First Steps Towards an Ethics of Robots and Artificial Intelligence

JOHN TASIOULAS

King's College London

ABSTRACT

This article offers an overview of the main first-order ethical questions raised by robots and Artificial Intelligence (RAIs) under five broad rubrics: functionality, inherent significance, rights and responsibilities, side-effects, and threats. The first letter of each rubric taken together conveniently generates the acronym FIRST. Special attention is given to the rubrics of functionality and inherent significance given the centrality of the former and the tendency to neglect the latter in virtue of its somewhat nebulous and contested character. In addition to exploring some illustrative issues arising under each rubric, the article also emphasizes a number of more general themes. These include: the multiplicity of interacting levels on which ethical questions about RAIs arise, the need to recognise that RAIs potentially implicate the full gamut of human values (rather than exclusively or primarily some readily identifiable sub-set of ethical or legal principles), and the need for practically salient ethical reflection on RAIs to be informed by a realistic appreciation of their existing and foreseeable capacities.

1. INTRODUCTION

For almost all of human history, robots existed only as imaginary beings endowed with a Jekyll-and-Hyde character. In one guise, promising to usher in a utopia free of illness, poverty, and the drudgery of work; in another, intent on enslaving or destroying humankind. Only in the middle of last century, however, did robots achieve a significant real-world presence when General Motors installed a robot, 'Unimate', in one of its plants to carry out manual tasks—such as welding and spraying—that were deemed too hazardous for human workers.[1] Today, robots are so commonplace in manufacturing that they are a major cause of unemployment in that sector.[2] But the use of robots in factories is only the beginning of a 'robot revolution'—itself part of wider developments powered by the science of Artificial Intelligence (AI)—that has had, or promises to have, transformative effects on all aspects of our lives.

Robots are now being used, or being developed for use, in a vast array of settings. Driverless cars have already been invented and are expected to appear on our roads within a decade. These cars have the potential to reduce traffic accidents, which currently claim more than a million lives each year worldwide, by up to 90%, while also reducing pollution and traffic congestion (Bonnefon, Shariff, Rhawan 2006). Robots are also used to perform domestic chores, including vacuuming, ironing, and walking pets. In medicine and social care, robots surpass doctors in diagnosing certain forms of cancer or performing surgery, and they are used in therapy for children with autism or in the care of the elderly. Tutor robots already exist, as do social robots that provide companionship, or even sex. In the business world, AI figures heavily in the stock market, where computers make most decisions automatically, and in the insurance and mortgage industries. Even the recruitment of human workers is turning into a largely automated process, with many rejected job applications never being scrutinized by human eyes. AI-based technology, some of it robotic, also plays a role in the criminal justice system, assisting in policing and decisions on bail, sentencing, and parole. The development of autonomous weapons systems (AWSs), which select and attack military targets without human intervention, promises a new era in military defence. And this is just a sample of recent developments.

In this article, I examine some of the key ethical questions posed by robots and AI (or RAIs, as I shall refer to them). The overall challenge, of course, is to harness the benefits of RAIs while responding adequately to the risks we incur in doing so. The need to balance benefit and risk is a recurrent one in the history of technological advance, but RAIs present it in a new and potentially sweeping form with large-scale implications for how we live among others—in relation to work, care, education, play, friendship, love—and even regarding how we understand what it is to be a

[1] http://my.ilstu.edu/~kldevin/Introduction_to_robotics2/Introduction_to_robotics6.html

[2] By 2012, for example, there were approximately 1,563 robots per worker in Japan's automotive industry (the figure for Germany was around 1,133 robots per worker), see Furman and Seamans (2018, 8).

human being and whether we should deploy these new technologies in the pursuit of 'human enhancement' or even, as with 'trans-humanism', in order to transcend our human condition. Prior to addressing these matters, we must first clarify some key notions. [3]

2. WHAT IS A ROBOT? WHAT IS ARTIFICIAL INTELLIGENCE?

A recent UNESCO report describes robots as artificial beings with four characteristics:

- *mobility*, which is important for functioning in human environments like hospitals and offices;
- *interactivity*, made possible by sensors and actuators, which gather relevant information from the environment and enable a robot to act upon this environment;
- *communication*, made possible by computer interfaces or voice recognition and speech synthesis systems; and
- *autonomy*, in the sense of an ability to 'think' for themselves and make their own decisions to act upon the environment, without direct external control (UNESCO 2017: 4).

The sophisticated robots that are our topic in this article operate on the basis of 'Artificial Intelligence' (AI). In the words of AI pioneer Marvin Minsky, this is 'the science of making machines do things that would require intelligence if done by men', such as face recognition or language translation. In understanding AI, two distinctions are important: (a) general and narrow AI, and (b) top-down and bottom-up AI. The first distinction relates to the scope of AI capabilities, the other to AI's technical functioning.

General AI refers to intelligent machines that are able to replicate a broad range of human intellectual capacities, and even to surpass them. These forms of AI, although familiar from science fiction characters such as *Star Wars*' C3PO, lie at best in the very remote future. To the extent that there has been any significant progress in AI in recent years, it has occurred in *narrow AI*. These are machines that replicate, or exceed, human capabilities with respect to a limited range of tasks, e.g. car-driving, medical diagnosis or language translation.

[3] For helpful overviews of developments robotics and AI, and the ethical issues they raise, Edmonds (2017) and Tegmark (2017, esp ch.3).

AI operates by means of algorithms, which are rules or instructions for the solution of various problems that are usually embedded in a computer, and for present purposes can be roughly grouped into two broad kinds, corresponding to two kinds of robots. *Top-down* (or deterministic or closed-rule) algorithms control a robot's behavior by means of a pre-determined program, with the result that the robot's behavior is highly predictable. Such algorithms have been used in the preparation of income tax forms and in certain kinds of automated medical diagnoses. *Bottom-up* (or stochastic) algorithms, by contrast, enable a robot to 'learn' from past experience and revise its algorithm over time (UNESCO 2017, 4, 17-19). An illustration of this 'machine learning' is Google's DeepMind algorithm, which taught itself to play Atari games such as Breakout, inventing new score-maximizing strategies that took its own programmers by surprise. Other examples are to be found in driverless cars, facial recognition systems used by police, and algorithms recommending items to buy based on one's purchasing history. There are different kinds of 'machine learning'. Some deploy 'neural networks', which are processing nodes connected to one another in layers and modelled on the functioning of the human brain. Robots of this second sort enjoy a level of 'autonomy' not only in the sense that their behavior need not depend on human decision-making, or may not be subject to human intervention or control, but in the more radical sense that it is not readily predictable by human beings.

We should, of course, exercise caution when throwing around terms like 'intelligence', 'reasoning', 'decision' and 'autonomy' in relation to AI. These terms must not obscure the fact that a vast chasm still separates RAIs and human beings. AI systems process information as a means of recognizing patterns and relations among symbols that enable certain problems to be solved. But they cannot (as yet) *understand* in any meaningful sense what these symbols stand for in the real world (Tegmark 2017, ch.3). Moreover, even if RAIs can be successful in achieving complex goals—like recognizing a face in a crowd or translating a document from one natural language into another—they lack anything like the human capacity to deliberate about what their ultimate goals ought to be. For some philosophers, this power of rational autonomy is the source of the special dignity that inheres in human beings and differentiates them from non-human animals. No RAIs known to us, or that are realistically foreseeable, are anywhere near exhibiting such rational autonomy. [4]

[4] For some scepticism about the hype surrounding Artificial Intelligence, by one of the world's leading computer scientists, see Jordan (2018).

3. ETHICAL QUESTIONS: FRAMES AND LEVELS

RAIs generate a variety of ethical questions which arise on at least three interconnected levels. One level concerns the *laws* that should be enacted to govern RAI-related activities. These laws are public standards that purport to be morally binding on all citizens in virtue of their formal enactment and which are standardly backed up by institutional enforcement mechanisms including, at the limit, punishments such as fines and imprisonment. One set of questions here concerns whether some particular aspect of RAIs should be subject to legal regulation at all; another set of questions concerns the extent to which we need to fashion specific laws to address issues thrown up by RAIs, as opposed to relying on more general legal standards. Do we need special traffic laws for driverless cars? How should the law on insurance and accident liability apply to them? Should there be criminal laws prohibiting certain kinds of robots or AI applications? In addition to domestic legislation on such matters, RAIs also raise pressing questions that require regional or international legal solutions, e.g. through treaties to outlaw AWSs or prevent the outbreak of an arms race in relation to them.

At a second level are questions about the kind of *social morality* that we should strive cultivate in relation to RAIs. This is a recognition of the fact that not all of the socially entrenched standards that properly govern our lives are, or should be, legal standards. We rely not only on the law to discourage people from wrongful behavior, such as murder or theft, but also on moral standards that are instilled in us from childhood and reinforced by society through informal mechanisms such as criticism and other extra-legal sanctions. Arguably, the very efficacy of legal regulation would be severely diminished if it could not rely upon a sustaining underlying ethical culture. Accordingly, we need to reflect on the shape of a morally sound culture in relation to RAIs.

