying the human operators of complex machines. We call it "the control problem", understood as the tendency of the human within a human–machine control loop to become complacent, over-reliant or unduly diffident when faced with the outputs of a reliable autonomous system. While the control problem has been investigated for some time, up to this point its manifestation in machine learning contexts has not received serious attention. This paper aims to fill that gap. We argue that, except in certain special circumstances, algorithmic decision tools should not be used in high-stakes or safety-critical decisions unless the systems concerned are significantly "better than human" in the relevant domain or subdomain of decision-making. More concretely, we recommend three strategies to address the control problem, the most promising of which involves a *complementary* (and potentially *dynamic*) coupling between highly proficient algorithmic tools and human agents working alongside one another. We also identify six key principles which all such human–machine systems should reflect in their design. These can serve as a framework both for assessing the viability of any such human–machine system as well as guiding the design and implementation of such systems generally.

Keywords Control · Artificial intelligence · Human-in-the-loop · Human–machine systems · Human–computer interaction · Human factors · Ironies of automation · Machine learning

✉ John Zerilli
jz303@cam.ac.uk

[1] Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

[2] Department of Computer Science, University of Otago, Dunedin, New Zealand

[3] Department of Philosophy, University of Otago, Dunedin, New Zealand

[4] Faculty of Law, University of Otago, Dunedin, New Zealand

Published online: 11 December 2019

1 Introduction

The recent trend toward automation by machine learning and artificial intelligence systems raises in a new guise old questions about the role of humans in human–machine systems (Johannsen 1982; Margulies and Zemanek 1982). The difficulties involved in sharing control with computer systems are of course familiar to industrial psychologists and systems analysts investigating the human operators of complex machines. In these contexts, the danger of human operators devolving responsibility to machines and failing to detect cases where they fail has been recognised for many years. The problem is that, as automation becomes smarter and cheaper, its operators have to assume an increasingly supervisory role (Meister 1999; Strauch 2018). In aviation, for example, the role of the pilot appears to have become easier, but a closer look reveals that the pilot's role has been transformed rather than simplified, with the pilot now performing a crucial monitoring function (Baxter et al. 2012; cf. Stanton 2015). Likewise in financial trading, "[t]he human trader's role is now largely one of setting strategies and monitoring their execution" (Baxter et al. 2012, p. 68). How does this shift from operator to supervisor affect the person who has undergone the shift, and the nature of the interaction between operator and machine?

What we shall term "the control problem" arises from the tendency of the human agent within a human–machine control loop to become complacent, overreliant or unduly diffident when faced with the outputs of a reliable autonomous system. Although it might be thought innocuous, decades of research confirm that the problem is actually pernicious, and perhaps even intractable (Banks et al. 2018b; Cunningham and Regan 2018; Greenlee et al. 2018). Somewhat alarmingly, it seems to afflict experts as much as novices, and is largely resistant to training (see Parasuraman and Manzey 2010 for reviews). Its effects may also be observed beyond the limits of strictly sociotechnical systems. For instance, it is well known that police officers, judges and jurors frequently overestimate the importance of forensic evidence—the so-called "CSI effect" (Marks et al. 2017; see also Damaška 1997).

Our interest is in how the control problem bears upon the proliferation of the newer types of machine learning systems. Machine learning is a form of data processing that extracts statistical patterns from large amounts of information. One of its more prominent applications is in the arena of decision support. State-of-the-art decision support systems are increasingly run on vast datasets (so-called "big data") and exploit an especially powerful form of machine learning known as "deep learning". Deep learning has two features worth noting here: first, as a form of machine learning, it is not in the mould of more traditional expert systems, which were programmed "by hand" to compute solutions to well-defined problems in a more or less deterministic manner; second, deep learning relies on a computational architecture modelled on the neurons and synapses of actual, biological brains—although obviously in a highly simplified form. These features are worth drawing attention to because they underscore how sophisticated the

autonomous systems under human care can be; and system complexity has been an abiding theme of research into the control problem from its beginning.

While the control problem has been investigated for some time, up to this point its manifestation in machine learning contexts has not received serious attention. This is significant, not because the problem necessarily has any distinctive characteristics in a machine learning context, but because there is a risk that lessons learned elsewhere will go unheeded in this new arena—an arena we have every reason to believe brings with it all the psychological pitfalls of earlier systems of industrial-scale automation. This paper aims to fill that gap by offering a critical analysis of the problem in light of the advent of sophisticated machine learning techniques. As a recent French report into artificial intelligence notes, "it is far easier for a judge to follow the recommendations of an algorithm which 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" (Villani 2018, p. 124). 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" (AI Now 2018, p. 13). 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 14 years before it was successfully challenged in court proceedings, the authors of the report attributing this to the "long-held assumption that the system had been rigorously validated" (AI Now 2018, p. 14). In her book, *Automating Inequality*, Virginia Eubanks (2017) 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 deciding whether to follow up 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. The proliferation of advanced machine learning tools in both government and private sector agencies clearly behoves us to examine the control problem in the unique context in which it now arises.

In addressing the problem, we have drawn on the literatures of both industrial psychology/engineering on the one hand—primarily "human factors" research (see below)—and artificial intelligence on the other. We argue that, except in certain special circumstances (in which great care must be taken), algorithmic decision tools should not be used in high-stakes or safety-critical decisions unless the systems concerned are significantly "better than human" in the relevant domain or subdomain of decision-making—a position towards which an increasing number of human factors experts have somewhat reluctantly been driven: see e.g. Banks et al. (2018b), Cunningham and Regan (2018), Walker et al. (2015), Cebon (2015). More concretely, we recommend three strategies to address the control problem, the most promising of which involves a *complementary* (and potentially *dynamic*) coupling

between highly proficient algorithmic tools and human agents working alongside one another. For any complex task, the choice between using a sophisticated system that makes only occasional errors but which reduces the human role to that of a monitor, and a simpler system that is reliable effectively 100% of the time but requires ongoing human participation to get the job done, can only be resolved on a case-by-case basis. In high-stakes settings, however, it is generally not advisable to choose the more sophisticated system unless it makes considerably fewer errors than a proficient human expert. Even though no technology is really 100% reliable, the dangers posed by human complacency diminish practically to zero the moment a system approaches a certain (admittedly very high) threshold of reliability. How many sophisticated systems actually reach this threshold is another question. For example, as at the time of writing, autonomous vehicles do not approach this level of capability (Banks et al. 2018a, b), but many subcomponents within standard (nonautonomous) vehicles clearly do, such as automatic transmission, automatic light control and first generation cruise control (Walker et al. 2015).[1] In more typical decision support settings, arguably diagnostic and case prediction software are approaching this better-than-human standard. There are at present AI systems which can spot early-onset Alzheimer's disease from control patients with over 80% accuracy up to a decade before the first appearance of symptoms, a feat vastly outperforming the ablest human pathologist attempting anything similar (Amoroso et al. 2017).[2] 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% accuracy, the outcomes of cases before the European Court of Human Rights when fed the facts of the cases alone (Aletras et al. 2016). Most impressively, a similar system had better luck in predicting the rulings of the US Supreme Court than a group of 83 legal experts, of whom almost half had previously served as the justices' law clerks (60% vs. 75% accuracy) (Brynjolfsson and McAfee 2017). If the disparity between the performance of such systems and that of well-trained and experienced human professionals widens any further, presumably it will not much matter if humans perfunctorily adhere to whatever these systems decide or advise in a particular situation.

