Ethics and Information Technology (2018) 20:15–26
https://doi.org/10.1007/s10676-018-9448-6

ORIGINAL PAPER

Patiency is not a virtue: the design of intelligent systems and systems of ethics

Joanna J. Bryson[1]

Published online: 16 February 2018

Abstract
The question of whether AI systems such as robots can or should be afforded moral agency or patiency is not one amenable either to discovery or simple reasoning, because we as societies constantly reconstruct our artefacts, including our ethical systems. Consequently, the place of AI systems in society is a matter of normative, not descriptive ethics. Here I start from a functionalist assumption, that ethics is the set of behaviour that maintains a society. This assumption allows me to exploit the theoretical biology of sociality and autonomy to explain our moral intuitions. From this grounding I extend to consider possible ethics for maintaining either human- or of artefact-centred societies. I conclude that while constructing AI systems as either moral agents or patients is possible, neither is desirable. In particular, I argue that we are unlikely to construct a coherent ethics in which it it is ethical to afford AI moral subjectivity. We are therefore obliged not to build AI we are obliged to.

Keywords Moral patiency · Moral agency · Ethics · Systems artificial intelligence · Strong AI

Introduction

The questions of robot or AI Ethics are difficult to resolve not because of the nature of intelligent technology, but because of the nature of Ethics. As with all normative considerations, AI ethics requires that we decide what "really" matters—our most fundamental priorities. Are we more obliged to our biological kin or to those with whom we share ideas? Do we value more the preservation of culture or the generation of new ideas? Unfortunately, asking "what really matters" is like asking "what happened before time": it sounds at first pass like a good question, but in fact makes a logical error. *Before* is not defined outside of the context of time. Similarly, we cannot circuitously assume that a system of values underlies our system of values. Consequently, the "correct" place for robots and other intelligent artefacts in human society cannot be resolved from first principles or purely by reason. It is not a fact that can be established through science.

In this article I argue that the core of all ethics is a negotiated or discovered equilibrium that creates and perpetuates a society. The descriptive argument of this article is that integrating a new capacity like artificial intelligence (AI) into our moral systems is an act of normative, not descriptive, ethics. Contrary to the claims of much previous philosophy of AI ethics (see e.g. Gunkel and Bryson 2014), there is no necessary or predetermined position for AI in our society. This is because both AI and ethical frameworks are artefacts *of* our societies, and therefore subject to human control.

Ethics has both descriptive and normative components (Fischer 2004). Descriptive ethics is based on the observation of what people seem to do—as such it is related to science and open to measurement and at least an assertion of facts. Normative ethics consist of recommendations concerning what should be done. Though these recommendations may be backed by descriptive facts, for example concerning likely consequences, by their nature they are not themselves facts. Indeed unlike science which is contingent on a single reality, normative ethics is contingent on not so much a present society as one that is desired to be achieved. Thus an account of normative recommendations should include also an account of the societal outcomes that are intended with its successful implementation.

This article contains both descriptive and normative components. The descriptive components hold independent of the normative recommendations, but the normative recommendations are entirely dependent on the descriptive content. To

✉ Joanna J. Bryson
jjb@alum.mit.edu

[1] University of Bath, Bath BA2 7AY, UK

be very clear from the outset, the moral question I address here is not whether it is possible for robots or other artefacts to be moral entities. Human culture can and does support a wide variety of moral systems. Many of these already attribute patiency to artefacts such as particular books, flags, or concepts. For example Islam considers that any copy of the Koran should be treated with respect, the United States has similar laws for its flag, and many wars are justified as defence of abstractions such as liberty. The more interesting and important question is normative—should we as technology, legal, or ethical experts recommend putting *intelligent* artefacts in that position? If so, who or what would benefit? Again, the moral status of robots and other AI systems is a choice, not a necessity. We can choose the types and properties of artefacts that are legal to manufacture and sell, and we can write the legislation that determines the legal rights and duties of any agent capable of knowing those rights and carrying out those duties.

My primary normative argument is that making robots such that they deserve to be moral patients could in itself be construed as an immoral action, particularly given that it is obviously avoidable since it is in fact a choice. To examine this we must consider not only human society, but also make potential moral-subject robots into second-order moral patients. I claim that it would be unethical to put artefacts in a situation of competition with us, to make them suffer, or to make them unnecessarily mortal. I do not claim that it is wrong to use machine intelligence to create—that is, to produce human culture. Nor do I claim that AI cannot take other moral actions. I do claim that there are many human values for which it is incoherent to think of them as extended through AI proxies in absence of a core of human moral agents. Therefore there are substantial costs but little or no benefits from the perspective of either humans or robots to ascribing and implementing either agency or patiency to intelligent artefacts beyond that ordinarily ascribed to any possession. The responsibility for any moral action taken by an artefact should therefore be attributed to its owner or operator, or in case of malfunctions to its manufacturer, just as with conventional artefacts (Bryson and Kime 2011).

In the next section I define terms for use at least local to this article, and also establish the justification I use to recommend one normative position over another. Next I examine the origins of our present ethical systems and intuitions. I discuss sociality as it existed before the concept of morality was innovated, then move on to discussing our explicit system of ethics.

Definitions and approach

Definitions and perspective

The usages of terms local to this article are chosen to attempt clear communication despite introducing a perspective many people find radical. I believe that incoherence has been introduced to AI and robot ethics debates partly because some terms are made to do "double duty." For example, *conscious* and *intelligent* have fairly clear psychological and even computational meanings, but as a confound are often assumed to be core to moral obligation. The usages I exploit here are intended as strict subsets of ordinary language usage—that is, I attempt local to this article to focus these terms' usage down to one essential meaning. I intend to use here only the formal philosophical term *moral agent* to mean "something deemed responsible by a society for its actions," only *moral patient* to mean "something a society deems itself responsible for preserving the well being of," and *moral subjects* to be all moral agents or moral patients recognised by a society (cf. Gray and Wegner 2009; Karlsson 2002; Kant 1785; Duranti 2004, p. 454). Note that *moral* agency is a distinct special case of agency more generally. *Agency* is just the capacity to effect change, for example chemical agents cause reactions. In itself the term implies no moral status.