At a third level, there are questions that arise for *individuals and associations* (e.g. businesses, universities, professional bodies, etc.) regarding their engagement with RAIs. Whatever social modes of regulation exist on these matters, individuals and associations will still need to exercise their own moral judgement. This may be because existing law and social morality lag behind technical developments, or because they are deficient in some way, or because they confer on individuals the leeway to make their own decisions on some matters. For those corporations which are at the cutting-edge of developments, the fast-changing and transformative character of RAIs may justify the elaboration of their own codes of ethics on these topics. Meanwhile, others have called for a 'Hippocratic oath' for data scientists to establish an ethical framework for their operations independently of applicable legal standards (Upchurch 2018).

Difficult questions arise as to how best to integrate these three modes of regulating RAIs, and there is a serious worry about the tendency of industry-based codes of ethics to upstage democratically enacted law in this domain, especially given the considerable political clout wielded by the small number of technology companies that are driving RAI-related developments. However, this very clout creates the ever-present danger that powerful corporations may be able to shape any resulting laws in ways favourable to their interests rather than the common good (Nemitz 2018, 7). Part of the difficulty here stems from the fact that three levels of ethical regulation interrelate in complex ways. For example, it may be that there are strong moral reasons against adults creating or using a robot as a sexual partner (third level). But, out of respect for their individual autonomy, they should be legally free to do so (first level). However, there may also be good reasons to cultivate a social morality that generally frowns upon such activities (second level), so that the sale and public display of sex robots is legally constrained in various ways (through zoning laws, taxation, age and advertising restrictions, etc.) akin to the legal restrictions on cigarettes or gambling (first level, again). Given this complexity, there is no a priori assurance of a single best way of integrating the three levels of regulation, although there will nonetheless be an imperative to converge on some universal standards at the first and second levels where the matter being addressed demands a uniform solution across different national jurisdictional boundaries.

Deepening this complexity is the fact that the fields of AI and robotics are both rapidly changing and the focus of considerable hype, making it hard to disentangle realistic future scenarios from mere science fantasy. In light of this, our ethical thinking at all three levels must be sensitive to the time-frame in question, sometimes addressing matters of immediate concern, other times anticipating future developments. A persistent danger is that we are distracted by potential developments that will arise, at best, in the very remote future, while neglecting pressing concerns in the here and now. In what follows, an attempt will be made to keep the focus on the here and now, as well as realistic future scenarios, although inevitably more speculative scenarios will also be broached.

4. FIVE MAJOR MORAL ISSUES—A F*I*R*S*T ANALYSIS

Many, if not all, of the moral questions raised by RAIs can be arranged under five main headings—functionality, inherent significance, rights and responsibilities, side-effects, and threats—with the first letter of each rubric conveniently generating the acronym 'FIRST'. Of course, the boundaries between the five distinct headings are not always sharp, and although I will usually refer to RAIs compendiously as a group, different kinds of RAIs will raise significantly different kinds of concerns under each of these five headings. The acronym is apposite because the issues discussed below are first-order questions about the rights and wrongs of our engagement with RAIs. There are, in addition, important second-order questions, regarding the procedures we should adopt in addressing these first-order issues, such as standards of transparency or democratic accountability. But these second-order matters are largely beyond the scope of this article. In what follows I focus primarily on functionality and inherent significance, giving only highly compressed treatments of the other three headings.

4.1 FUNCTIONALITY

The first issue is whether a proposed RAI, e.g. a driverless car, is functional. I take 'functionality' here in an expansive sense that is not neutral with respect to the moral quality of the ends pursued by a RAI or the means it adopts in pursuing them. Functionality concerns a RAI's ability to: (a) achieve a *worthwhile* goal, e.g. transporting passengers to their desired destination, and to do so: (b) *effectively* i.e. with a reliable degree of success, (c) *efficiently* i.e. without undue expenditure of resources, and (d) in a *morally appropriate way* i.e. without violating moral norms as an inherent part of its operation, irrespective of the intent of the designer, e.g. rights to life or privacy or norms of environmental protection. Although all these dimensions raise important questions, let us focus on the last one, which throws up two large questions: (1) what are the moral standards that apply to RAIs?, and (2) how can they be built into the operation of RAIs?

A famous attempt to address the first question is Isaac Asimov's 'Three Laws of Robotics':

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. (Asimov 1950, 40).

But Asimov's laws immediately run into problems. One is a lack of clarity about the concepts of 'injury' and 'harm' in the first law. If a robot bodyguard injures a would-be assassin in the course of protecting an innocent person, has it 'injured' or 'harmed' him? The interests of the assassin have obviously been impaired, but has he been wronged? We need to distinguish between non-moralized and moralized conceptions of harm or injury (or, in the legal version of the distinction, *damnum* and *injuria*). When we do so, it seems unlikely that a complete ban on RAIs harming human beings in the non-moralized sense will be sustainable. Indeed, even requiring that RAIs never wrong a human being may underestimate the complexity of the dilemmas RAIs may legitimately confront.[5] A familiar dilemma concerns how a self-driving car should respond to situations where there is a choice between avoiding harm to its passenger—e.g. by swerving out of the path of an oncoming truck—versus avoiding harm to other humans (drivers, passengers or pedestrians) who are at risk of death or injury if the car swerves to save its passenger. This 'trolley problem' receives conflicting responses, but any plausible answer seems to entail the all-things-permissibility of a wrong being done to a human being by a RAI. Interestingly, empirical studies indicate that most people agree that passengers should be sacrificed in order to save a greater number of bystanders, yet most would also prefer to ride in a car that always save its passenger (Bonnefon, Shariff, Rahwan 2006).[6] If so, identifying the correct answer to the trolley problem may turn out to be practically irrelevant, as not enough people would buy the first type of car to make it worth producing. Again, Asimov's second and third laws may be questioned if we are persuaded that advanced RAIs, with seemingly human-like intellectual and emotional traits, acquire something approximating human personhood and the rights, including to self-defence, that flow from it.

Asimov's principles are an early and rudimentary attempt at constructing an ethics for RAIs. But similar elementary difficulties plague more recent efforts, such

[5] Cf. also the fifth framework ethical principle outlined in House of Lords AI Committee 2018, 125: 'The autonomous power to hurt, destroy or deceive human being should never be vested in artificial intelligence'.

[6] For a discussion of ethical issues related to how self-driving cars should handle accidents, see Nyholm 2018.

as the Asilomar AI Principles formulated in 2017. Some of these principles verge on the truistic, e.g. principle 6 which requires that 'AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible'. Others are unexceptionable at the price of being unhelpfully vague, e.g. principle 15 which states that 'the economic prosperity created by AI should be shared broadly, to benefit all of humanity'. Two other tendencies of the principles are worth highlighting, since they are often replicated in other declarations of ethical principles for RAIs. The first is the implicit assumption that there is an enumerable catalogue of evaluative considerations that are especially engaged by RAIs. Thus, principle 11 demands compatibility of AI systems with 'ideals of human dignity, rights, freedoms, and cultural diversity'. But it is questionable that any meaningfully specific list of RAI-salient values is in order. Why not include additional values such as charity, respect for the natural environment or concern for the common good, among others? There is no reason *ab initio* to suppose that the ethical values potentially applicable to RAIs fall short of the entire range of human values. There is, of course, some recognition of this when the principles invoke other values, such as the common good. But here a second worrying feature crops up, which is the tendency to reduce the enumerated values to widely held *beliefs* about value. Hence, principle 23 on the common good states: 'Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization'. There is a conflation here of two distinct notions of the common good: (1) ethical values that are in fact widely shared among human beings, and (2) that which would objectively benefit all human beings. The latter is a normative idea, the former an empirical one whose normative implications, if any, need to be worked out in tandem with genuinely normative principles. The problem does not go away if an appeal is made to law, rather than widely held beliefs. The European Commission's recently published *Ethics Guidelines for Trustworthy AI*, for instance, accord a foundational role to human rights law.[7] Leave aside the fact that this law does not reflect all of the ethical considerations (e.g. environmental values) bearing on AI, that it does not tend to be directly binding on non-state actors, and that not all of its provisions bind all states (e.g. because they have not ratified relevant human rights treaties). The

[7] "We believe in an approach to AI ethics based on the fundamental rights enshrined in the EU Treaties, the EU Charter and international human rights law. Respect for fundamental rights, within a framework of democracy and the rule of law, provides the most promising foundations for identifying abstract ethical principles and values, which can be operationalised in the context of AI". European Commission 2019, 9.

more fundamental point is that such laws—despite the powerful moral charge conferred by the words 'human rights'—are not basic ethical standards. Instead, like any other set of laws, they are themselves to be formulated and evaluated—and, sometimes, to be found seriously wanting—in terms of basic ethical standards, including the morality of human rights (see Tasioulas (forthcoming)).