[1] Notice, incidentally, that it is therefore not just when a decision tool is architecturally opaque that a human operator should potentially be retained in the decision control loop, as is already widely appreciated (e.g. IEEE 2017; House of Lords 2018). Even a fully technically transparent decision system may enjoin human agency to a greater or lesser extent.

[2] The AI system in this case computed measures of structural brain connectivity from fMRI brain scans, and used these as inputs to a classifier. Accuracy was computed on unseen brain scans; the classifier's performance was significant at the $p < 0.001$ level. More generally, assessing the accuracy of machine learning systems can be a complex task, but suffice it to say that more meaningful accuracy rates would need to consider base rates, which are not always cited in the relevant studies. It is well known that when a base rate is lower than the false positive rate of a test, false positives will exceed true positives even for an extremely accurate test (the so-called "false positive paradox"). Stepping back a little, however, because our point here is just that in some highly formulaic and process-driven domains it appears that machines perform better than humans, the underlying base rates will not be strictly relevant. So long as machines perform better than humans in these domains (as our examples illustrate), these results should hold regardless of base rates.

2 Background to the Control Problem

A human–machine system (HMS) may be defined as the synthesis of a biological–psychological system and a technological-mechanical system characterized by functional interdependence (Johannsen 1982). The object of any HMS is to provide a "function, product or service as an output with reasonable costs, even under conditions of disturbances influencing man, machine or both" (Johannsen 1982, p. xiii). Importantly, it is has long been recognized as ideal for the human element in this system to be satisfactorily absorbed in the role being played—to reach an adequate level of job satisfaction (Moray 1979)—even if this conflicts with the overall aims of the HMS (e.g. in providing a service at reasonable cost).

HMSs were first investigated in relation to predominantly manual control tasks, initially in aircraft piloting, but then later in ship steering, car driving and industrial process control (see Kelley 1968; Edwards and Lees 1974; Sheridan and Ferrell 1974 for early reviews). This research continues today under the branch of psychology known as "human factors". Human factors research draws on various strands of inquiry, including sociology, physiology, control theory, systems engineering and cognitive science, the latter a branch of (cognitive) psychology that investigates mental processes through the use of models inspired by computer science (Newell and Simon 1972; Rouse 1982). Apart from providing models of cognitive processes, computers have been a common denominator in practically all HMSs from the time they were first studied, with human–computer interaction (HCI) a key focus from the start (Pazouki et al. 2018). Here the standard topics have concerned optimal task allocation, interface design and software ergonomics generally (Rouse 1981; Hatvany and Guedj 1982; Williges and Williges 1982). Computer-aided decision-making, which is our concern here, can therefore be considered a special branch of HCI and human factors research.

When we talk about "control" of HMSs, we are using the term in a broad sense—broader than the sense typically understood in control theory and human factors—encompassing tasks which have traditionally been regarded, strictly speaking, as distinct from control, such as problem-solving (see e.g. Johannsen 1982). Thus by "control" we understand both *fault diagnosis and management* (solving problems as they occur in real time, with a view to restoring normal operation) as well as *planning* (anticipating future problems and devising appropriate strategies to combat them). The control problem was arguably first identified in papers by Wickens and Kessel (1979) and Wiener and Curry (1980), but it did not receive its definitive and celebrated formulation until Lisanne Bainbridge (1983) paper came along with the succinctly telling title: "Ironies of Automation". The chief irony with which her paper grappled is "that the more advanced a control system is, so the more crucial may be the contribution of the human operator" (1983, p. 775). 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 monitor in a HCI is as salient today as when the paper first appeared:

> if the decisions can be fully specified then a computer can make them more quickly, taking into account more dimensions and using more accurately speci-

> fied 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"*. (1983, p. 776, emphasis added)

As we see things, this residual monitoring function of the human operator generates at least four kinds of difficulties that should be treated separately. The first relates to the cognitive limits of human processing power (the "capacity problem"). Its statement in Bainbridge followed directly on from the italicized portion of the preceding quote:

> if the computer is being used to make the decisions because human judgement 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. (1983, p. 776)

Humans are often at a severe epistemic 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 is impossible for a monitor to keep abreast of what is happening in real time because the trades occur at speeds that simply exceed the abilities of human monitors to keep track. As Baxter et al. (2012, p. 68) 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". Analogous problems arise in aviation with respect to the use of autopilot systems (Baxter et al. 2012). Cebon (2015, p. 10) notes that autopilot systems are becoming "…so sophisticated that they only fail in complex 'edge cases' that are impossible for the designers to foresee. Consequently pilots cannot be trained to handle them". Along with the attentional problem (see next paragraph), cognitive constraints account for the main difficulties operators experience in reasserting control of a HMS when the system malfunctions. While one may question whether the opacity of a system impedes its intelligibility quite as much as the capacity problem seems to imply (see e.g. Zerilli et al. 2018), the capacity problem presents a formidable HCI challenge regardless. Today, even a fully proficient software technician would be loath to understand the multi-vector logic governing the generation of a neural network's outputs.

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 towards 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…. (Bainbridge 1983, p. 776)

Automation has a significant impact on situation awareness (Stanton 2016). This is perhaps most clearly illustrated in respect of autonomous vehicles. Inattentive drivers operating a vehicle while it is in autonomous mode are less able to anticipate takeover requests and may be ill-prepared to resume control in an emergency

(Stanton 2015; Cunningham and Regan 2018; Banks et al. 2018a, b). Instantaneously transitioning from low to high workload poses great difficulties for most people (Walker et al. 2015).

The third difficulty relates to the *attitudes* of human operators in the face of sophisticated technology (the "attitudinal problem"). Except for a few brief remarks (Bainbridge 1983 p. 776), this problem was not really addressed in Bainbridge's paper (cf. Wiener and Curry 1980). It has, however, been the subject of active research in the years since (see e.g. Skitka et al. 2000; Parasuraman and Manzey 2010; Pazouki et al. 2018). Here the conundrum 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…[T]he operator assumes that the system is reliable and therefore failure detection deteriorates" (Pazouki et al. 2018, p. 299). There is some evidence that complacency is worse under conditions of multiple task load, "when manual tasks compete with the automated task for the operator's attention" (Parasuraman and Manzey 2010, p. 387). In other words, "[t]he operator's attention allocation strategy appears to favor his or her manual tasks as opposed to the automated task" (Parasuraman and Manzey 2010, pp. 387–388). While this makes it sound as if complacency is an attentional issue, in truth it is an attitudinal one because (so it seems) the monitor only risks being "distracted" by other tasks when they believe the system is reliable enough to be left alone. When the system is not regarded as reliable in the first place, the effect does not occur. (As we discuss later, the control problem arises only from the use of highly reliable but imperfect systems: it does not arise from the use of less reliable systems). This very likely explains why the effect is reversed when the operator monitors a less reliable system, which predictably elicits higher vigilance (Parasuraman and Manzey 2010; Banks et al. 2018a, b). Related to automation *complacency* is automation *bias*, occurring when human operators "trust the automated system so much that they ignore other sources of information, including their own senses" (Pazouki et al. 2018, p. 299). Both complacency and bias "describe a conscious or unconscious response of the human operator induced by overtrust in the proper function of an automated system" (Parasuraman and Manzey 2010, p. 406).