When I talk about something as *intelligent*, I mean only that it can detect contexts appropriate for expressing one of an available suite of actions. If I call something *cognitive* then it is both intelligent and can learn new contexts, actions, and/or associations between these (cf. Winston 1984; Carlson 1994). If I call something *conscious* I only imply that it can act on explicit memory—memory of individual events (Bryson 2012). Local to this paper, these are treated as psychological terms that neither require moral status nor necessarily in themselves imply any other human-like trait. Using these extremely specific definitions, we can now meaningfully discuss questions such as "Does being intelligent necessarily make one a moral agent?" This is not to neglect the many other definitions of these terms or to say that they are 'wrong'—any word's meaning is exactly how it is used (Caliskan et al. 2017). The purpose of these definitions is to label and advance a set of concepts specifically rallied to answer the questions of whether or when AI should be considered as—or constructed to be—a moral patient.

Ethics will mean here the entire set of behaviours that maintains a society, including by defining it. This is the most radical departure from convention presented in this section. I will discuss its ramifications below, but first I will clarify. By this definition, ethics does include components that have clear universal utility, e.g. prohibition on murder or theft. But each individual society's ethical system will also consist of components dedicated to creating its unique identity, or carving out an ecological or economic niche. Identity may seem less essential than universal morality, but determining the boundaries of a society may be essential to its persistence. Identity is linked to *autonomy*, the capacity of an entity to be distinguished. For example a nation is not autonomous if it cannot control its borders (Armstrong and Read 1995; Cooke 1999). The way I am defining ethics here,

the interaction between individual and societal autonomy is key. Members of a society will often produce a system of *public goods*—that is, goods that benefit all members, such as security or transport. Public goods may be key to a society's autonomy, and may even define a society (Bryson et al. 2014). For *altruism* I will use a standard definition from biology and economics: the willingness to pay a cost (such as the cost of constructing public goods) in order to benefit others. Such willingness is adaptive in the biological sense (meaning it can be promoted by evolution) if the sum of the benefits to the beneficiaries times the relatedness of the beneficiaries to the altruist is less than or equal to the cost to the altruist ($\text{cost}_i < \sum_j \text{benefit}_j \times \text{relatedness}_{ij}$, Hamilton's Law, Hamilton 1964).

One critique of this definition of ethics is that normally we would like to believe in a single ethical standard against which societies are able to improve. My definition does undermine this exact formulation, but does not mean we cannot make moral comparisons between societies, only that we have to specify a metric for any specific such comparison. So for example eliminating slavery or expanding suffrage results in a more egalitarian society or a more coherent set of laws, not just a "more ethical" society. The Nazis were not unethical, they had an elaborate system of ethics that while briefly facilitating rapid expansion of power, also killed many people (including co-nationals), and attracted its own destruction. Again, I am not claiming my definitions are necessarily the right or common ways to use these words, but that they are precise and useful for the present article's discussion.

A key aspect of my normative argument hinges also on the definition of *artefact*. Again, local only to this article, I limit the term *artefact* to objects deliberately created by moral agents. This means that things referred to as 'artefacts' here are also specific to the society that deigns an agent moral, and therefore only occur in species that have established concepts and norms concerning responsibility and deliberation. Artefacts are *for this article* discontinuous from nature; the point at which a culture develops the concepts *deliberate* and *responsible* is when this discontinuity occurs and artefacts can begin to be made.

Approach to normative recommendations

As stated earlier, any normative recommendation requires specifying the intended societal outcome. My recommendation is not simply derived from descriptive precedent. The advent of potentially-autonomous decision-making human artefacts is novel, and requires constructing new social and ethical accommodation. Descriptive ethics may take us some way by establishing precedent, but few consider precedent sufficient or even necessary for establishing what is right. *Is* does not imply *ought* (Hume 1739).

I will work then from two fairly familiar, hopefully uncontroversial objectives:

1. The moral system should be coherent. This derives from the same principle as that unenforceable laws are not useful (McNeilly 1968). *Ought* is generally held to imply *can* (Stern 2004).
2. Where possible there should be minimal restructuring of existing norms, so that introducing new norms will be less likely to create social disruption or medium-to-long-term instability. This axiom is based on the example of Common Law (Mahoney 2001), but is admittedly less definitive than the first. It also kicks a can down the road, by allowing me to propose a descriptive criterion after all for my candidate metric, provided only that it doesn't conflict with the first objective.

The nature of machines as artefacts means that the question of their morality is not simply a question of what moral status they deserve (Miller 2015). Rather, at the same time that we ask what moral status we ought to assign intelligent artefacts, we must also ask what moral status we ought to build those artefacts to meet. This second aspect of our concurrent, tightly-coupled responsibilities has been neglected even by those scholars who have observed the constructive nature of the first (e.g. Coeckelbergh 2010; Gunkel 2014). As I said, *ought* does require *able*—computationally intractable and indeed logically incoherent systems such as Asimov's laws are excluded (Myers 2010; Bryson 2017). So are ecologically unsustainable objectives.

Our capacity to design an artefact defines the term, which means that obligations regarding intelligent artefacts, unlike those regarding natural entities, can be met not only through constructing the socio-ethical system but also through specification of the intelligent artefacts themselves. This fact defies the intuition of many who cannot conceive of intelligence in non-human contexts, or who conceive of it on a single, Lamarckian scale that converges to human-like. The historical correlation of language, episodic memory, and reasoning with the prototypical moral subjects—human adults—is taken as necessary or even causal, as if there were particular badges or features of human moral status that could be excised from our gestalt and still deserve the same treatment as a citizen of our society.

In an effort both to reduce this confusion, and also to consider what would be minimally disruptive per the second objective just mentioned, the next section of this article discusses not what *should* matter to us, but why some things *do*. Before considering where we might want to slot robots into our contemporary ethical frameworks and society, I start by considering ethics and moral patiency from an evolutionary perspective. I do this less to inform our intuitions than to explain them.