A sound ethical approach to RAIs, therefore, must go beyond invoking widespread beliefs or established law, including human rights law, to engage the full gamut of relevant ethical values. The tendency to conflate the normative with the empirical, conventional or legal is perhaps to be expected among those with a technological and 'data-driven' mind-set, who for understandable reasons tend to be prevalent in the robotics and AI community. It can lead them to the disastrous conclusion that ethical standards are to be identified through the deployment of empirical methods—for example, 'crowd-sourcing'—for ascertaining widespread ethical beliefs. Down this path, however, lies the degeneration of ethics into a branch of the public relations industry.

The second question, about how to build ethical norms into the workings of RAIs, is no less challenging. Some entertain high ambitions for robot moral sages commanding an expert knowledge of morality that far surpasses the average human. Julian Savulescu and Hannah Maslen contend that 'artificial ethical agents', thanks to their superhuman speed, extensive data-bases, and lack of characteristic human vices, such as selfishness, could 'actually aid or even replace humans when it comes to difficult moral decision-making' (Savulescu and Maslen 2015). Along similar lines, RAIs have been proposed as a means of overcoming the biases that notoriously afflict human judges in the sentencing of criminals—such as sentencing more leniently or harshly depending on the time of day or, more worryingly, in response to the class, ethnicity or race of offenders. [8]

Others, such as the authors of a recent UNESCO report, are sceptical about the moral perfectibility of RAIs.[9] Such scepticism has two inter-related sources. First, that moral decision-making confronts a potential infinity of relevantly different situations that no algorithm or process of machine learning is sensitive enough to engage

[8] See, for example, Sunstein (forthcoming), on how algorithms can help correct for the Current Offender Bias—the tendency to place an excessive emphasis on the fact of the current offence—in decisions about bail. More generally, for the claim that human cognition is more of a 'black box' compared to the levels of transparency that can be potentially achieved in relation to detecting discrimination by algorithms, see Kleinberg, Ludwig, Mullainathan and Sunstein (forthcoming).

[9] 'It does not seem probable that any machine that lacks emotions like empathy… could deal with this variation of morally relevant facts and preferences'. UNESCO (2017, 44).

with adequately. And, second, that sound moral reasoning requires the cultivation of emotional responses on the part of the reasoner, such as guilt, indignation, and empathy, that are properly attuned to their objects. It is these responses that enable us to register the moral significance of certain situations, e.g. the need to act with urgency in situations warranting fear of an imminent threat. But, arguably, they are inherently beyond the capacities of beings that do not share in a human consciousness and way of life. Both lines of thought have been stressed by a diversity of traditions in moral philosophy, most prominently in recent times by neo-Aristotelian virtue ethics and feminist theory. But they also receive some support from the behavioural and brain sciences, which suggest that the capacity for emotion is not a distinct 'module' that is added to our cognitive machinery, but an integral part of the overall architecture of our brains (Pessoa 2018).

Whether RAIs can become moral experts will partly depend on what is the correct philosophical account of morality. If the correct view is something like utilitarianism, which rests on a single very general ethical principle and requires daunting calculations of future consequences, the prospects for RAI sages may seem bright. If, by contrast, the correct moral approach requires context-specific judgment drawing upon a rich palette of moral emotions attuned to a plurality of values, and gives no role to mechanically applicable general principles, then the prospects seem correspondingly bleak.

Even if we set aside futuristic speculations about RAIs as moral experts, and focus instead on their compliance with basic moral norms in the performance of various specific tasks, the truth about morality will have an important bearing on how such norms are best integrated into the workings of robots. Of course, much will depend on what sort of tasks we wish robots to perform and what kind of settings they will operate in, e.g. whether this will be with or without potential human supervision or override. However, it would be unrealistic to suppose that we have to resolve disagreements in moral philosophy before we program ethical principles into the workings of RAIs. This is because proponents of different moral philosophies can still have good reason to converge on some core of basic moral standards, even if they would offer different justifications and (in many cases, interpretations) of them.

As we saw, two kinds of approaches to instilling ethical standards into the operations of RAIs can be distinguished, although elements of both approaches could be blended in any given RAI. The first is a top-down approach that involves rendering certain principles—such as the Geneva Conventions on the conduct of war, in the case of AWSs, or norms of conversational reasonableness and non-offensiveness for 'chatbots'—into algorithmic form. This is a daunting task, one which, if successfully carried out, would enable a RAI to make sophisticated judgments of proportionality when using lethal force or to distinguish between playful humour and offensive slurs. Another approach is more bottom up in character. It would proceed on the basis of a form of 'machine learning' in which, for example, the RAI might be exposed to a vast number of past decisions made by legal experts in the relevant field and then proceeds to extrapolate to decisions of its own in future scenarios. Taking their cue from virtue ethics, some have even argued that the right way to instil ethics into RAIs is to raise them as we do children, on the basis that the development of good character requires a decent upbringing (Rini 2017).

Although RAIs promise to assist us in meeting certain challenges, including overcoming human imperfections and limitations in carrying out important tasks, they are often defective in complying with appropriate standards for achieving an otherwise valuable goal. As we noted above, AI is powered by enormous data-sets. But the means by which data is accumulated is often morally dubious. One notorious example is the targeted advertising carried out by online platforms like Facebook and Google, from which they derive around 90% of their revenue. It might be convenient, if also unnerving, to have advertisements appear in your online newsfeeds that are tailored to your tastes and interests. But this process involves algorithms operating on data gleaned from the websites you visit, the emails you send, and mobile tracking. And there is a serious question as to whether people using these platforms are aware that they are 'data cows' being relentlessly milked for commercially valuable information, let alone whether they have meaningfully consented to it. As a result, these business models have rightly sparked concerns that they violate privacy rights or constitute forms of economic exploitation. Similar concerns extend to many other activities, including platforms such as that developed by the Axon corporation, which store more than 20 million gigabytes of public safety-related data taken from police body cameras (Goode 2018).

Giving individuals the right to control what is done with their personal data is not an all-purpose solution to the ethical problems created by the gathering and use of such data. One reason for this is that data may be needed to advance a vital social good—for example, to prevent the outbreak of a contagious disease or to anticipate a terrorist attack. In such cases, giving individuals a veto over whether their data is accessed or used in certain ways seems disproportionate to the value of the good that is forgone. However, the pursuit of social goods in this way in the absence of individual consent may require the fulfilment of other stringent conditions, such as transparency in the aims and methods of data-users, mechanisms for holding them to account, and so on.

Another defect that those working in the field have struggled to eradicate is that of algorithmic bias, which can arise even if the means of capturing data do not violate moral norms, such as those of privacy and non-exploitation. RAIs are driven by algorithms that are trained on datasets and operate by generalizing from them to future scenarios. One problem arises from the fact that the training data may itself be defective as a basis for accurate judgments or decisions. This is unsurprising, since the data are generated by the very fallible creatures (human beings) whose shortcomings, many of which are congealed in the form of harmful and unjust social patterns of behaviour, the RAIs were developed to overcome in the first place. In particular, the data may be statistically skewed, e.g. not inclusive of minority groups or embodying prejudices and historical patterns of discrimination. Recent examples of real-life algorithmic bias include an algorithm used by an English police force that discriminated against people from poorer areas in deciding whether to keep offenders in custody, job search tools that favoured men over women for high income jobs, and facial recognition algorithms, used in a range of applications from internet image searches to police enforcement, that have much higher error rates for women and non-white people (Oswald, Grace, Unwin and Barnes 2018; Burgess 2017; O'Neil 2016; and Buolamwini and Gebru 2018). A particularly egregious illustration is the risk-assessment algorithm COMPAS, which has been used by some American courts in making sentencing decisions. Its aim is to predict the likelihood of an offender re-offending within the next two years, an aim it fulfils at a 70% accuracy rate. However, its errors exhibited a pronounced racial bias: COMPAS was twice as likely to make an erroneous finding of high risk in the case of a black offender than a white offender (a 'false positive'), meanwhile, white offenders were twice as likely to be erroneously labelled as low risk as compared with black offenders (a 'false negative') (Larson, Mattu, Kirchner and Angwin 2016). [10]

Now even if, say, a risk-assessing algorithm is effective in accurately predicting the likelihood that suspects possessing certain characteristics will offend, this does

[10] For a more recent discussion that places the COMPAS algorithm controversy within the wider philosophical debate about fairness and discrimination, see Binns (2018). See also, more generally, Barocas and Selbst, (2016).

not put an end to our ethical worries. For one thing, there is something problematic about a person having a judgment made about him on the basis of how *other* people, who share various of his characteristics, behaved in the past. Does not respect for his personal autonomy require that he be assessed on the basis of the distinct individual that he is, his previous history of words and deeds, rather than those of others who share his (perhaps unchosen) characteristics? The problem is aggravated in cases where the relevant characteristics, such as race, gender or poverty, are themselves the focus, or the products, of a long history of grave injustice. The result is a 'vicious cycle' in which the RAI perpetuates and exacerbates the history of injustice that generated its dataset, e.g. by focusing on minority or poor suspects in making arrests, leading to greater numbers of minority convictions, which in turns leads to greater discrimination against minority suspects, in an endless downward spiral.