The fourth and final difficulty relates to the *currency* of human skills (the "currency problem"):

> 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….[With regard to cognitive skills] efficient retrieval of [process] knowledge from long-term memory depends on frequency of use….[T]his type of knowledge develops only through use and feedback about its effectiveness. People given this knowledge in theoretical classroom instruction without appropriate practical exercises will probably not understand much of it, as it will not be within a framework which makes it meaningful, and they will not remember much of it as it will not be associated with retrieval strategies which are integrated with the rest of the task. (Bainbridge 1983, pp. 775–776)

These four problems may be seen as four distinct but interrelated aspects of the one control problem. For example, the capacity problem can be expected to intensify the human tendency towards overestimating the value of a machine's outputs, thus compounding the attitudinal problem. In this paper we shall follow the trend of most human factors research since 1983 by confining our attention primarily to the attitudinal problem. Although the problems are distinct, their very interrelatedness means that prescriptions regarding one may go some way towards alleviating (some of?) the others (e.g. see our discussion of the value of *dynamism* in HMSs, in Sect. 5).

3 Locating the Control Problem Within the Landscape of Control-Related Issues

The control problem is nestled among a set of (at least) six interrelated questions about HMSs. Identifying the main questions together enables us to situate the control problem within a broader framework of inquiry into human control in HMSs. In this section we say something brief about the first four questions. In the following two sections, we pursue answers to the final two questions at comparatively greater depth. The questions we have identified are as follows:

(i) What is meaningful human control of a HMS?
(ii) Is human control, so understood, always necessary within a HMS?
(iii) Can the role of the human operator be safely reduced to that of monitor alone?
(iv) If not, why not?—What is "the control problem"?
(v) When, or under what conditions, does the control problem arise?
(vi) Can the control problem be solved?

Regarding (i), a system for us is under "meaningful human control" when, *at a minimum*, it behaves the way it should, i.e. in accordance with the wishes of its operators (cf. Santoni de Sio and van den Hoven 2018). For a system to be under meaningful human control, however, also implies that it is under *effective* control, such that its operators have the wherewithal to correct the system or abort its operations in sufficient time to avert the worst effects of its deviance. This is why we stated earlier that our notion of control extends to fault diagnosis and management (resolving problems with a view to restoring normal operation) as well as planning (devising strategies to deal with contingencies). Regarding (ii), we assume that it is always desirable for a HMS to be under effective human control.

Our discussion of the control problem already signals a negative reply to question (iii) whenever high-stakes or safety-critical decisions are involved: humans perform very poorly at prolonged monitoring tasks (Molloy and Parasuraman 1996; Banks et al. 2018a). Human attentional resources typically "shrink to fit" task demands (Walker et al. 2015). The attitudinal and attentional effects of the control problem combined are sufficiently detrimental to explain why a HMS that reduces the human controller to a monitor of displays or passive recipient of outputs is necessarily hazardous. However, because the effects of complacency can be reversed by replacing a reliable system with a less reliable system, some purely monitor-based setups do

seem to avoid the problem (note: they *avoid*, rather than *solve*, the problem). And of course systems that approach better-than-human reliability do not pose a control problem as such (e.g. automatic transmission, case prediction software, etc.).

Because we have already defined the control problem (iv), all that remains is for us to elaborate on the answers we have intimated to the final two, most important, questions, i.e. (v) and (vi). We address these in the following two sections.

4 When, or Under What Conditions, does the Control Problem Arise?

The conundrum of control is that the more reliable a system becomes, the more difficult it is for a human supervisor to maintain an adequate level of engagement with the technology to ensure safe resumption of manual control should the system malfunction. In relation to current "Level 2" autonomous vehicles (see below)—which allow the driver to be hands- and feet-free but not *mind-free*, because the driver still has to watch the road—Stanton (2015, p. 9) puts the point vividly: "even the most observant human driver's attention will begin to wane; it will be akin to watching paint dry". This is a manifestation in manual systems of a more general problem of control over automated decision support systems, viz., the tendency to defer to systems that approach (but do not reach) very high reliability and predictability. Conversely, *decreases* in automation reliability generally seem to *increase* the detection rate of system failures (Bagheri and Jamieson 2004). Starkly put, automation is "most dangerous when it behaves in a consistent and reliable manner for most of the time" (Banks et al. 2018b, p. 283). Decades' worth of research in aviation, shipping, driving and industrial process control supports this assessment. The only options to deal with the predicament therefore appear to be (at least in high-stakes or safety-critical contexts):

(a) The use of less reliable systems for tasks whose execution may be more expedient when automated to any standard of proficiency than when not at all, since less reliable systems do not pose the control problem;
(b) To implement only *partial* automation through *task decomposition* (see Sect. 5); and
(c) To wait for a system to reach near perfect (better-than-human) reliability before deployment.

Options (a) and (c) are self-explanatory. We shall discuss (b) in greater detail in the following section (as well as a few other strategies, such as "catch trials", in Sect. 7). For now we note that (b) and (c) are not mutually exclusive: any decision task that has automatizable subcomponents may employ near-risk-free technology working alongside an active, purposefully-engaged human operator, with human and machine fully autonomous within their particular spheres of operation. Furthermore, options (a) and (b) are not mutually exclusive either: any decision task that has automatizable subcomponents may likewise employ patently *suboptimal* technology working alongside an active, purposefully-engaged human operator (although here

the machine will obviously *not* be fully autonomous within its sphere of operation). Because it may seem counterintuitive, we shall justify the inclusion of option (a) (and (a + b)) as part of our menu of options in the following section.

Since decision tasks can be cut more or less finely [as option (b), above, assumes], one might wonder whether the control problem presents quite the same challenge when the situation involves the automation of a few subcomponents of a decision, in contrast to the automation of an entire decision. We could assume, for instance, that *border control* is one big decision—i.e. whether to admit, or not admit, persons moving between state boundaries—involving customs clearance, passport verification, drug detection, and so on. When the *entire* process is automated by one large, distributed border control software package, and this system works reliably well most of the time—but still requires human monitors to invigilate display panels and the like—is the control problem any worse than it would be when some automatizable subcomponents of the overall decision are carved out for discrete automation (and automated, once again, by *mostly*, but not completely, reliable systems)?