Substantive claims concerning life and intelligence

Stability in action selection

As with all human and other ape (Whiten and van Schaik 2007) behaviour, our ethics is rooted both in our biology and our culture. Nature is a scruffy designer with no motivation or capacity to cleanly discriminate between these two sources of behaviour, except that what must change more quickly should be represented more plasticly (Depew 2003; Hinton and Nowlan 1987). As human cultural evolution has accelerated our societies' paces of change, increasingly our ethical norms are represented in highly plastic forms such as legislation and policy (Ostas 2001).

The problem with a system of action selection so extremely plastic as explicit decision making is that it can be subject to *dithering* —switching from one goal to another so rapidly that little or no progress is made on either (Humphrys 1996; Rohlfshagen and Bryson 2010). Dithering is a problem potentially faced by any autonomous actor with multiple goals that at least partially conflict and must be maintained concurrently. Conflict is often resource-based, for example visually attending to two children at one time, or needing to both sleep and work. An example of dithering in early computers was *thrashing*—a process of alternating between two programs on a single CPU where each required access to the majority of main memory. Poor system design could result in an operating system allocating a slice of time to each process shorter than the time it took to be read into main memory from disk, preventing either program from achieving any of its real functions. More generally, dithering implies changing goals—or even optimising processes—so frequently that more time is wasted in the transition than is gained in accomplishment.

Perhaps to avoid dithering, we as humans prefer to regulate social behaviour even in an extremely dynamic present by planting norms in a "permanent," bedrock past, like the anchoring of tall buildings built over a swamp. For example, American law is often debated in the context of the US constitution, despite being rooted in British Common Law and therefore a constantly changing set of precedents. Ethics is often debated in the context of holy ancient texts, even when the ethical questions at hand concern contemporary matters such as abortion or robots about which there is no reference or consideration in the original documents. Societies tend to believe that basic principles are rational, fixed, and universal. Enormous changes in social order such as universal suffrage or the end of legalised human slavery are simply viewed as corrections, bringing about the originally-intended rather than a newly-improved (or worse, locally-convenient) order.

In fact our ethical structures and morality do co-evolve with our society (Waal 1996). When the value of human life relative to other resources was lower, murder was more frequent and less sanctioned, and political empowerment was less widely distributed (Johnson and Monkkonen 1996; Pinker 2012). When women can support themselves and their children independently, infidelity is viewed less harshly (Price et al. 2014). What it means to be human changes, and our ethical systems have to accommodate that change.

Fundamental social behaviour

As I implied when defining *ethics*, an ethical systems will contain components addressing two problems:

1. Defining a society—discriminating it from others, and
2. Maintaining a society internally.

The first problem may underpin our psychological obsession with ingroup-outgroup dynamics. I have suggested elsewhere that a society may be defined by the public goods it creates and defends, thus the scale of a coherent economy may limit the size of a society (Bryson et al. 2014, cf. Powers et al. 2011). The second problem could however at least in theory be universal, and as such could also be a candidate for describing how AI might become a moral subject. Maintaining a society internally is also the topic of the rest of this section.

I begin by considering the most basic component of social behaviour: whether that behaviour is for or against society—pro- or anti-social. Assessing morality is not trivial, even for apparently trivial, 'robotic' behaviour of single cell organisms, which also behave pro- and anti-socially. For example MacLean et al. (2010) demonstrate the overall social utility of organisms behaving in a way that at first assessment seems to be obviously anti-social—free riding off of pro-social agents that manufacture costly public goods. Single-cell organisms produce a wide array of shared goods ranging from shelter to instructions for combating antibiotics (Rankin et al. 2010). MacLean et al. (2010) focus on the production of digestive enzymes by the more 'altruistic' of two isogenic yeast strains. Having no stomachs, yeast must excrete such enzymes outside of their bodies. The production of these enzymes is costly, requiring difficult-to-construct proteins, and the production of pre-digested food is beneficial not only to the excreting yeast but also to any other yeast in its vicinity. The production of these enzymes thus meets the common anthropological and economic definition of *altruism*: paying a cost to express behaviour that benefits others (Fehr and Gächter 2000).

In the case of single-cell organisms there is no 'choice' as to whether to be free-riding or pro-social. This is genetically determined by their strain, but the two sorts of behaviour are accessible from each other during reproduction (the construction of new individuals) via common mutations

(Kitano 2004; Youk and Lim 2014). For such systems, natural selection performs the 'action selection' between goals by determining what proportion of which strategy lives and dies. What MacLean et al. (2010) show is that selection can operate such that the lineage as a whole benefits from mixing both strategies (cf. Akçay and Cleve 2016). The 'altruistic' strain in fact *overproduces* the public good (the digestive enzymes) at a level that would be wasteful, while the 'free-riding' strain of course underproduces. Thus the greatest good—the most efficient exploitation of the available resources—is achieved by the species as a whole.

Why can't the altruistic strain evolve to produce the right level of public goods? This returns to my earlier point about rates of plasticity. The optimal amount of enzyme production is determined by available food and this will change more quickly than the physical mechanism for enzyme production in a single strain could evolve. However death and birth can be fast and cheap in single-cell organisms. A mixed population composed of multiple strategies, where the high and low producers will always over and under produce respectively, and where their proportions can be changed very rapidly, is thus an agile solution. Thus the greater good for the species is served by the 'selfishness' of many of its members, but would not be so served without the presence of altruists.

Human society also appears to up *and* down regulate investment in public goods (Bryson et al. 2014). We may increase production of public goods by calling their creation 'good', and associating 'good' with a social status that is beneficial in the socio-economic contexts where more public goods are beneficial. Meanwhile, self interest and individual learning from direct reenforcement can be relied on to motivate and maintain the countervailing population of underproducers. For human society too the 'correct' amount of investment may vary quickly due to shifts in socio-economic and political context. For example, national military investment may be worthwhile under threat of invasion, but investment in local businesses may be more advantageous at other times. This implies that the *reduction* of other's 'good' behaviour can itself be of public utility in times when society benefits from more individual productivity or self-sufficiency (cf. Trivers 1971; Rosas 2012). If so, we would expect that in such contexts it may also be easier for human institutions to change their overall assessment of which public goods require investment than to change their exact rate of output for all individuals (Bryson et al. 2014).