This problem of algorithmic bias cannot be solved by simply winnowing out data that is explicitly about sensitive categories such as age, class, gender or race, since other seemingly innocuous types of data may correlate with (or be proxies for) these categories. One important part of the answer to such challenges is to scrutinize carefully the purposes for which an algorithm is deployed (Fry 2018, 62). For example, it may be that in some contexts what really matters is the overriding issue of predicting outcomes, whereas in other contexts this is of lesser importance. Arguably, for example, determining which suspects should be given bail allows greater scope for a decision based on the predicted likelihood of a suspect committing an offence than does a sentencing decision, since the former does not carry anything like the same level of moral opprobrium as the latter. The challenges already mentioned to the just operation of algorithms are compounded by the fact often, for commercial reasons, neither the algorithm nor the data on which it is trained are made public. Moreover, in the case of bottom-up algorithm, it can be opaque even to the people operating the RAI precisely what algorithm is governing its activity.

If a RAI is functional in these ways discussed above, a question arises as to the level at which the requirement of functionality should be pitched, especially in comparison to what human beings can achieve. The answer will depend on the value of the goal being served and the risks we should be prepared to run in order to achieve it. In the case of some tasks, such as those performed by house-cleaning or companionship robots, it may be enough that a RAI performs adequately even if not quite as well as humans. Regarding other tasks, such as driving, medical diagnosis or the sentencing of criminals, what is at stake may be judged so important—life, liberty and justice—

that RAI functionality must be at least as good, or perhaps significantly better, than that ordinarily achieved by human beings. Superior performance by RAIs is something that may also be needed to offset the unwelcome side-effects and threats that reliance on them may also generate, such as job losses or sabotage by malicious agents (see 4.4 and 4.5, below). A complicating factor in identifying the minimum acceptable level of functionality is the baseline of comparison: is it what humans are in principle capable of achieving or what they are actually likely to achieve? For example, in a world in which millions of people lack access to basic education and health care, the lessons of a robot tutor or diagnoses by a robot physician may have significant value even if they are decidedly inferior to the services of human tutors and physicians that individuals cannot access, whether for financial or other reasons. It would be perverse to deprive people of vital benefits provided by RAIs in the name of an idealized level of human service that is likely to remain beyond their reach.

A further consideration we need to keep in mind is the dynamic quality of the moral standards bearing on RAIs; in particular, the way in which they may evolve over time as a result of technological developments and the new profile of opportunities and costs that they may come to generate. What I have in mind here is not a simple-minded one-way determination of our moral standards by technological developments, the sort of view that erroneously leads some to claim that 'privacy is dead' just because there is no foolproof way of preventing violations of privacy. Rather, the idea is that moral standards which were worked out in an era prior to the emergence of capabilities associated with RAIs may need to be reassessed in light of their emergence and the benefits they promise or the risks they pose. If, for example, the development of RAIs significantly increases the possibility of predicting the outbreak of epidemics by means of digital disease detection, we may need to develop less constraining norms regarding the kinds of digital surveillance RAIs may undertake to protect us from such threats. Forms of surveillance that, in the past, may have legitimately been judged violations of the right to privacy may, in this new technological environment, no longer be so (Vayena and Tasioulas 2016).

4.2 INHERENT SIGNIFICANCE

Even if RAIs can achieve sufficient functionality in a given task or role, questions may arise about the inherent significance of assigning that task or role to RAIs and not humans. Sometimes, the elimination of the 'human factor' may be beneficial.

In the case of an elderly person who needs assistance when bathing, for example, a robot carer can minimize the risk of embarrassment.[11] But the elimination of the human factor can also be troubling. This is already evident in a widespread concern that people should know whether the 'other' they are interacting with online or over the telephone is a human being or a machine. One context where the relevance of the human factor seems especially acute is that of RAIs making decisions that have serious consequences for human beings. What is mainly at issue here are especially (1) decisions based on bottom-up or stochastic algorithms which are therefore not reliably predictable, and (2) decisions that turn upon some kind of personal evaluation of the human being who is affected, e.g. their merit, desert or entitlements.

In reflecting on this concern, it will be useful to have in front of us a vivid, if compressed and contestable, reminder of the significance of being a human, especially as contrasted with being an artefact such as a machine. In this regard, we can do little better than the following statement by David Wiggins:

> *[O]ur sharing in a given specific animal nature and a law-sustained mode of activity is integral to the close attunement of person to person in language and integral to the human sensibilities that make interpretation possible. Secondly, that sharing in a specific animal nature and mode of activity is a precondition of the human solidarity (where present) that excoriates the treatment of a human being—of one of us—as a mere thing or a mere tool. And, thirdly, that affinity in nature and activity is integral to our picture—a non-deterministic picture—of our capacity, singly and collectively, to determine, within a framework not of our own choosing and replete with meanings that are larger than we are, our direct and indirect ends. Within this framework we can find our place and exercise our capacities. We see ourselves not as things with a function—what on earth could a person, as a person, be for?—but as autonomous, self-moving, animate beings, beings who find themselves in the world and seek to leave their own mark upon it, make the best of what they find there, and look (if they are lucky) for something that each one of us can come to think of as his or her own proper work or calling (Wiggins 2016, 91).*

In this passage, Wiggins accords a threefold significance to a shared human nature: *hermeneutically*, it enables a certain kind of mutual understanding of one

[11] For a thoughtful discussion of the use of RAIs in health and social care, in light how technological advances impact on human relations in the context of modernity, see Coeckelbergh (2015).

person by another, whether or not mediated by language; *ethically*, it is a basis for a certain kind of solidarity premised on a shared inherent value that is inconsistent, among other things, with the treatment of fellow humans as mere tools for the furtherance of our individual ends; and, *metaphysically*, humans have the rational autonomy to determine their life-shaping ends, as opposed to artifacts whose nature and activity is given by some fixed and specific purpose set by others.

Let us now return to our concern with the inherent significance of assigning decision-making functions to RAIs. Consider, for example, the plight of long-term unemployed people whose job applications are routinely rejected by the automated systems that now dominate workforce recruitment. After months or even years of applying unsuccessfully for jobs, those individuals may never once have their application read and evaluated by a fellow human. Even if we assume that the relevant algorithm meets a good standard of functionality, i.e. it is just as effective, efficient and compliant with norms of appropriateness as the average human recruiter, the fact that it is a non-human mode of decision-making is worrisome. It is hard to pin down the worry very precisely, but the thought is roughly that the job seeker is subjected to a cold, alienating, and ultimately potentially disrespectful process because his application never comes to the attention of a fellow human being. So much is suggested in this extract from a recent *Guardian* article:

> *"It's a bit dehumanising, never being able to get through to an employer," says Robert, a plumber in his forties who uses job boards and recruiters to find temporary work. Harry, 24, has been searching for a job for four months. In retail, where he is looking, "just about every job" has some sort of test or game, anything from personality to maths, to screen out applicants. He completes four or five tests a week as jobs are posted. The rejections are often instant, although some service providers offer time-delay rejection emails, presumably to maintain the illusion that a person had spent time judging an application that had already failed an automated screen (Buryani 2018).*

Or consider two other illustrations that engage even more vital interests: the sentencing of criminals by RAIs and the use of autonomous weapons systems (AWSs). Of course, there is a serious question as to whether RAIs can achieve a level of functionality comparable, or superior, to that of human judges and soldiers. [12] But even if

[12] For a comparatively upbeat view of the possibilities here, see Turner (2018).

they can, we might be troubled by the fact that a RAI's decisions impact so drastically on human life and liberty. Now, a sceptic might ask: what justified complaint can an accused or enemy soldier possibly have about the *mere fact* that they were sentenced, or killed, by a non-human being? Is the worry here nothing more than nostalgia masquerading as an ethical qualm? Perhaps one way of articulating the concern is by means of the thought that decisions about the life and liberty of others are so significant, something of value is lost if they are not made by an agent who can take responsibility for them. And, paradigmatically, for us, that agent is a human being—someone who can understand and empathise with our plight as a fellow human and reach a decision in light of their own autonomous assessment of the reasons in play for treating us one way rather than another. Here, all three of Wiggins' dimensions of the human being are in play. A corollary of this is that there is no decision-maker of whom one can demand *their* reasons for reaching the decision they did and whom one can hold accountable given an evaluation of those reasons. There is a valuable human solidarity and reciprocity—human beings recognizing each other as fellow human beings and forming their attitudes and decisions about each other on that basis—that is lost in the context of dehumanized, fully automated decision-making.