Currently, of course, border control decisions are only partially automated in this sense: SmartGate allows for fully automated electronic passport control checks, but customs officials and sniffer dogs still litter most immigration checkpoints, and their job is to handle such parts of the overall decision task as cannot be effectively automated. The same issue arises in automotive engineering. 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, while Google's self-driving car does everything except turn itself on and off. The thought is that if an actively engaged human is retained somewhere in the control loop, contributing purposefully to the decision task, the human may be less susceptible to automation complacency and bias than when their only role is monitoring.

But actually the research we cited earlier suggests that matters are not quite so simple. Complacency appears to be *worse* under conditions of multiple task load so long as the automated subcomponent, being reliable most of the time, engenders misplaced trust in its ability to be left alone. Crucially, this effect can be reversed if the system is replaced with one that does not engender the same degree of confidence—i.e. a less reliable system (Parasuraman and Manzey 2010, pp. 387–388). These findings indicate that the control problem is not necessarily affected by the size or share of the automated subcomponent relative to the whole decision procedure. As long as the autonomous system in question is more reliable than not, the control problem rears its head, with the only options available for remediating its effects being the three outlined above.

On the other hand, this is not to deny that differences between partial and full automation may extend to differences in how human operators typically *perceive* the respective capabilities of these systems. There is evidence that operator trust is positively 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" (Banks et al. 2018b, p. 283). As the level of automation increases, however, this allocation gets blurred to the point that drivers find it difficult forming accurate assessments of the vehicle's capabilities, and on the whole are inclined to overestimate them (Banks et al. 2018b). Counteracting this effect may be the greater readiness to believe that a smaller, less sophisticated device—with fewer working parts and opportunities for system glitches—will be compensatingly less temperamental. In fact we suspect that this presumption is probably justified in the case of decision subsystems in SAE Level 0 vehicles such as automatic transmission, automatic light control and first generation cruise control (Walker et al. 2015). These subsystems may be so reliable, approaching near perfect (better-than-human) dependability, as to effectively neutralize the control problem's sting in most cases. So in the end, it seems, larger and more sophisticated technologies that are mostly reliable *probably do* pose a greater control problem than smaller ones.

5 Can the Control Problem be Solved?

If the question is interpreted literally, the answer to "Can the control problem be solved?" appears to be straightforwardly negative: the control problem cannot *literally* be solved. There is nothing we can do which *directly* targets, still less *directly* alleviates, 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 HMSs, we may 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 is instructive here. There is no reason to suspect that machine learning and other state-of-the-art decision support systems are less likely to induce complacency effects than other forms of automated effort (e.g. Cummings 2004; Edwards and Veale 2017, p. 51). With this in mind, one important human factors recommendation is to foster mutual accommodation between human and computer competencies through a *dynamic* and *complementary* allocation of functions that optimally preserves attentional resources (Stanton and Marsden 1996). Ideally, only those parts of a decision should be automated that leave the human operator with something vital and absorbing to do (Bainbridge 1983). Optimal workload, moreover, would prevent demand explosion in scenarios where the operator must intervene quickly to rectify a situation—effective intervention can be enormously difficult when the operator has to shift from low to high level cognitive effort within a very short window (Walker et al. 2015).

Let us call this the "dynamic/complementary allocation of function" (DCAF) approach. DCAF assumes that human performance can be enhanced when automation *augments* rather than *replaces* human skills. It need not, however, assume that augmentation is always preferable to replacement. In fact, for the DCAF approach to work, some systems clearly need to replace the human agent and be left to operate autonomously. Human–machine decision systems that contain automated subcomponents work best when the human operator is allowed to concentrate their

energies on the chunks of the task better suited to human rather than autonomous execution—a setup which 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 for the most part but still require a human monitor—and it is clear where *this* path leads. Indeed one of the advantages of complementarity is precisely that, by carving up a big decision into smaller and smaller chunks, the more likely it will be that better-than-human systems can be found to handle them.

Complementarity between human and computer is crucial. In a passage worth citing in full, Pohl (2008, p. 73) notes that

> …intelligent software systems can be particularly helpful in complementing human capabilities by providing a tireless, fast and emotionless problem analysis and solution evaluation capability. Large volumes of information and multifaceted decision contexts tend to easily overwhelm human decision-makers. When such an overload occurs we tend to switch from an analysis mode to an intuitive mode in which we have to rely almost entirely on our ability to develop situation awareness through abstraction and conceptualization. While this is perhaps our greatest strength it is also potentially our greatest weakness, because at this intuitive meta-level we become increasingly vulnerable to emotional influences.
>
> The capabilities of the computer are strongest in the areas of parallelism, speed and accuracy. Whereas the human being tends to limit the amount of detailed knowledge by continuously abstracting information to a higher level of understanding, the computer excels in its almost unlimited capacity for storing data. While the human being is prone to impatience, loss of concentration and panic under overwhelming or threatening circumstances, the computer is totally oblivious to such emotional influences. *The most effective implementation of these complementing human and machine capabilities is in a tightly coupled partnership environment that encourages and supports seamless interaction*. (emphasis added)

At the same time, DCAF emphasises that the allocation of functions in a HMS should be flexible enough to support *dynamic* interaction, with hand-over and hand-back for shared competencies (as occurs when a driver disengages cruise control and thereby resumes control of acceleration). Of course dynamism will be dangerous if there is hand-over between agents that are ill-matched in their competencies. Dynamism will only work in circumstances where the human and machine are nearly equivalently proficient (with the machine perhaps only marginally better than the human). Apart from anything else, dynamic interaction allows the human to maintain their manual control skills, and so may go some way towards alleviating the currency problem.

It should be clear that DCAF falls under option (b) in our three-item menu of options for managing the control problem. It should also be clear that, as we presaged in Sect. 4 (and explained in the preceding paragraph), DCAF (b) will almost always employ near-risk-free technology (c)—so that (b) and (c) are not mutually

exclusive. For example, in relation to autonomous vehicles, Stanton (2015, p. 9) describes the DCAF challenge as follows: "We need to design vehicle automation to have graduated and gradual hand-over and hand-back tasks if it is to successfully support human drivers. Vehicle automation needs to work towards providing a chatty co-pilot, not a silent auto-pilot!" For this to be the case, however, the driver cannot worry about whether what the "chatty co-pilot" is saying is true. In the DCAF approach, the human *is* a co-pilot of sorts, with real work to do, so that the fidelity of their autonomous counterpart/s needs to be assured.