Is does not imply *ought*. The roots of our ethics do not entirely determine where we should or will progress. But roots do affect our intuitions. Our intuitions towards inclusion of artefacts in our society are probably driven by the extent to which we identify with such artefacts (Bryson and Kime 2011). This goes back to the biological account for altruism given in the definitions section: we are by nature willing to pay a higher cost for those more related to us. For humans, this 'relatedness' seems to extend also to those whose ideas we share (Plotkin 1995; Gardner and West 2014). This would allow us to be a phenomenally agile species, rapidly generating new societies to exploit available opportunities, particularly if (as seems to be true, Coman et al. 2014) we can prompt each other to focus on particular identities in particular circumstances.

Others have proposed using our intuitions as a mechanism for determining our obligations with respect to robots and AI (Dennett 1987; Brooks 2002; Prescott 2017). Because of their origins in our evolutionary past, and the simple observation of how patiency can be attributed to plush toys (Bryson and Kime 2011), I do not trust this strategy to create coherent ethics. I do however trust those with vested interests—such as interests in selling weapons, robots, or even books—to exploit such intuitions (Bryson 2010; Bryson et al. 2017). Although established precedent is close to my second objective proposed earlier for the justification we seek for a normative recommendation, I consider picking a precedent (in-group identification) that divides as much as it unites to be unsatisfactory. Such divisions seem particularly dated given that we can expect communication technology to increase the potential size of our social group (Roughgarden et al. 2006; Bryson 2015). In the next section I turn as an alternative established source of criteria for making a normative recommendation to philosophy, which I exploit in the sections following to propose a more coherent, minimally disruptive path to situating AI in our society, and (therefore) our ethics.

Normative claims concerning robots and AI

Freedom and morality

"[Moral] action is an exercise of freedom and freedom is what makes morality possible."—Johnson (2006). For millennia morality has been recognised as something uniquely human, and therefore taken as an indication of human uniqueness and even divinity (Forest 2009). But if we throw away a supernaturalist and dualistic understanding of human mind and origins, we can still maintain that human morality is at least rooted in the one incontrovertible aspect of human uniqueness—language—and our unsurpassed competence for cultural accumulation that language both exemplifies and further enables (Bryson 2008). The cultural accumulation of new concepts gives us more ideas and choices to reason over, and our accumulation of tools gives us as individuals more power to derive substantial changes to our environment from our intentions.

Some of these tools include social concepts that have proved useful fulcrums for the leverage we need to construct our complex societies. These include 'self', 'society',

'justice', 'responsibility', 'freedom', and 'intention'. As asserted earlier in the section on definitions, it is at the point of the invention of these concepts that we can discriminate artificial from natural intelligence. An artefact is something for which the design is intentional. That intention—the authorship of our action—ordinarily is seen as entailing responsibility, even in the face of determinism (Fischer 1999).

If human morality depended simply on human language then our increasingly language-capable machines would be excellent candidate moral subjects. But I believe that freedom—which I take here to mean *the socially-recognised capacity to exercise choice* is the essential property of a moral actor (cf. Tonkens 2009; Rosas 2012). Dennett (2003) argues that human freedom is a consequence of evolving complexity beyond our own capacity to provide a better account for our behaviour than to attribute it to our own individual responsibility. This argument entails a wide variety of interesting—and not necessarily desirable—consequences. For example, as our science develops and our behaviour becomes more explicable via other means (e.g. insanity) fewer actions might be taken as moral. This principle might also be seen to encourage the irresponsible construction of opaque institutions or obfuscated source code for robots or other AI systems in order to avoid individual or institutional responsibility (Siponen 2004; Bryson et al. 2017).

I will nevertheless here be conservative and follow from Dennett's suggestion to generalise morality beyond human ethics. Again only local to this article, I define moral actions *for an individual agent* to be those for which:

1. A particular behavioural context affords more than one possible action for that agent,
2. At least one available action is considered *by a society* to be more socially beneficial than the other options, and
3. The agent is able to recognise which action is socially beneficial—or at least socially sanctioned—and act on this information.

Note that this definition captures society-specific morals as well as the individual's role as the actor.

With this definition I again reach to the biological heritage of our present ethics by deliberately extending morality to include actions by other species which may be sanctioned by *their* society, or by ours. For example, non-human primates will punish individuals that violate their social norms, e.g. for being excessively brutal in punishing a subordinate (Waal 2007), for failing to vocally 'report' available food (Hauser 1992), or for sneaking copulation (Byrne and Whiten 1988).[1] Similarly, this definition allows us to say dogs and even cats can be good or bad when they obey or disobey human social norms they have been trained to recognise, provided they have demonstrated a capacity to select between relevant alternative behaviours, and particularly when they behave as if they expect social sanction when they select the proscribed option. I make this inclusive reach to prepare for a consideration of ethics from the perspective of a society of artefacts.

With respect to AI, there is no question that we can train or simply program machines to recognise more or less socially-acceptable actions, and to use that information to inform action selection (Cakmak et al. 2010; Riedl and Harrison 2016). So we can certainly build AI to take moral actions. But this in itself does not determine moral agency. The question is, who would be responsible for those actions? An agent that takes a moral action is not necessarily the moral agent—not necessarily the or even a locus of responsibility for that action. A robot, a child, a pet, even a plant or the wind might be an agent that alters some aspect of an environment. Children, pets, and robots may know they could have done 'better.' We can expect the assignment of responsibility for moral acts by intellgent artefacts to be similarly subject to debate and variation. Moral responsibility is only attributed to those a moral community has recognised as being in a position of responsibility. Households may differ in their assignment of culpability to children and pets, so may civilisations. Presently in the OECD at least, small children and pets are certainly not considered legally responsible for their actions.