To this we can add a further observation. It will often be the case that the relevant reasons bearing on how another should be treated can be balanced in a variety of ways, with a range of decisions being rationally eligible, but no one decision being the single correct answer. In sentencing criminals, for example, the range of sentences open to the judge will typically vary in severity within certain rough bounds, with no exact amount of punishment being determinately singled out as the one which is uniquely justified. In response to this situation, some will prize consistency in the handling of similar cases above all else, claiming that criminals who commit the same wrong must receive exactly the same amount of punishment. RAIs, they might insist, are especially well-placed to treat 'like cases alike' in this way. But others see a place for discretion on the part of the judge—and the consequent variation in sentences that this generates. For those who adopt the second approach, there is value in a merciful judge being able to express their values and their character by, for example, choosing a more lenient sentence from a range of eligible options. There is value in a criminal justice system that offers offenders the possibility of discretionary mercy. The granting of mercy here is a kind of gift-giving, by one person to another, perhaps reflecting the hope of the former that the latter has genuinely repented of their past wrong-doing. In the case of the automated sentencing system, this value would be severely curtailed or eliminated. The robot is not an individual, with values and a character of its own, who can respond to the offender's plea for mercy as one human being to another, choosing a more lenient sentence when a harsher one is also rationally open.

Now, even if there is a generalized concern about RAIs making important decisions bearing on the interests of human beings, there are still two further issues that need to be addressed. The first is the weight that the human factor possesses in any given case. Is it simply one reason among others, or does it rise to the level of an obligation, perhaps even one associated with a human right? A report by the Rathenau Institut has proposed a human right to meaningful human contact (as well as a right not to be measured, analysed, or coached) (UNESCO 2017, 39), one that is presumably threatened in cases of the three kinds discussed above. Meanwhile, the UNESCO report previously mentioned rejects AWSs on the basis of a 'guiding principle that machines should not be making life or death decisions about humans', since delegating the decision to kill a human to a machine is an affront to human dignity (UNESCO 2017, 54).[13] Of course, even if a weighty value is being sacrificed, it may be that, all-things-considered, the benefits of robot judges or AWSs justify that sacrifice in some of these cases. More radically still, some contend that we should welcome the eventual erasure of the 'human-machine divide' through the emergence of human-machine hybrids or cyborgs (see 4.3, below). More realistically, however, even if we accord inherent significance to the human factor, its weight is liable to vary from one domain to another: perhaps in cancer diagnosis it is of no or negligible value, with everything being subordinated to the question of which is the most accurate method of diagnosis, whereas in criminal sentencing it seems intelligible to accord significance to the human factor, even if this results in an overall reduction in the soundness of sentencing decisions.

The second issue relates to ways in which concerns about the loss of the human factor might be tempered while still giving RAIs an important role. Minimally, it might be thought, those affected by the decisions of robot judges should be able to seek disclosure of the basis on which those decisions were made. [14] This last require-

[13] For a useful discussion of robots designed to kill, see R. Sparrow, 'Killer Robots', *Journal of Applied Philosophy* 24 (2007); 62-77.

[14] This is in line with the 'right to an explanation' some claim arises from the 'right not to be subjected to automated decision-making' in the EU's General Data Protection Regulation (2016); but for some scepticism about whether any such (practicable) right has been established, see Wachter, Mittelstadt, and Floridi (2017).

ment is more demanding than it may appear, and not just because commercial incentives lead to algorithms being shrouded in secrecy but also because creators of RAIs often candidly admit that they do not fully understand how their creations, operating on the basis of bottom-up algorithms, manage to perform in the successful ways that they do. In addition, there may be scope for human supervision and override regarding sentences passed by robot judges. Or, in the job application case, it might be that some random sample of applications is evaluated by human recruiters, so that the system is not entirely devoid of direct human input. All these suggestions head towards the general conclusion that RAIs might play a valuable role in *assisting* humans to make important decisions rather than completely replacing them in performing such tasks. But it is doubtful that this could serve as an all-purpose solution, since there might be scenarios, such as battlefield decisions by AWSs, where comprehensive human supervision is simply incompatible with securing the benefits promised by RAIs.

I have so far focused on one broad category of cases in which the inherent fact that a RAI is performing a function has potential ethical significance: that in which an authoritative decision is made that impacts severely on human interests. But concern about the human factor also arises in other sorts of cases, not least those in which RAIs are treated as partners in relationships of friendship or intimacy. The absence of the human factor here can provoke a feeling of creepiness and revulsion. Its source lies in the introduction of a robot, which is ultimately a machine or instrument, into a form of relationship that—up until now—we have reserved for humans. Some will respond that the 'human factor' has here degenerated into a visceral 'yuck factor' and carries no more normative weight than the revulsion homophobes or racists feel towards same-sex or mixed-race couples. But such a dismissal would be too quick. Ideally, in a romantic relationship, the parties value each other not simply as means, but as ends in themselves, partly in virtue of their ability freely to reciprocate feelings and emotions such as love and affection. In this context, dating a robot with a human form may not be quite as absurd or weird as a romantic attachment to an iPhone or smart fridge, but it's arguably on a continuum with them.

Imagine now a social world in which large numbers of people prefer RAIs over humans as their friends and romantic partners, perhaps because they are programmable to the user's own specifications and lack the independence of mind that can be a source of friction, disappointment, and betrayal in the case of human friends and lovers. But the very word 'user' here highlights what such a relationship lacks as compared with the value potentially available in a personal relationship between autonomous human beings. An instrument is being made to play a role that belongs to a person. That the instrument may have certain good effects on the happiness or state of mind of the user does not erase the fact that it remains a *mere* instrument and not a person. The flip side of this point is the feeling that someone who prefers robot friends to human friends has not truly succeeded in growing up. It is somewhat as if a middle-aged man still treated his childhood teddy bear as his 'best friend'.

Of course, even if you agree that something of value is lost here, nothing automatically follows about whether people should choose relationships with robots, or whether such relationships should be condemned by our social morality or prohibited by law. After all, it may be that a relationship with a robot is the only kind of relationship realistically available to some people or that the loss of the human factor is compensated by other benefits the relationship brings. Or it may be that we should simply respect people's free choice to enter into such relationships even if we judge these relationships to be deeply deficient. But the starting-point for addressing any of these questions is deeper reflection on the value of something we have largely taken for granted: the presence of the human factor in our everyday lives. [15]

4.3 RIGHTS AND RESPONSIBILITIES

If a RAI is functional and there are no compelling inherent reasons against deploying it, questions arise as to whether it possesses a moral status that confers upon it rights and responsibilities. As an artifact, created to further our ends, it seems doubtful that a RAI can possess the inherent value required to ground such a status. If RAIs came close to replicating our general capacity for rational autonomy, there would be a case for according them a comparable moral status to human beings, with corresponding rights as well as responsibilities. Indeed, it may be that in coming years the human/machine divide itself will gradually blur and even disappear with the appearance of various hybrids or cyborgs (RAIs integrated with human beings). In the words of a proponent of 'transhumanism', Ray Kurzweil:

> *Computers started out as large remote machines in air-conditioned rooms tended by white coated technicians. Subsequently they moved onto our desks, then under*

[15] For explorations of some of the issues posed by the possibility of roots friends / lovers, see Danaher and McArthur Eds (2017).

> *our arms, and now in our pockets. Soon, we'll routinely put them inside our bodies and brains. Ultimately we will become more nonbiologial than biological (Kurzweil 2002. For a fuller discussion, Kurzweil 2005).*

Such developments could have radical implications for the content of the moral standards that apply to RAIs, such as a right to self-defence that could justify killing or harming humans who pose a threat. However, as we have already seen, general AI seems a very distant prospect, even if it cannot be totally ruled out as a logical possibility.

The more pressing question is whether a good case exists for attributing legal personality to RAIs, with corresponding legal rights and responsibilities, on analogy with other 'artificial persons' recognized by law, such as corporations.[16] The issue is not that of treating RAIs the same as human beings for all legal purposes since the legal personality of RAIs need not precisely match that enjoyed by ordinary human beings, or 'natural persons'. It may consist in a different, and probably considerably smaller, bundle of rights and obligations. And the relevant bundle may vary in content from one kind of RAI (self-driving cars) to another (health care RAIs).

The case for attributing legal personality to RAIs seems more plausible in relation to RAIs operating with bottom-up algorithms, given that their behavior is not fully predictable. In the case of RAIs operating on the basis of top-down algorithms, which render their behaviour highly predictable, the argument for attributing legal responsibility to manufacturers, owner or users seems compelling. Along these lines, the European Parliament has entertained the possibility of 'creating a special legal status for robots in the long run, so that at least the more sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently' (European Parliament 2017, para 59(f)). This proposal is reminiscent of the old English common law notion of a 'deodand', according to which animals or things, including (then) new technologies such as trains, could be forfeited if they caused damage to another person, without any fault being attributed to their owner. As a result of the forfeit, the value of the thing in question could be used to compensate anyone who had been harmed by it. The doctrine was eventually

[16] For some recent discussions, Schwitzgebel and Garza (2015); Gunkel (2018); Bryson, Diamantis, and Grant, (2017); and Turner (2018).

abolished in the nineteenth century and replaced by no-fault liability and insurance schemes to compensate those harmed by newly emerging technologies.