What if this fidelity cannot be assured? We have said that *as a rule* a decision tool should not replace a human agent in a high-stakes/safety-critical setting unless the tool reaches a certain crucial threshold of reliability; and we have applied this recommendation iteratively, carrying it over to DFAC decision contexts in high-stakes/safety-critical settings such that, for any automatizable *subcomponent* of a decision procedure, a tool should not replace the human agent responsible for that (sub)decision unless the tool meets our very high (better-than-human) standard. But what if this standard cannot be met? Can less-than-reliable systems be deployed here? In other words, are there exceptions to the general rule we have defended, or special circumstances where the general rule gives way? The short answer is: yes. As we have explained, the control problem does not arise from the use of patently suboptimal automation, only from *generally* dependable automation. Therefore, depending on the exact nature of the HCI at issue, a *less*-than-reliable system might safely replace the human agent charged with deciding some matter within a larger decision structure, for example, passport verification within the larger border control decision structure. The use of less-than-reliable systems is of course option (a) in our menu, and as exhibited in this example, nested within a broader decision structure consisting of subcomponents, as predicated by (b) (i.e. (a) and (b) are not mutually exclusive, again as we noted in Sect. 4). But now, why *would* a patently suboptimal decision tool be deployed here at all (or anywhere else, for that matter)? It is one thing to say that it avoids the control problem. It is quite another to say that a system which is so suboptimal it does not engender human confidence should *for that very reason* be used to assist human decision-makers. Clearly we owe our readers an explanation for including option (a) within our menu of strategies for dealing with the control problem.

We think deployment of suboptimal tools may still prove useful in circumstances where the tools have access to information to which the human does not, or otherwise "decide" things in ways that humans generally *cannot*. Such systems very literally augment human capacities: human and machine in effect *share* control. The clearest examples of this form of technology are recidivism risk prediction tools used in law enforcement. Not all such instruments come with the problematic biases of PredPol (used in so-called hot-spot policing) and COMPAS (predicting the likelihood of offender recidivism) built into them. Some may be genuinely useful in reducing crime at the same time as *reducing* the prison population (see e.g. Kleinberg et al. 2018). These systems answer questions of the form: How should we distribute police officers over a locality having such and such geographical characteristics? What is the likelihood that this prisoner will reoffend if released on parole? And so on. Consider how these systems decide such matters. Often they use logistic

regression or more advanced actuarial techniques to mine patterns from very large databases. This is not a feat unaided humans can hope to match. There are also some phenomena within human decision-making that algorithms can help to counteract—e.g. decision fatigue and decision inertia, of which some of the classic studies actually involve judges' parole decisions (e.g. Danziger et al. 2011).

Be that as it may, however, we would still urge that great caution be exercised before any form of suboptimal automation is used in high-stakes/safety-critical settings. Many of these systems (like COMPAS) are after all tools which have attained notoriety for their problematic biases and inherent technical limitations (e.g. Blomberg et al. 2010; Larson et al. 2016; Dressel and Farid 2018). And as some of these systems gradually begin to overcome their limitations, our worry is that the control problem will gradually re-emerge, taking human operators unawares and decision subjects along with them. It will be all too easy for a judge with decision fatigue, for example, to simply rely on what a predictive risk instrument "objectively" recommends. It may turn out that guidelines recommending, for example, that decision-makers consult their own judgment first before consulting an algorithm, using the algorithm merely as a check on their intuitions, could assist in offsetting some of the effects of automation complacency and bias. (Note that this would come close to telling decision-makers *not* to use the algorithm—hardly a "solution" to the problem). The Wisconsin Supreme Court, when discussing protocols around judicial use of the COMPAS recidivism algorithm required that sentencing judges be given a list of warnings about the tool as a condition of relying on its predictions. More empirical research is required to see whether such approaches really do work.

To conclude our analysis of option (a), some of these features of suboptimal decision systems should be stated more explicitly. Option (a)—whether or not combined with option (b)—always represents a specific type of HCI. It is noteworthy that whenever a patently suboptimal tool is used, the human agent does not readily stumble into the control problem, and may therefore be assured of an active and meaningful role working alongside it (assuming a certain level of diligence, aptitude and motivation on the part of the human). But the interaction here will not quite be the same as that envisaged for HMSs under the DCAF approach. Under DCAF, the allocation of functions is complementary. This will not generally be the case under option (a). Assume that a decision, D, comprises the (sub)decisions d_0, d_1, d_2 and d_3. Under DCAF, the human will take care of (let us say) d_0, d_1 and d_2 while the system will take care of d_3. This means the human can concentrate their energies *entirely* on d_0, d_1 and d_2 and essentially *ignore* d_3, because the system is superior in making decisions of the type d_3. This cannot happen when suboptimal software is used. In such cases, it is not as if the human can take care of d_0, d_1 and d_2 while the system takes care of d_3; the human must *also* take care of d_3, albeit with the *assistance* of a suboptimal decision tool. Both human *and* machine participate in d_3 (i.e. human and machine effectively *share* control). Indeed, in light of what we said earlier, in many cases both human and machine may aptly be described as doing (or deciding) the same thing in different ways: for the system may be trying to answer the same question as the human (will she reoffend if released on parole? etc.), only with access to information which the human does not have, or in ways (or at speeds) which humans are unable to match. So there need be no complementarity of *functions* here,

as both human and machine may be performing the *same* function (viz., d_3), albeit distinctly—a classic case of "multiple realization" (Zerilli 2017). We might even say that instead of a complementary coupling of skills between human and computer, what we have under option (a) will often be a *supplementary* coupling—where each agent adds something unique to the decision problem in point of how the agent goes about resolving it (somewhat analogous to the role of an expert witness who "assists" the judge in determining the appropriate sentence for an offender).

Table 1 summarizes the logical space of decisions that may be automated, in whole or in part, and our recommendations for whether or not they should be. Figure 1 depicts both the presence and danger of complacency as a function of system reliability. Notice that at a certain point of reliability, the presence of complacency no longer matters. Notice also what the figure does *not* capture: it does not tell us what over-reliance on an algorithm looks like, or otherwise calibrate for a healthy or sceptical level of dependence on an algorithm.[3] This—which could well be called the *measurement* problem—is the subject of separate research in human factors, although it is probably fair to say that most design interventions and optimal task allocation and HCI paradigms are about preventing over-reliance rather than detecting it, no doubt (at least partly) because the hallmarks of over-reliance will differ from system to system and context to context (see e.g. Endsley 2017).

5.1 System Reliability

The DCAF approach might be thought of as a variant of the "Privacy by Design" (PbD) approach—something like "Control by Design". PbD proponents want data protection principles taken into account throughout the whole lifecycle of information systems development (Bygrave 2017). Analogously, one could say that the DCAF approach seeks to "hardwire" human factors considerations into the development of HMSs from the earliest stages of modeling. For algorithmic decision support tools, we would suggest that the following six principles can serve as a framework both for assessing the viability of any such human–machine system as well as guiding their design and implementation:

- *Division of labour* Decisions with automatizable subcomponents should reflect a clear allocation of responsibilities between the human- and computer-operated parts of the decision.
- *Complementarity* The allocation of responsibilities should proceed in such a way that those subcomponents better suited for human handling are not automated, and those better suited for computer handling are not manually controlled. While an unhelpful aversion to algorithms can be reduced by giving users power to adjust a decision system's outputs (Dietvorst et al. 2016), human interference also tends to introduce errors (Fildes et al. 2009). Humans should stick to what they do best, such as communication, symbolic reasoning, conceptualiza-

[3] We owe this point to an anonymous reviewer of the journal.