My recommendation for AI (below) will be similar. However, the core observation here is that a moral community defines itself and its moral agents—not by simple assertion of an individual, but by consensus of the society formed. A growing child will demand agency, and as they grow these demands generally become both more costly to deny and safer to accede. However, children by our nature become adults, the components of our societies. We tend to build AI systems in their final form—as search engines, surveillance cameras, spell checkers, automobiles, or whatever product we are marketing. Should we produce a product to be a

[1] While reports of social sanctions of such behaviour are often referred to as 'anecdotal' they are common knowledge for anyone working with primates. I personally, despite having been forewarned, was once forced to violate a Capuchin monkey norm: *possession is ownership*. I was sanctioned (barked at) by the entire colony—not only those who observed the affront directly, but all those in hearing range of those observers.

moral agent? Can our society sustain itself if responsibility is delegated to entities that can be specified, built, bought, and sold?

Artefacts and responsibility

We have now reached the heart of machine responsibility. This is the point at which there is no simple descriptive solution, but rather we are looking to establish norms and laws that will lead to an ethics that is both sound and stable. We could designate to intelligent artefacts any position of responsibility we choose, and indeed several such positions are presently being considered by various legal systems (Bryson et al. 2017). To motivate my normative recommendations, allow me to ask a relevant question: Would it be moral for us to construct a machine that would of its own volition choose any but the most moral action?

This is a trick question, the key to which returns to the definition of freedom I took from Dennett. For it to be rational for us to describe an action by a machine to be "of its own volition", we must already have sufficiently obfuscated its decision-making process such that we cannot otherwise predict its behaviour, and thus be reduced to applying sanctions to it in order for it to learn to behave in a way that our society prefers. Otherwise, if the machine acted as we intended, the responsibility would be ours just as if we had performed the action with any other tool.

Note that if we assume as I've asserted that *ought* implies *can*, there's an issue of whether we can coherently and ethically produce AI that suffers from sanctions. I will return to this point below, but first I wish to discuss obfuscation. I do not consider training action selection via deep learning, reinforcement learning or any other statistical technique to be necessarily obfuscating in this sense. Even if we do not know the exact 'meaning' of every individual components of an internal representation, the basic principles of optimisation that underlie machine learning are well-understood and the probable outcomes known to in my mind be sufficient for moral clarity (see e.g.Wilson et al. 2016). Further, there are basic and well-established procedures for creating test suites and exploring responses to input to ensure a system incorporating machine learning or any other sort of programming, or even human judgement, is performing to a standard that we approve (Jones 2008; Chessell and Smith 2013; Dwork et al. 2012; Feldman et al. 2015). Similarly, I do not consider the fact that unexpected effects 'emerge' during the operation of complex systems to alter the designers' responsibility to observe and account for such effects. Neither do present courts of law (Bryson et al. 2017).

As I asserted earlier, the step change from nature to artefact is our intentional acts of creation, for which we as moral agents are almost by definition and certainly by convention considered responsible. When executing an intentional action, deliberately blinding oneself to an outcome is not ordinarily seen as ending responsibility; rather it is termed *wilful negligence*. The same should hold true for not following adequate procedures to ensure transparency in the construction of intelligent artefacts (Wortham and Theodorou 2017; Bryson and Winfield 2017). Since we have perfect control over when and how a robot is created, we also have responsibility for it. Assigning responsibility to the artefact for actions we designed it to execute would be to deliberately disavow our responsibility for that design. Currently, even where we have imperfect control over something as in the case of young children, owned animals, and operated machinery, causing harm by losing control entails at least some level of responsibility to the moral agent, the legal person. If you deliberately drive into someone you commit murder, if you do it accidentally you commit manslaughter. You are responsible but the sanctions are reduced.

Why, for example, are we responsible for intelligent beings such as children, but only up to a fixed age? Because society has found this to be the best method for organising itself, though there is some dispute and therefore variability about the exact point at which the child becomes responsible. Thus legal persons are responsible for maintaining control over their dogs, cars, and children, and individuals can be held accountable for negligence and manslaughter performed by those things over which control was not sufficiently maintained (Lia 2015). Again, these laws are not based on some pure, indisputable, formal, mathematical fact—though they can be informed by science, for example of developmental psychology. Fundamentally though they reflect the best way of maintaining our society that our society has been able to both discover and agree to enforce. Subparts of our society, for example clubs or families, may institute additional rules that maintain and hopefully enhance the lives of members of these smaller societies.

Beneficiaries of machine patiency

Why—or in what circumstances—should robots be given the moral agency we deny children? Should we be allowed to obscure our own control of the machines we make? Create and sell legal products without being responsible for understanding their behaviour? Could there even be a reason to pass off or hand on responsibility for an artefact that *has* been well-designed and is transparent to us?

Deriving normative recommendations for how we should adjust our ethical systems to encapsulate the AI we create requires reasoning about multiple levels of ethical obligation and multiple possible ethical strategies. In the yeast example I gave earlier, 'anti-social' free riding actually optimised the overall investment of a society—a spatially-local subset

of a species inhabiting a particular ecological substrate—in a way that helped it compete with other nearby species. Behaviour possibly disadvantageous very local to free riders was less-locally advantageous to the species as a whole. Similarly, it is at least possible that in times of severe economic and political crisis, turning to a competitive strategy that destroys some levels of social organisation and their associated public goods may be an effective survival strategy for what remains of the society after this destruction. The definition of morality introduced above depends then on social benefit. Could there be social benefit for any society in abdicating our responsibility as authors of our artefacts?

I will focus here on just two of the conceivable societies, the two at the extremes of the possibilities: one entirely focussed on humans and human organisations such as we at least legally inhabit now, and the other a society of independent robotic devices functioning as legally autonomous individuals in a subcommunity in a world otherwise much like the present day. The assumption is that these two worlds are technologically equivalent, but in the second we have constructed and marketed robots as legal products that are designated also as legal persons, giving them sufficient cognitive state (both memory and motivation) to self organise along the same lines as human communities, for example as a union. I will now consider briefly for each of these who benefits and who does not from designating moral agency and patiency to AI.