Given the assumption that the attribution of legal personality to RAIs would not be grounded in their independent moral personhood, the question of whether RAIs should be designated legal persons falls to be determined by reference to the overall balance of benefits and burdens for human beings, and other morally considerable beings, of adopting this legal innovation (Bryson, Diamantis, and Grant, 2017). Critics have drawn attention to various pitfalls of granting RAIs legal personality. One pitfall is the possibility of abuse (a consideration belonging to the category of threats, discussed at 4.5, below): the legal personality of RAIs might be used by other unscrupulous agents, e.g. their manufacturers and operators, in order to evade any responsibilities on their part. Another arises where there is no relevant human agent 'standing behind' the actions of a RAI: how will the RAI compensate others for its breach of obligation? (Bryson, Diamantis, and Grant, 2017). Would the successful litigant be given ownership of the RAI, or might RAIs possess assets, e.g. bank accounts, that could be claimed as compensation? But these are not conclusive objections, and they certainly do not definitively rule out the advisability of limited forms of legal personality for some RAIs.

Even if there are serious obstacles in the way of according legal personality to RAIs, it is arguable that bottom-up RAIs are sufficiently different from most machines that existing law must be revised to take into account their capacity for autonomous behaviour. Along these lines, the UNESCO report proposes a threefold scheme: (1) use of existing law in assigning responsibility in relation to top-down robots; (2) for bottom-up robots, use codes of practice and ethical guidelines in addition to the law; and (3) for bottom-up robots that can harm human beings, e.g. AWSs or self-driving cars, consideration of the degree to which autonomous decisions should be left to the robot and where meaningful human control should be required (UNESCO 2017, 48-9). It is doubtful, however, that any such neat schema is ultimately defensible. The injunction in (3) seems plausible, if rather vague, but (1) is misguided. There is no reason to suppose that only legal standards are needed in the case of top-down RAIs, any more than only legal responsibilities arise for the users of a standard car.

Underlying all of these moral and legal issues is the need to solve a difficult technical problem: how to ensure the 'traceability' of RAIs in order to be able to assign moral or legal responsibility in relation to them. Traceability involves being able to determine the causes that led a RAI to behave in the way that it did, and securing it seems especially difficult in the case of RAIs that involve bottom-up algorithms. Given the potential threat that RAIs pose to human interests, it seems hard to justify permitting their creation and use in any given case in the absence of an adequate answer to this challenge.

4.4 SIDE-EFFECTS

Side-effects, both good and bad, will inevitably result from the use of RAIs. One positive side-effect, for example, is that of affording people greater opportunities to develop their personal relationships or to pursue leisure activities. On the other hand, RAIs may cause significant levels of unemployment and, ultimately, serious levels of social deprivation, inequality, and unrest. These are, properly speaking, side-effects, since even RAIs intended to replace human labour do not have as part of their operational goal to cause unemployment or its attendant societal problems. These are at most foreseen but unintended consequences.

Economists disagree about the potential impact of the widespread use of RAIs on human employment. One study has found that up to 50% of American jobs, including jobs performed by lawyers, doctors and accountants, are at risk of being automated (Frey and Osborne 2013). In the United Kingdom, more than one in three jobs could be taken over by RAIs in the next twenty years, with the impact disproportionately affecting those in repetitive, low-paid jobs (Tovey, 2014). But these developments may not necessarily cause significant unemployment, as the history of technological innovation reveals: new jobs, many of which we currently have no inkling, might emerge, partly as a result of productivity gains made by novel technology, and often responding to new wants that technological advances have themselves helped generate (Milanovic 2016; Autor 2015). Others take a more pessimistic view, especially if they anticipate that RAIs will acquire superhuman capacities that will render pretty much all human workers obsolete (Brynjolfson and McAfee 2014; Drury 2018; Aeon 2018). Whatever the truth turns out to be, it is likely that over the shorter term many people will lose their jobs to RAIs and will face great difficulty in retraining for whatever new jobs—perhaps centred on skills requiring judgment, creativity or emotional intelligence—might eventually emerge.

These possibilities force us to reassess the value of work in human life. Partly this value has to do with earning income, and so one popular response to the danger of RAI-induced unemployment is the policy of a Universal Basic Income (UBI) financed by the increased productivity achieved through the deployment of RAIs (Van Parijs and Vanderborght 2017).[17] A UBI would provide a regular and unconditional cash payment to all irrespective of whether they work or fulfil any further conditions such as having a disability or actively seeking work. But work also potentially serves values aside from the generation of income: it is an important source of accomplishment and self-worth, it fosters virtues of responsibility and self-discipline, and it provides a focus for valuable social engagement. Will we need to limit, in some way, the incursion of RAIs into our economic life in order to preserve adequate human access to these values? Perhaps we should use RAIs mainly as tools to assist, rather than replace, human workers. Would limiting the use of RAIs to preserve a sphere for human achievement be self-defeating, insofar as these 'achievements' would be shadowed by the knowledge that RAIs could have done an equivalent, or even far superior, job? Or are there feasible, and perhaps even preferable, ways of realizing these values through other pursuits, such as family life, art, religious worship or sport? Or, in a world liberated from the necessity of human work, might we discover an altogether new set of values to give meaning to our lives?

Unemployment is just one worrying potential side-effect of RAIs. A more diffuse worry, along broadly similar lines, is that over-reliance on RAIs may lead to the atrophying of valuable skills and to the diminution of our sense of responsibility for our own lives and choices. Doctors, drivers and pilots, for example, may begin to lose the skills that they need in order to perform well in emergencies; ordinary people may become excessively reliant on RAIs for guidance in making day-to-day decisions about such matters as the food they eat, the newspapers they read, and the political parties they vote for at elections. Moreover, the greater our reliance on RAIs in making decisions, the more we may discover that our lives are increasingly shaped by the sorts of considerations to which automated decision-making is most sensitive. This need not mean that the 'outcomes' produced are inferior to those that would have resulted had a wider range of considerations been in view, but it does mean that the contours of our lives will increasingly be a function of the capabilities of RAIs, rather than deliberation about the full array of pathways made eligible by the salient value considerations.[18]

There are other serious worries, including potential corrosive effects on the

[17] Other possible responses include the provision of wage subsidies and guaranteed government employment; for a discussion of these three strategies, Furman and Seamans 2018, 21-25.

[18] For a discussion of this sort of side-effect, in relation to AI encouraging a move from more 'equitable' to more 'codified' forms of legal adjudication, see Re and Solow-Niederman (2019).

quality of our relations with other humans. The more that our lives revolve around interactions with machines—often endowed with anthropomorphic forms and voices—that are fashioned to service our desires, the more we risk succumbing to the temptation to extend the same instrumentalizing attitude towards our fellow humans. This is a concern that is especially acute in relation to RAIs used for companionship or sex. Others, as we have seen, contend that we should welcome the eventual erasure of the 'human-machine divide', a process that will be hastened by the appearance of human-machine hybrids or cyborgs.

4.5 THREATS

RAIs that are specifically designed to carry out malign goals, such as privacy-violating surveillance, financial fraud or terrorist attacks, can pose serious threats to our interests and values. Such RAIs are not functional, in the sense identified above (4.1). But threats can also arise from RAIs designed to perform worthwhile tasks being sabotaged or subverted in their functioning—for example, if their algorithms are purposefully fed with false or corrupted data, or if they are hacked by malevolent agents. A world in which your smart phone spies on you or AWSs fall into the hands of terrorists is hardly a far-fetched prospect. And the threats come not only from criminals, terrorist groups or corporations, but perhaps above all from governments, often working in collaboration with groups of other kinds. A striking recent example of RAIs being used as instruments of authoritarian rule is the Social Credit System set up by the Chinese government whereby individuals receive a 'citizen score', based on data collected about them, which is used to determine eligibility for jobs, foreign travel and other benefits (*The Economist* 2016).