Table 1 The logical space of automatizable decisions and their attendant control problems

For any decision D: { d_0, d_1, d_2 ... d_n}		
Option (a) Patently suboptimal system D (whole decisions)	Option (b) Decomposition/allocation d_0, d_1, d_2 ... d_n (semiautonomous decisions)	Option (c) Better-than-human system D (whole decisions)
Supplementary coupling/ human in the loop e.g. medical diagnostic tools **Proceed with caution**	Complementary coupling/ human in the loop (b + c) (DCAF) e.g. ACC (below) **Recommended** OR Supplementary coupling/ human in the loop (a + b) e.g. recidivism risk prediction instruments **Proceed with caution** OR *Mostly* reliable system/ human in the loop e.g. SAE Level 2 AV **Danger! Control problem!**	No human in the loop e.g. SAE Level 5 AV (eventually); case prediction sofware **Recommended**

tion, empathy and intuition—provided these are the skills actually required for a particular decision (Pohl 2008). Intuition may be excellent in some decision contexts (who will be the best actor for this part?) but the last resort of a scoundrel in other contexts (what is the likelihood that this brown-skinned man with a history of mental health issues and no priors will re-offend?). A related point is that what may count as a virtue in a human may well count as a vice in a machine. "Graceful degradation" has long been regarded as one of the advantages of humans over machines (Fitts 1951), but as Bainbridge (1983, p. 777) pointed out, "[t]his is not an aspect of human performance to be aimed for in computers…automatic systems should fail obviously."[4] In short, complementarity means humans and machines have clearly defined and clearly separated roles, where the human is effectively barred from interfering with the machine's

[4] This is not always true. An object classifier should tell you how much a given object is like a chair, rather than move discretely from "chair" to some other category of objects: to that extent, classifiers certainly *are* designed with a view to graceful degradation.

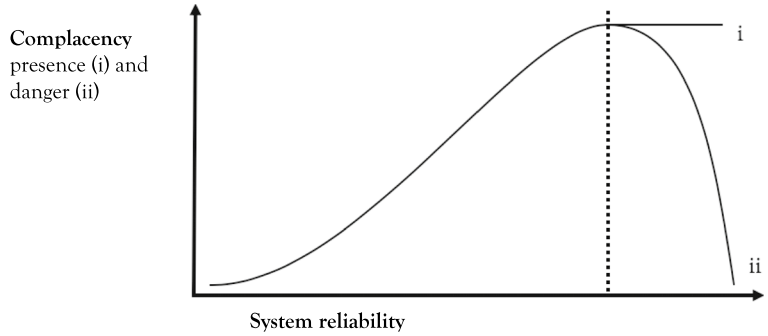

Fig. 1 Presence (i) and (ii) danger of complacency as a function of system reliability. The dashed line represents near perfect (better-than-human) reliability

outputs and perhaps in many cases even knowing what the outputs are. Breaking the task up in this way also increases the chances of finding optimally reliable software to handle the automated parts of a decision.

- *Dynamism* The allocation of responsibilities should incorporate hand-over and hand-back protocols where this flexibility contributes to optimal performance. This assumes that some decisions can be safely handled by both humans and computers, i.e. where humans and computers have shared competencies within particular subdomains. Hand-over and hand-back may also go some way towards alleviating the currency problem, as operators are thereby afforded an opportunity to practice and maintain their manual control skills (Bainbridge 1983).
- *Co-evolution* User requirements co-evolve over time, and decision support tools should reflect this. Decision support tools within HMSs should be designed for adaptability and change. This means designers should not over-specify how a system will work, but allow its users to tailor the system so that it best meets their particular needs (Walker et al. 2015, p. 201).
- *Pragmatism* Decision support tools "should be congruent with existing practices which may on occasion appear archaic compared to what technology now offers" (Walker et al. 2015, p. 201). When cell phones first appeared, people did not throw out their telephone directories and address books in short order. The older technology held on for a while longer until mobiles were subsumed into the Internet of Things. As new decision software gets tested and then rolled out, we think the same approach is advisable.
- *Context-sensitivity* Each decision tool, situated within its own unique decision context, may prioritise these principles and negotiate their various trade-offs differently.

We now turn to consider an algorithmic decision tool that exemplifies several of these principles in its design.

6 Case Study: New Zealand's Accident Compensation Corporation (ACC)

Since 1974, New Zealand has run a universal no-fault accident compensation scheme for personal injury. Because the scheme is compulsory, all citizens/residents (or temporary visitors) who have suffered personal injury can expect to be covered regardless of whether or not another party is at fault. Most of the ACC's payments cover treatment/medical costs, but they also regularly cover lost earnings and home and/or vehicle modifications for those with more serious injuries. The ACC processes around 2 million claims per year, of which (on average) about 96% are accepted (ACC 2018).

Over the years, the ACC has largely relied on manual control for processing claims. In the past this has involved ACC staff members sorting through and assessing individual claims one by one. Even with improvements to case handling procedures over the years, such as technology allowing electronic submission, *all* claims have required some degree of manual processing (ACC 2018). At the time of writing, the ACC plans to introduce an improved claim registration and cover assessment process by the end of 2018. It aims to make the claims approval process quicker and more efficient, removing the need for manual control in standard cases altogether. The ACC hopes that by harnessing the power of big data—12 million claims submitted between 2010 and 2016—it can both reduce the wait time for approvals as well as more efficiently distribute the more complex claims to ACC teams for final determination.

There are two key features of the new claims handling process. First, the ACC's machine-learning algorithm (developed in-house) has been designed to identify such characteristics of a claim as are strictly relevant to whether it can be accepted. Thus, "[s]imple claims—where the information provided shows that an injury was caused by an accident—will be fast-tracked and immediately accepted" (ACC 2018). For example, the system would fast-track "someone going to the emergency clinic to have a cut stitched," but not someone presenting with "multiple severe injuries" (ACC 2018). Claims that are not accepted at this step will be passed along to ACC teams for manual processing. Second, the system is not able to *decline* claims—it can only *accept* claims that show up as straightforwardly involving accidents based on information provided by the claimant. In giving the tool this limited jurisdiction, its designers have ingeniously converted what is potentially a high-stakes decision into an extremely low one: for the tool cannot decide adversely, only inclusively.