The perspective of human well being

Many people have suggested after Kant that failing to treat something that appears to us to be human as if it were human would be a moral wrong *towards other humans*, because it encourages our propensity to dehumanise (Gunkel 2017 for a recent review). While it would be both foolish and unnecessary to argue against Kant, what these arguments overlook is that there are two ways to address this problem—either by treating artefacts as human, or by making their inhumanity transparent (Theodorou et al. 2017). The only extant national-level robot ethics policies recommends the latter (Boden et al. 2011), as do I. Another possible benefit of robot moral subjects is that, for some of us, it gives us pleasure or feeds our egos to construct objects that we owe moral status (Helmreich 1997; Bringsjord et al. 2012). Some of us also project ourselves into our creations, and see AI as a route to immortality more perfect than other forms of procreation (Goertzel 2010). Of course, that perceived identity is demonstrably false (Choe et al. 2012; Claxton 2015).

To me the only persuasive argument is that it is possible that in the long term treating intelligent artefacts as moral agents would be a simpler way to control an opaquely-complex intelligence, and that the benefits of that opaquely complex intelligence might outweigh the costs of losing some of our own moral responsibility and therefore moral status. However this necessity has yet to be demonstrated. Given the costs of abdicating responsibility, we should not abdicate based on speculation (Bryson et al. 2017).

Some are currently arguing that enforcing good practice in transparency and accountability for AI may slow the rate of progress. Empirically, it often turns out that designing a system to be transparent makes it easier to maintain and extend, so any such penalties even if they exist may only do so in the short term (Theodorou et al. 2017; Zeng et al. 2017). But even if the penalties of transparency prove real and long-lasting, we need to consider the benefits of any accelerated progress against the costs of losing human responsibility, costs which may include losing social cohesion. The principal such cost I see is the facilitation of the unnecessary abrogation of responsibility by sellers or operators of AI. For example, customers could be fooled into wasting resources needed by their children or parents on a robot, or citizens could be fooled into blaming a robot rather than a politician for unnecessary fatalities in warfare (Sharkey and Sharkey 2010; Bryson and Kime 2011; Bryson 2000). A corporation could displace responsibility for its decision to use automation rather than human employment onto the automation itself, creating a legal lacuna—a set of far poorer, purely-synthetic entities set up to be held responsible for tax and legal liability (Bryson et al. 2017). If such an entity went bankrupt or were jailed, it would disuade no one into changing their or its behaviour.

The more I study the legal aspects of this problem, the more convinced I am that academics facilitating the fantasies of trans- and post-humanists (Geraci 2010) run the risk of encouraging political and economic decisions that could seriously disrupt our ability to govern, as well as our economy. The artificial entities likely to be most advantaged are transnational corporations, and there are few effective mechanisms of government at the transnational level. Without a capacity to govern our economy we lose the collective ability to encourage arts and innovation beneficial to all. With unaccountable power and economic inequality, the majority of humanity feels and may indeed become too at-risk to innovate and contribute to their local economy and public goods. Empirically, this leads to social disruption (Atkinson 2015).

The perspective of AI well being

Although this argument has been overlooked by some critics (notably, Gunkel 2012; Prescott 2017), the policies I promote (e.g. Bryson 2010, 2009) have always explicitly considered the welfare of potential intelligent artefacts. I cast these as second-order moral patients. Why should we design artefacts to be in the position of competing with us for resources; of longing for higher social status (as all

evolved social vertebrates do); of fearing injury, extinction, or humiliation? We are able to ensure that AI is properly and continuously backed up. We can and do build it to have no concern for social status, nor sense of purpose.

In short, we can afford to stay agnostic about whether an artefact can have qualia, because we can avoid constructing motivation systems encompassing suffering. We know we can do this because we already have. There are many proactive AI systems now, and none of them suffer. There are already machines that play go, chess, checkers, do arithmetic, refrigerate, and clean clothes better than we do, but none of these aspires to world domination. We can limit AI—or at least legally-produced commercial AI—to be as it is now, something to which no obligations are owed directly. There can therefore be no ethical costs to the AI of maintaining AI in its present status of non-suffering, unless we postulate rights of the 'unbuilt'.

Tonkens (2009) makes a very similar point to mine concerning AI well being, which Rosas (2012) disputes. I believe the root of the conflict here is that Rosas believes morality must be rooted in social dominance structures. The definition of morality I introduced in the previous section eliminates this confound. For evolved intelligence, dominance structure may be an inevitable part of the selective process; therefore we may expect the dysphoric aspects of subjugation may also be universal in evolved beings. Certainly therefore human ethical systems, as a part of social regulation, have much to say concerning dominance. But in designed artefacts we can safely eliminate this dysphoric aspect of subservience. Even negative self assessment by a robot has no need to lead to self harm or degradation, just restraint in risk taking and a request for repairs.

Recommendations

In the Introduction I suggested two criteria for ethical systems: coherence, and a lack of social disruption. I can think of no coherent reason to create agents with which we should compete. Every value we have, from æsthetics to peace to winning, comes from our evolutionary origins as apes. What coherent reason can we have to 'pass the baton' to machines made to share and compete on the basis of these goals? Even if we take the technologically-dubious case of machine immortality, what would we be making immortal? Any self-learning technological agent would rapidly evolve preferences that suit its machine nature, not ours. Would an initially-human-like capacity for computation be worth sacrificing human potential for in order to create something eventually as similar to us as crabgrass (Moore 1947)? Returning to Hamilton's Law, the answer might be "yes" if we could assume that our technology was likely to survive our species or even our civilisation, but to date digital technology formats tend to survive less than $\frac{1}{16}$ as long as individual humans.

Bryson et al. (2002) argue that the right way to think about intelligent services (there in the context of the Internet, but here I will generalise) is as extensions of our own motivational systems. We are currently the principal agents when it comes to our own technology, and I believe it is our ethical obligation to design both our AI and our legal and moral systems to maintain that situation. Legally and ethically, AI works best as a sort of behavioural prosthetic to our own needs and desires. If we wish to extend the lifespan of our civilisation, I recommend focussing on ways to do this while maintaining a flourishing human society at the motive core.