The response that this kind of governmental threat does not really apply to democracies is unconvincing. It grossly underestimates the ways in which politicians and bureaucrats in democracies might seize on widespread fears, for example, about terrorism or immigration, to enact repressive measures and prolong their hold on power. Another major concern is the rise of 'preventive' or 'actuarial' approaches to policing whereby AI is used to predict future crimes, and would-be criminals are apprehended on the basis of what they are likely to have done rather than any crime they have actually committed, with all the attendant risks to civil liberties vividly portrayed in the film *Minority Report* (see, for example, Ferguson 2017). Moreover, an especially insidious phenomenon in this context is the way in which some suspect forms of government surveillance and corporate data-gathering are intertwined, as was dramatically shown by Edward Snowden's revelations about the NSA's access to data gathered by Google, Facebook, and Microsoft.[19]

One of the gravest threats posed by RAIs is to the proper functioning of democracy itself. [20] Democracy not only requires that all citizens have a vote, but also that they are able to exercise their vote after free and informed deliberation and debate on the issues that are at stake. It requires a free flow of information to enable democratic deliberation to shape policy formation and to ensure that office-holders are held accountable. Concerns have arisen in recent years about the use of RAIs to compromise these democratic processes. Methods used to this effect have included the micro-targeting of individuals with bespoke political advertisements based on data illicitly gleaned from social media platforms, such as Facebook, or using robot accounts — 'bots'—that masquerade as human in order to saturate Twitter and other platforms with propaganda, or the creation of virtually undetectable audio-visual forgeries. The threat here is not simply a matter of the illicit provenance of the data, nor even that the messages and images crafted may be deceptive or manipulative, but also the way in which such activities contribute to the formation of distinct informational 'universes' for different categories of voters, thereby eroding the common public sphere that is vital to democratic deliberation (Bartlett 2018). As to the magnitude of the threat to democracy, Onora O'Neill, has issued a stark warning:

> *Not deceiving is one of the fundamental duties. When I think about technology, I wonder whether we will have democracy in 20 years because if we cannot find ways to solve this problem, we will not. People are receiving messages and content which is distributed by robots, not by other human beings, let alone by other fellow citizens. It is frightening (O'Neill 2018; see also Helbing et al 2017).*

When allied to the potential side-effects of RAIs—undercutting both our sense of personal responsibility to our fellow citizens and our sense of the distinctiveness of humanity—the risks for democracy appear significant.

Ways of addressing the threat to democracy include technology-specific mea-

[19] For a discussion, including the point that government agencies often circumvent legal restrictions on data-gathering by buying, demanding, or hacking data held by corporate agents, see Pasquale 2015, 48-51.

[20] For an overview of the implications, both positive and negative, of digital technology for democracy, see Susskind (2018).

sures such as enhanced privacy protection for personal data, greater transparency regarding the use of data by online platform providers, and more stringent registration processes for social media accounts. O'Neill even raises the possibility of some form of internet censorship akin to that practiced by the Chinese government, which may be a cure worse than the disease it aims to treat. But it is important also to address structural features of our political systems that may interact with RAIs to subvert democracy. For example, the United States' lax campaign-financing laws make it easier for resources to be funnelled into the large-scale dissemination of 'fake news'. More generally, we face a conundrum. What is arguably needed to counter most of the societal threats posed by RAIs is enhanced democratic accountability; but this may involve us in a race against time, since one of the gravest of those threats is precisely to the functioning of democratic processes. We need democratic solutions to the problems posed by RAIs before they are used to destroy democracy itself.

Of course, according to some, the greatest threat posed by RAIs is the one from which this article began: that they become so much more intelligent than humans that they eventually subjugate or eradicate us in pursuit of their own ends. This doomsday scenario has been emphasized by prominent figures in the field of RAIs, such as Bill Gates and Elon Musk, as well as by leading scientists, including the late Stephen Hawking, who observed towards the end of his life:

> *A super-intelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we're in trouble... You're probably not an evil ant-hater who steps on ants out of malice, but if you're in charge of a hydroelectric green energy project and there's an anthill in the region to be flooded, too bad for the ants. Let's not place humanity in the position of those ants (Griffin 2015).*

Some have dismissed such warnings as irresponsibly speculative on the basis that general AI, let alone general AI of a superhuman form, is not a realistic prospect in the foreseeable future. As Daniel Dennett (2019) has put it, 'we're making tools not colleagues'. On this view, doomsday scenarios are fantasies that distract us from other, urgent RAI-related problems that we confront. But another response focuses on the assumption that underlies Hawking's warning. Why assume that the goals of super-intelligent RAIs will be troublingly unaligned with ours? If RAIs develop truly superhuman abilities, won't these include the abilities to reason about, and conform to, morality?[21] In other words, it is an impoverished concept of 'intelligence' to confine it to the capacity to achieve complex goals, regardless of their moral value. Imagine, then, a world in which we are governed by just and benevolent RAIs that far surpass any human in intelligence and goodness. Is this scenario the ultimate fulfilment of the promise of RAIs to serve humankind or a deep betrayal of our interests and values? That there is a case for the latter conclusion is evident. Partly, this is a matter of being subjected to forms of governance based on principles and considerations that potentially outrun the ability of most, or perhaps all, human beings to grasp, which is a massive deviation from the Enlightenment idea of rule under standards that can be rationally apprehended and approved by those subjected to them.[22] After all, given that human morality is attuned to the possibilities and limitations inherent in our human predicament, why should it be supposed that super-intelligent RAIs, who do not share a human nature, would be disposed to give much weight to anything we could recognize as moral considerations? But leaving this worry aside, a key value that the putative just and benevolent RAI governors would presumably have to acknowledge is that of human freedom, not only at the level of the individual human making personal life choices, but at the level of groups of human beings exercising their communal self-determination. It is difficult to see how their rule could avoid seriously undermining this freedom.[23] Perhaps, in light of these considerations, a benign race of super-RAIs would turn out to have the grace to leave humans to forge their own path, subject to the enforcement of a minimal set of norms that averted the more devastating manifestations of human error.

5. CONCLUSION

The task of developing a sound approach to the ethics of RAIs operates at multiple levels, including legal regulation, social morality, and personal moral standards, which interact in complex ways. The main first-order questions arising at these three

[21] For a more elaborate defence of the opposite, non-alignment thesis, as to the values of RAIs, see Bostrom (2014).

[22] A point emphasized in Kissinger (2018) 'The most difficult yet important question about the world into which we are headed is this: What will become of human consciousness if its own explanatory power is surpassed by AI, and societies are no longer able to interpret the world they inhabit in terms that are meaningful to them?'

[23] A further, more speculative consideration here is that rule by non-humans would be an attack on loyalty to, or identifying with, one's species, see Williams 2006, 149-152.

levels, I have suggested, can be for the most part illuminatingly collected under the FIRST schema, as pertaining to the rubrics of functionality, inherent significance, rights and responsibilities, side-effects, and threats. However, I have cast doubt on the idea that there is some helpful segment of existing moral or legal principle that primarily or exclusively bears on the ethics of RAIs. Instead, RAIs should be seen as potentially engaging the full gamut of human values under the five rubrics identified by the FIRST schema. Although work has already been done under all five of these rubrics, we are at an early stage of thinking through the ethics of RAIs. Moreover, the inherent significance of the human factor is a matter that has not as yet, for understandable reasons, received anything like the level of sustained attention it deserves. We need to grapple with the very idea of such significance and how it acquires varying forms of moral valence, depending upon features such as the domain of decision-making (e.g. cancer diagnosis or criminal sentencing) or the form of relationship (e.g. lawyer or lover) that is at issue. In addressing questions that arise under these five rubrics, it is essential to be guided and constrained by a realistic appreciation of the existing and foreseeable future capacities of RAIs, and not to allow our ethical thought to be hi-jacked by utopian (or dystopian) speculation based on possibilities that lie, at best, in the remote future, even if they do not strictly fall outside the realm of scientific and technological possibility.

ACKNOWLEDGEMENTS

I am grateful to Roger Brownsword, Rebecca Lowe, Claudia Chwalisz, Francesca Rossi, Jose Such, David Nelken, participants at a KCL/Queen's Ontario colloquium in legal philosophy, and—especially—Hannah Maslen, Annette Zimmermann, and an anonymous referee for helpful comments on previous drafts. Work on this paper was enabled by a grant from the Future of Life Institute. I am also grateful for the excellent research assistance provided by Napoleon Xanthoulis. A condensed version of this article is published as Tasioulas (forthcoming).

REFERENCES

Aeon. (2018). "Humans Need Not Apply." (video) *Aeon*, <https://aeon.co/videos/the-robots-are-coming-for-our-jobs-why-the-human-workforce-is-at-risk> [Accessed on 13 June 2019].

Asilomar AI Principles. (2017). <https://futureoflife.org/ai-principles/> [Accessed on 13 June 2019].

Asimov, I. (1950). "Runaround.". I, *Robot*. New York City: Doubleday.

Autor, D. (2015). "Why are There Still So Many Jobs? The History and Future of Workplace Automation." *Journal of Economic Perspectives* 29 (3): 30.

Barocas, S. and Selbst, A. (2016). "Big Data's Disparate Impact." *California Law Review* 104: 670-732.

Bartlett, J. (2018). Th*e People vs Tech: How the Internet is Killing Democracy (and How We Save It)*. London: Penguin.

Binns, R. (2018). "Fairness in Machine Learning: Lessons from Political Philosophy." *Journal of Machine Learning Research* 81: 1-11 <https://arxiv.org/abs/1712.03586> [Accessed on 13 June 2019].

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2006). "The Social Dilemma of Autonomous Vehicles." *Science* 352: 1573-1576. <http://science.sciencemag.org/content/352/6293/1573> [Accessed on 13 June 2019].

Bryson, J, Diamantis, M. and Grant, T. (2017). "Of, for and by the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence and Law* 25: 273-29.

Buolamwini, J. and Gebru, T. (2018). "Gender Shades: Intersectional Accuracy Disparity in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1-15. <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf> [Accessed on 13 June 2019].