The process runs as follows. Each claim moves through a series of automated system checks. At each "checkpoint," one of two things can happen: the claim passes, or it does not. If it passes, it proceeds to the next checkpoint. If it does not pass, the system flags it for manual processing. There are three checkpoints: one for *validation and eligibility*; a second for *accident description*; and a third for *cover decision* (final determination). At the validation and eligibility checkpoint, the system checks that the claimant has provided all essential information (e.g. location and date of accident, type of injury, healthcare provider, claimant's

employment and residency status, etc.). If the information is complete, the system passes the claim on to the accident description checkpoint. If the information is incomplete, the system flags it for manual processing. At the accident description checkpoint, the system attempts to categorize the claim in accordance with a predetermined taxonomy of claim types ("fall," "rugby accident," etc.). The system searches the claim form for words that correlate with recognized claim types. If the claim can be categorized in this way, it is passed on to the final checkpoint for determination. If it cannot be categorized, it is flagged for manual processing. At the final checkpoint, two statistical models are employed:

> The "Probability of Accept" model is informed by a statistical model that uses data from 12 million previous, anonymised claims to calculate the probability a new claim should be approved. Each claim is then given a score, and ACC sets a threshold for scores that will be automatically accepted or not. A claim that scores above the threshold set by ACC will be automatically accepted for cover.
>
> A claim that scores below the auto-acceptance threshold would then be run through the "Complexity" model, which categorises the claim on a scale of low-complexity through to high-complexity. Each claim is then given a complexity score, and ACC sets a threshold for complexity scores that will be automatically accepted or not. A claim that scores below the threshold set by ACC will be automatically accepted.
>
> A claim that scores above the auto-acceptance threshold for complexity would then be referred for further manual processing by an ACC staff member.
>
> *Example: If a client submits a claim for treatment relating to multiple severe injuries and post- traumatic stress following a motor vehicle accident, their claim is likely to receive a high complexity score and would be referred for handling by ACC teams.* (ACC 2018)

How well does the ACC tool incorporate the six human factors principles we outlined in the previous section? We think it performs commendably on at least three of our stated principles:

- *Division of labour* There is here a clear allocation of responsibilities between the human- and computer-operated parts of the decision regarding whether to approve claims.
- *Complementarity* The allocation of responsibilities proceeds in such a way that those subcomponents of the decision better suited for human handling are not automated, and those better suited for computer handling are not manually controlled. Human controllers do not interfere with the automated parts of the process, which are therefore left to operate as essentially risk-free zones of automated decision-making. Humans are not able to perform such aspects of the "accident description" and "cover decision" subcomponents of the decision as are handled by the algorithm either as *accurately* or as *quickly* as the algorithm.

In the ACC model, humans only intervene at the point where the algorithm reaches the limit of its competence to determine a claim.

- *Context-sensitivity* The allocation of responsibilities does not incorporate hand-over and hand-back protocols (i.e. it is not *dynamic*), but in this particular setup, such flexibility would not contribute to optimal performance. This means that the ACC tool has been deployed in a context-sensitive manner, *dispensing with* or *trading off against* principles that are not especially salient in the specific circumstances of deployment.

7 Are There Other Ways to Address the Control Problem?

Earlier we mentioned that the control problem appears to be resistant to training, and that experts are as prone to complacency as novices. Before concluding, we should say a little about some of the other commonly suggested strategies for overcoming the control problem.

First, there is some evidence that increasing accountability mechanisms can have a positive effect on human operators whose primary responsibility is monitoring an autonomous system. In an important study, Skitka et al. (2000, p. 701) found that "making participants accountable for either their overall performance or their decision accuracy led to lower rates of automation bias". 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 *complacency* (as opposed to bias). Parasuraman and Manzey (2010, p. 396) also warn that the results of this study are not conclusive. But clearly auditing protocols, and perhaps more creative accountability measures, such as "catch-trials"—in which system errors are deliberately generated to keep human invigilators on their toes—could be quite useful in counteracting automation bias. Catch trials look particularly promising. While more research on their efficacy is needed before concrete guidelines can be promulgated recommending their adoption, they do seem to offer a viable means of getting humans more actively engaged in the control task.[5] But in any case, much like other touted solutions to the control problem, they do not offer a *literal* 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 do not induce the same complacency and bias that attend more reliable systems. Thus catch trials really fall under option (a) in our menu of strategies above.

Second, might having a *group* of humans in the loop, working together and able to keep watch on one another, alleviate automation bias? Apparently not:

[5] Catch trials might not be such a good thing if case workers came to think that "odd" looking cases were "probably not real".

> 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…. (Parasuraman and Manzey 2010, p. 392)

Finally we should note that Parasuraman and Manzey (2010, p. 387) added these observations in connection with the effects of training regimes:

> Although extended practice does not eliminate automation complacency, other training procedures may provide some benefit. In particular, given that complacency is primarily found in multitasking environments and represents attention allocation away from the automated task, training in attention strategies might mitigate complacency.

We are not aware of research empirically substantiating this latter claim, but would anyway caution against task allocations that reduce human agents to monitors of largely static displays of information (if that is the implicit proposal here). Other studies cited by the authors, albeit in the context of automation bias, indicate that even explicit briefings about risk factors in HCI do not mitigate the strength of automation bias, and this seems to be true in both multitasking and singletasking environments. An idle mind is not an empty mind, but rather a wandering mind, and when the stakes are high, the risk of complacency is still too great to be managed by rubber-stamp training interventions that make only a questionable difference to deep-seated psychological propensities. Perhaps there is then reason to be sceptical about the effectiveness of "warnings" or other guidelines about how best to use decision tools in judicial settings (as we mentioned earlier in regards to COMPAS). But it is a live research question.

8 Conclusion

Automation introduces more than just automated parts: it very often also transforms 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. To date, little has been said about whether and to what extent the same problem arises from the use of ever more sophisticated algorithmic decision tools, including those exploiting cutting-edge machine learning techniques such as deep learning. We have endeavoured to show how insights from human factors research have great relevance to policy specialists working in AI regulation and policy. Among the factors which 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 effective control to an algorithm just because it works well in most instances. We argue that, except in special cases, whenever an automatizable decision is high-stakes or safety-critical, the decision support tool under consideration should not be deployed

unless it has genuinely earned its keep. Failing that, a dynamic and complementary allocation of functions between actively engaged human operators and simpler but more nearly perfectly reliable autonomous systems should be considered the safest course.

Acknowledgements The authors wish to thank the participants of two roundtables, one held in Oxford, November 23-24, 2017, in partnership with the Uehiro Centre for Practical Ethics, University of Oxford, the other in Dunedin, December 11-12, at the University of Otago.

Funding This research was supported by a New Zealand Law Foundation Grant (2016/ILP/10).

Compliance with Ethical Standards

Conflict of interest AK works for Soul Machines Ltd under contract. JZ, JM and CG have no other disclosures or relevant affiliations apart from the appointments above.

References

Accident Compensation Corporation. (2018). *Improving the claim registration and approval process. Version 1.0.* 4 July 2018.

AI Now. (2018). *Litigating algorithms: Challenging government use of algorithmic decision systems.* New York: AI Now Institute.

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science, 2*(93), 1–19.

Amoroso, N., La Rocca, M., Bruno, S., Maggipinto, T., Monaco, A., Bellotti, R., Tangaro, S., the Alzheimer's Disease Neuroimaging Initiative. (2017). Brain structural connectivity atrophy in Alzheimer's disease. arXiv:1709.02369v1.

Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced "complacency". In D. A. Vicenzi, M. Mouloua, & O. A. Hancock (Eds.), *Human performance, situation awareness, and automation: Current research and trends* (pp. 54–59). Mahwah, NJ: Erlbaum.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779.