As mentioned above, one of the best arguments I know against this human-based perspective is that maltreating something that reminds us of a human might lead us to treat other humans or animals worse as well (Parthemore and Whitby 2014). The United Kingdom's *EPSRC Principles of Robotics* specifically address this problem in its fourth principle, and in two ways (Boden et al. 2011; Bryson 2017). First, robots should not have deceptive appearance—they should not fool people into thinking they are similar to empathy-deserving moral patients. Second, their AI workings should be 'transparent' (Wortham and Theodorou 2017). This implies that clear, generally-comprehensible descriptions of an artefact's goals and intelligence should be available to any owner, operator, or other concerned party. This Principle was adopted despite considerable concerns about the requirement for both therapeutic and simple commercial/entertainment robots to masquerade as moral patients and companions (cf. Miller et al. 2015). Because of this consideration, the fourth Principle deliberately makes transparency *available* for informed long-term decisions, but not constantly *apparent*. The goal is that most healthy adult citizens should be able to make correctly-informed decisions about emotional and financial investment. As with fictional characters and plush toys (Ullán et al. 2014), we should be able to both experience beneficial emotional engagement, *and* to maintain explicit knowledge of an artefact's lack of moral subjectivity.

One thread of theory for the construction of human-level AI[2] holds that it may be impossible to create the sort of intelligence we want or need unless we completely follow the existing biologically-inspired templates which therefore

[2] Please note that all arguments in this paper apply to all AI, including Artificial General Intelligence (AGI). AGI was originally a pejorative meant to imply that AI had 'failed' because the discipline was not pursuing the correct goals. Now it has come to mean two contradictory things—a system that can be all knowing, and a system that is human-like. There is no chance of the combinatorial complexity of all possible knowledge or planning being overcome such that there can be an omniscient AI; this is a computational impossibility. The problems of making an artefact completely human-like I have dealt with earlier in the main text.

must include social striving, pain, etc. So far there is no evidence for this position, in fact we are persistently creating super-human AI capacities without these attributes (Bryson 2015; Bryson and Winfield 2017). But if it is ever demonstrated, even then we would not be in the position where our hand was forced—that we *must* permit patiency and agency. Rather, we will then, and only then, have enough information to stop, take council, and produce a literature and eventually legislation, regulation, and social norms on what is the appropriate amount of moral subjectivity to permit given the benefits it would provide.

Conclusion

As Johnson (2006, p. 201) puts it "Computer systems and other artefacts have intentionality—the intentionality put into them by the intentional acts of their designers." It is unquestionably within our society's capacity to define robots and other AI as moral agents and patients. In fact, many authors (both philosophers and technologists) are currently working on this project. It may be technically possible to create AI systems that would meet contemporary requirements for moral agency or patiency. But even if it were possible, neither of these two statements makes it either necessary or desirable that we should do so. Both our ethical systems and our artefacts are amenable to human design.

The primary, and descriptive argument of this article is that making AI moral agents or patients is an intentional and avoidable action. The secondary, normative argument which is definitionally still open to debate, is that avoidance would be our most ethical choice.

Acknowledgements I would like to thank everyone who has argued with me about the above, but particularly David Gunkel and Will Lowe. Earlier versions of this paper have appeared in Gunkel et al. (2012) and Stojanov (2016), thanks to the participants of those meetings for discussion. Thanks also to Virginia Dignum, Bipin Indurkhya, Tom Grant, and Mihailis Diamantis, Robert Sparrow, the members of CITEC at Bielefeld especially Helge Ritter, and the excellent anonymous reviewers of this journal.

Funding Funding was provided by the AXA Research Fund. Gold open access was provided at a cost to the University of Bath.

References

Akçay, E., & Van Cleve, J. (2016). There is no fitness but fitness, and the lineage is its bearer. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *371*(1687), 20150085.

Armstrong, H., & Read, R. (1995). Western European micro-states and EU autonomous regions: The advantages of size and sovereignty. *World Development*, *23*(7), 1229–1245.

Atkinson, A. B. (2015). *Inequality: What can be done?* Cambridge, MA: Harvard University.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorell, T., Wallis, M., Whitby, B., & Winfield, A. (2011). Principles of robotics. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC).

Bringsjord, S., Bringsjord, A., & Bello, P. (2012). Belief in the singularity is fideistic. In A. H. Eden, J. H. Moor, J. H. Sraker, & E. Steinhart (Eds.), *Singularity hypotheses: The frontiers collection* (pp. 395–412). Berlin: Springer.

Brooks, R. A. (2002). *Flesh and machines: How robots will change us*. New York: Pantheon Books.

Bryson, J. J. (2000). A proposal for the Humanoid Agent-builders League (HAL). In Barnden, J. (Ed.)., *AISB'00 Symposium on Artificial Intelligence, Ethics and (Quasi-)Human Rights*, pp. 1–6.

Bryson, J. J. (2008). Embodiment versus memetics. *Mind & Society*, *7*(1), 77–94.

Bryson, J. J. (2009). Building persons is a choice. *Erwägen Wissen Ethik*, 20(2):195–197 (commentary on Anne Foerst, *Robots and Theology*).

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63–74). Amsterdam: John Benjamins.

Bryson, J. J. (2012). A role for consciousness in action selection. *International Journal of Machine Consciousness*, *4*(2), 471–482.

Bryson, J. J. (2015). Artificial intelligence and pro-social behaviour. In C. Misselhorn (Ed.), *Collective agency and cooperation in natural and artificial systems: Explanation, implementation and simulation*, Philosophical studies (Vol. 122, pp. 281–306). Springer: Berlin.

Bryson, J. J. (2017). The meaning of the EPSRC principles of robotics. *Connection Science*, *29*(2), 130–136.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*(3), 273–291.

Bryson, J. J., & Kime, P. P. (2011). Just an artifact: Why machines are perceived as moral agents. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (pp. 1641–1646), Barcelona: Morgan Kaufmann.

Bryson, J. J., Martin, D., McIlraith, S. I., & Stein, L. A. (2002). Toward behavioral intelligence in the semantic web. *IEEE Computer*, 35(11):48–54. Special issue on *Web Intelligence*.