Buranyi, S. (2018). "Dehumanising, impenetrable, and frustrating: the grim reality of job-hunting in the age of AI." *Guardian* March 4th <https://www.theguardian.com/inequality/2018/mar/04/dehumanising-impenetrable-frustrating-the-grim-reality-of-job-hunting-in-the-age-of-ai> [Accessed on 13 June 2019].

Brynjolfson, E. and McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York City, NY: WW Norton & Co.

Burgess, M. (2017). "Just Like Humans, Artificial Intelligence Can Be Sexist and Racist." *Wired* April 13. <http://www.wired.co.uk/article/machine-learning-bias-prejudice> [Accessed on 13 June 2019].

Coeckelbergh, M. (2015). "Artificial Agents, Good Care, and Modernity." *Theoretical Medicine and Bioethics* 36: 265-277.

Danaher, J. and McArthur, N. Eds. (2017), *Robot Sex: Social and Ethical Implications*. Cambridge MA: MIT Press.

Dennett, D.C. (2010). "Will AI Achieve Consciousness? Wrong Question." *Wired* (Feb 9) <https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question/> [Accessed on 13 June 2019]

Drury, C. (2018). "Mark Carney Warns Robots Taking Jobs Could Lead to Rise of Marxism." *The Independent* April 14 <https://www.independent.co.uk/news/uk/home-news/mark-carney-marxism-automation-bank-of-england-governor-job-losses-capitalism-a8304706.html> [Accessed on 13 June 2019].

Edmonds, D. (2017). "Can We Teach Robots Ethics?" *BBC News*, October 15, 2017 <http://www.bbc.co.uk/news/magazine-41504285> [Accessed on 13 June 2019].

European Commission (2019). *Ethics Guidelines for Trustworthy AI.* European Commission: Brussels. <https://ec.europa.eu/futurium/en/ai-alliance-consultation> [Accessed on 13 June 2019].

European Parliament (2017). Report with recommendations to the Commission on Civil Law Rules on Robotics, Jan 27. <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN> [Accessed on 13 June 2019].

Ferguson, A.G. (2017). *The Rise of Big Data Policing*. New York City, NY: New York University Press.

Frey, C.B. and Osborne, M.A. (2013). "The Future of Employment: How Susceptible Are Jobs to Computerisation." <https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf> [Accessed on 13 June 2019].

Fry, H. (2018). *Hello World: How to be Human in the Age of the Machine*. New York: Doubleday.

Furman, J. and Seamans, R. (2018). "AI and the Economy." <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3186591> [Accessed on 13 June 2019].

Goode, L. (2018). "Facial Recognition Software is Biased Towards White Men, Researcher Finds." *The Verge* (Feb 11). <https://www.theverge.com/2018/2/11/17001218/facial-recognition-software-accuracy-technology-mit-white-men-black-women-error> [Accessed on 13 June 2019].

Griffin, A. (2015). "Stephen Hawking: Artificial Intelligence Could Wipe Out Humanity When It Gets Too Clever as Humans Will be Like Ants." *The Independent* (October 8) <https://www.independent.co.uk/life-style/gadgets-and-tech/news/stephen-hawking-artificial-intelligence-could-wipe-out-humanity-when-it-gets-too-clever-as-humans-a6686496.html> [Accessed on 13 June 2019].

Gunkel, D. (2018). *Robot Rights*. Cambridge, MA: MIT Press.

Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R.V., and Zwitter, A. (2017). "Will Democracy Survive Big Data and Artificial Intelligence?" *Scientific American* (Feb 25 <https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/?redirect=1> [Accessed on 13 June 2019].

House of Lords AI Committee (2018). *AI in the UK; ready, willing and able?* House of Lords Paper 100. <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm> [Accessed on 13 June 2019].

Jordan, M. (2018). "Artificial Intelligence—The Revolution Hasn't Happened Yet." *Medium* April 18, <https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7> [Accessed on 13 June 2019].

Kissinger, H.A. (2018). "How the Enlightenment Ends." *The Atlantic* (June) <https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/> [Accessed on 13 June 2019].

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. (forthcoming). "Discrimination in the Age of Algorithms." <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669> [Accessed on 13 June 2019].

Kurzweil, R. (2002). "We Are Becoming Cyborgs." <http://www.kurzweilai.net/we-are-becoming-cyborgs> [Accessed on 13 June 2019].

——— (2005). *The Singularity is Near: When Humans Transcend Biology*. London: Duckworth.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). "How We Analysed the COMPAS Recidivism Algorithm." *ProPublica* May 23 <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> [Accessed on 13 June 2019].

Milanovic, B. (2016). "Three Fallacies that Make You Fear a Robot Economy." *Evonomics*, Sept 1. <http://evonomics.com/three-fallacies-robot-economy-branko/> [Accessed on 13 June 2019].

Nemitz, P. (2018). "Constitutional Democracy and Technology in the Age of Artificial Intelligence." *Philosophical Transactions Royal Society A* 376 (2133).

Nyholm, S. (2018). "The Ethics of Crashes with Self-Driving Cars: A Road Map, I." *Philosophy Compass* 17.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin.

———(2018). "I Wonder Whether We Will Have Democracy in 20 Years." Interview with Elena Cué. *Huffington Post* May 1st. <https://www.huffingtonpost.com/entry/onora-oneill-i-wonder-whether-we-will-have-democracy_us_5a4f8a12e4b0cd114bdb324f> [Accessed on 13 June 2019].

Oswald M., Grace, J., Urwin, S., and Barnes, G.C. (2018). "Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality." *Information & Communications Technology Law* 27 (2018): 223.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge MA: Harvard University Press, 2015.

Pessoa, L. (2018). "Robot cognition requires machines that both think and feel." *Aeon* April 13. <https://aeon.co/ideas/robot-cognition-requires-machines-that-both-think-and-feel> [Accessed on 13 June 2019].

Re, R.M. and Solow-Niederman, A. (forthcoming). "Developing Artificially Intelligent Justice." *Stanford Technology Law Review* 22.

Rini, R. (2017). "Raising Good Robots." *Aeon* April 18 <https://aeon.co/essays/creating-robots-capable-of-moral-reasoning-is-like-parenting> [Accessed on 13 June 2019].

Savulescu, J. and Maslen, H. (2015). "Moral enhancement and artificial intelligence: moral AI?" *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*. Eds. Romportl, J., Zackova, E., and Kelemen, J. Switzerland: Springer International Publishing.

Schwitzgebel, E. and M. Garza, M. (2015). "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39.

Sparrow, R. (2007). "Killer Robots." *Journal of Applied Philosophy* 24: 62-77.

Sunstein, C. (forthcoming). "Algorithms, Correcting Biases." *Social Research*.

Susskind, J. (2018). *Future Politics: Living Together in a World Transformed by Tech*. Oxford: Oxford University Press.

Tasioulas, J. (2019). "AI and Robot Ethics." *Ethics and the Contemporary World*. Ed. Edmonds, D. London: Routledge.

———(forthcoming). "Saving Human Rights from Human Rights Law." *Vanderbilt Journal of Transnational Law*.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. London: Penguin.

The Economist. (2016). "Big Data, Meet Big Brother: China Invents the Digital Totalitarian State." *The Economist* December 17. <https://www.economist.com/briefing/2016/12/17/china-invents-the-digital-totalitarian-state> [Accessed on 13 June 2019].

Tovey, A. (2014). "Ten Million Jobs at Risk from Advancing Technology." *Telegraph* (Nov. 10) <https://www.telegraph.co.uk/finance/newsbysector/industry/11219688/Ten-million-jobs-at-risk-from-advancing-technology.html> [Accessed on 13 June 2019].

Turner, J. (2018). *Robot Rules*. London: Palgrave MacMillan.

UNESCO (2017). Report of COMEST on Robotics Ethics. Paris, UNESCO. <http://unesdoc.unesco.org/images/0025/002539/253952E.pdf> [Accessed on 13 June 2019].

Upchurch, T. (2018). "To Work for Society, Data Scientists Need a Hippocratic Oath With Teeth." *Wired*, April 8th. <http://www.wired.co.uk/article/data-ai-ethics-hippocratic-oath-cathy-o-neil-weapons-of-math-destruction> [Accessed on 13 June 2019].

Van Parijs, P. and Vanderborght, Y. (2017). *Basic Income: A Radical proposal for a Free Society and a Sane Economy*. Cambridge, MA: Harvard University Press.

Vayena, E. and Tasioulas, J. (2016). "The Dynamics of Big Data and Human Rights: The Case of Scientific Research." *Philosophical Transactions of the Royal Society A* 28: 374.

Wachter, S. Mittelstadt, B., and Floridi, L. (2017). "Why a Right to an Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7: 76-99.

Wiggins, D. (2016). "Sameness, Substance, and the Human Person." *Continuants: Their Activity, Their Being and Their Identity—Twelve Essays*. Oxford: Oxford Unversity Press.

Williams, B. (2006). "The Human Prejudice." *Philosophy as a Humanistic Discipline*. Princeton: Princeton University Press.