Banks, V. A., Erikssona, A., O'Donoghue, J., & Stanton, N. A. (2018a). Is partially automated driving a bad idea? Observations from an on-road study. *Applied Ergonomics, 68,* 138–145.

Banks, V. A., Plant, K. L., & Stanton, N. A. (2018b). Driver error or designer error: Using the perceptual cycle model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Safety Science, 108,* 278–285.

Baxter, G., Rooksby, J., Wang, Y., & Khajeh-Hosseini, A. (2012). The ironies of automation…still going strong at 30? In E. C. C. E. Conf (Ed.), *Proc* (pp. 65–71). Aug.: Edinburgh.

Blomberg, T., Bales, W., Mann, K., Meldrum, R., Nedelec, J. (2010). Validation of the COMPAS risk assessment classification instrument. Center for Criminology and Public Policy Research College

of Criminology and Criminal Justice Florida State University. https://arxiv.org/pdf/1311.2901.pdf.
Brynjolfsson, E., & McAfee, A. (2017). *Machine platform crowd: Harnessing our digital future*. New York: Norton.
Bygrave, L. A. (2017). Hardwiring privacy. In R. Brownsword, E. Scotford, & K. Yeung (Eds.), *The Oxford handbook of law, regulation, and technology* (pp. 754–775). New York: Oxford University Press.
Cebon, D. (2015). *Responses to autonomous vehicles. Ingenia, 62,* 10.
Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. *AIAA Intelligent Systems Technical Conf*. https://doi.org/10.2514/6.2004-6313.
Cunningham, M., Regan, M. (2018). Automated vehicles may encourage a new breed of distracted drivers. The Conversation, Sep. 25.
Damaška, M. R. (1997). *Evidence law adrift*. New Haven: Yale University Press.
Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences, 108*(17), 6889–6892.
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170.
Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4,* 1–5.
Edwards, E., & Lees, F. P. (Eds.). (1974). *The human operator in process control*. London: Taylor and Francis.
Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law and Technology Review, 16*(1), 18–84.
Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors, 59*(1), 5–27.
Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St Martin's Press.
Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25,* 3–23.
Fitts, P. M. (1951). *Human engineering for an effective air navigation and traffic control system*. Washington D.C.: National Research Council.
Greenlee, E. T., DeLucia, P. R., & Newton, D. C. (2018). Driver vigilance in automated vehicles: Hazard detection failures are a matter of time. *Human Factors, 60*(4), 465–476.
Hatvany, J., & Guedj, R. A. (1982). *Man-machine interaction in computer-aided design systems*., Proceedings IFAC/IFIP/IFORS/IEA Conferernce Analysis, design and evaluation of man-machine systems Oxford: Pergamon Press.
House of Lords Select Committee on Artificial Intelligence. (2018). *AI in the UK: Ready, willing and able?* https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.
IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically aligned design (version 2)*. https://ethicsinaction.ieee.org.
Johannsen, G. (1982). Man-machine systems: Introduction and background. Proceedings of IFAC/IFIP/IFORS/IEA Conference on Analysis, design and evaluation of man-machine systems, Baden-Baden, Sept. Oxford: Pergamon Press.
Kelley, C. R. (1968). *Manual and automatic control*. New York: Wiley.
Larson, J., Mattu, S., Kirchner, L., Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica.org* May 23, 2016.
Leinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics, 2018,* 237–293.
Margulies, F., & Zemanek, H. (1982). *Man's role in man-machine systems*. Proceedings IFAC/IFIP/IFORS/IEA Conference Analysis, design and evaluation of man-machine systems. Oxford: Pergamon Press.
Marks, A., Bowling, B., & Keenan, C. (2017). Automated justice? Technology, crime, and social control. In R. Brownsword, E. Scotford, & K. Yeung (Eds.), *The Oxford handbook of law, regulation, and technology* (pp. 705–730). New York: Oxford University Press.
Meister, D. (1999). *The history of human factors and ergonomics*. Mahwah, NJ: Erlbaum.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors, 38,* 311–322.

Moray, N. (Ed.). (1979). *Mental workload: Its theory and measurement*. New York: Plenum Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood-Cliffs, NJ: Prentice Hall.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410.

Pazouki, K., Forbes, N., Norman, R. A., & Woodward, M. D. (2018). Investigation on the impact of human–automation interaction in maritime operations. *Ocean Engineering, 153,* 297–304.

Pohl, J. (2008). Cognitive elements of human decision making. In G. Phillips-Wren, N. Ichalkaranje, & L. C. Jain (Eds.), *Intelligent decision making: An AI-based approach* (pp. 41–76). Berlin: Springer.

Rouse, W. B. (1981). Human–computer interaction in the control of dynamic systems. *ACM Computing Surveys, 13,* 71–99.

Rouse, W. B. (1982). *Models of human problem solving: Detection, diagnosis, and compensation for system failures*., Proceedings of IFAC/IFIP/IFORS/IEA conference Analysis, design and evaluation of man-machine systems Oxford: Pergamon Press.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5,* 15. https://doi.org/10.3389/frobt.2018.00015 .

Sheridan, T. B., & Ferrell, W. R. (1974). *Man-machine systems: Information, control, and decision models of human performance*. Cambridge, MA: MIT Press.

Skitka, L. J., Mosier, K. L., & Burdick, M. (2000). Accountability and automation bias. *International Journal of Human–Computer Studies, 52,* 701–717.

Society of Automotive Engineers. (2016). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. J3016_201609*. Warrendale: SAE International.

Stanton, N. A. (2015). Responses to autonomous vehicles. *Ingenia, 62,* 9.

Stanton, N. A. (2016). Distributed situation awareness. *Theoretical Issues in Ergonomics Science, 17*(1), 1–7.

Stanton, N. A., & Marsden, P. (1996). From fly-by-wire to drive-by-wire: Safety implications of vehicle automation. *Safety Science, 24*(1), 35–49.

Strauch, B. (2018). Ironies of automation: Still unresolved after all these years. *IEEE Transactions on Human–Machine Systems, 48*(5), 419–433.

Villani, C. (2018). *For a meaningful artificial intelligence: Towards a French and European strategy*. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

Walker, G. H., Stanton, N. A., & Salmon, P. M. (2015). *Human factors in automotive engineering and technology*. Surrey: Ashgate.

Wickens, C. D., & Kessel, C. (1979). The effect of participatory mode and task workload on the detection of dynamic system failures. *IEEE Transactions Systems Man Cybernetics, 9*(1), 24–31.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics, 23*(10), 995–1011.

Williges, R. C., & Williges, B. H. (1982). *Human–computer dialogue design considerations*., Proceedings IFAC/IFIP/IFORS/IEA Conference Analysis, design and evaluation of man-machine systems Oxford: Pergamon Press.

Zerilli, J. (2017). Multiple realization and the commensurability of taxonomies. *Synthese, 196*(8), 3337–3353.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology, 32*(4), 661–683.