Bryson, J. J., Mitchell, J., Powers, S. T., & Sylwester, K. (2014). Understanding and addressing cultural variation in costly antisocial punishment. In M. A. Gibson & D. W. Lawson (Eds.), *Applied evolutionary anthropology: Darwinian approaches to contemporary world issues* (pp. 201–222). Heidelberg: Springer.

Bryson, J. J., & Winfield, A. F. T. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, *50*(5), 116–119.

Byrne, R. W., & Whiten, A. (Eds.). (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys*. Oxford: Oxford University.

Cakmak, M., Chao, C., & Thomaz, A. L. (2010). Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 108–118.
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.
Carlson, N. R. (1994). *Physiology of Behavior* (5th ed.). Boston: Allyn and Bacon.
Chessell, M., & Smith, H. C. (2013). *Patterns of information management*. London: Pearson Education.
Choe, Y., Kwon, J., & Chung, J. R. (2012). Time, consciousness, and mind uploading. *International Journal of Machine Consciousness*, *04*(01), 257–274.
Claxton, G. (2015). *Intelligence in the flesh: Why your mind needs your body much more than it thinks*. New Haven: Yale University.
Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, *12*(3), 209–221.
Coman, A., Stone, C. B., Castano, E., & Hirst, W. (2014). Justifying atrocities. *Psychological Science*, *25*(6), 1281–1285.
Cooke, M. (1999). A space of one's own: Autonomy, privacy, liberty. *Philosophy & Social Criticism*, *25*(1), 22–53.
Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT.
Dennett, D. C. (2003). *Freedom evolves*. New York: Viking.
Depew, D. J. (2003). Baldwin and his many effects. In B. H. Weber & D. J. Depew (Eds.), *Evolution and learning: The Baldwin effect reconsidered*. Cambridge, MA: MIT.
Duranti, A. (ed.). (2004). *A companion to linguistic anthropology*. Malden, MA: Blackwell.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* pp. 214–226. ACM.
Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*(4), 980–994.
Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM.
Fischer, J. (2004). Social responsibility and ethics: Clarifying the concepts. *Journal of Business Ethics*, *52*(4), 381–390.
Fischer, J. M. (1999). Recent work on moral responsibility. *Ethics*, *110*(1), 93–139.
Forest, A. (2009). Robots and theology. *Erwägen Wissen Ethik*, *20*(2), 195–197.
Gardner, A., & West, S. A. (2014). Inclusive fitness: 50 years on. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1642), 20130356.
Geraci, R. M. (2010). The popular appeal of apocalyptic AI. *Zygon*, *45*(4), 1003–1020.
Goertzel, B. (2010). AI against ageing–AIs, superflies, and the path to immortality. *Singularity Summit* (pp. 14–15). San Francisco: CA, USA.
Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*(3), 505.
Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. Cambridge, MA: MIT.
Gunkel, D. J. (2014). A vindication of the rights of machines. *Philosophy & Technology*, *27*(1), 113–132.
Gunkel, D. J. (2007). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 11*, 605.
Gunkel, D. J., & Bryson, J. J. (2014). Introduction to the special issue on machine morality: The machine as moral agent and patient. *Philosophy & Technology*, *27*(1), 5–8.
Gunkel, D. J., Bryson, J. J., & Torrance, S. (eds.). (2012). *The Machine Question: AI, Ethics and Moral Responsibility*. AISB/IACAP World Congress. The Society for the Study of Artificial Intelligence and Simulation of Behaviour, Birmingham, UK.
Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*, 1–52.
Hauser, M. D. (1992). Costs of deception: Cheaters are punished in rhesus monkeys (Macaca mulatta). *Proceedings of the National Academy of Sciences of the United States of America*, *89*(24), 12137–12139.
Helmreich, S. (1997). The spiritual in artificial life: Recombining science and religion in a computational culture medium. *Science as Culture*, *6*(3), 363–395.
Hinton, G. E., & Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems*, *1*, 495–502.
Hume, D. (1739). *A treatise of human nature*. London: John Noon.
Humphrys, M. (1996). Action selection methods using reinforcement learning. In P. Maes, M. J. Matarić, J.-A. Meyer, J. Pollack, & S. W. Wilson (Eds.), *From Animals to Animats 4 (SAB '96)*. Cambridge, MA: MIT.
Indurkhya, B., & Stojanov, G. (Eds.). (2016). *Ethical and Moral Considerations in Nonhuman Agents*. AAAI Spring Symposium Series. Stanford: AAAI Press.
Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*, 195–204. https://doi.org/10.1007/s10676-006-9111-5.
Johnson, E. A., & Monkkonen, E. H. (1996). *The civilization of crime: Violence in town and country since the Middle Ages*. Champaign, IL: University of Illinois Press.
Jones, C. (2008). *Applied software measurement: Global analysis of productivity and quality* (3rd ed.). New York: McGraw-Hill Education Group.
Kant, I. (1785). *Grundlegung zur Metaphysik der Sitten*. Leipzig: Hartknoch.
Karlsson, M. M. (2002). Agency and patiency: Back to nature? *Philosophical Explorations*, *5*(1), 59–81.
Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, *5*, 826–837.
Liao, H.-P. (2015). Stop calling my daughter's death a car accident. *Wired*.
MacLean, R. C., Fuentes-Hernandez, A., Greig, D., Hurst, L. D., & Gudelj, I. (2010). A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biology*, *8*(9), e1000486.
Mahoney, P. G. (2001). The common law and economic growth: Hayek might be right. *The Journal of Legal Studies*, *30*(2), 503–525.
McNeilly, F. S. (1968). The enforceability of law. *Noûs*, *2*(1), 47–64.
Miller, K., Wolf, M. J., & Grodzinsky, F. (2015). Behind the mask: Machine morality. *Journal of Experimental & Theoretical Artificial Intelligence*, *27*(1), 99–107.
Miller, L. F. (2015). Granting automata human rights: Challenge to a basis of full-rights privilege. *Human Rights Review*, *16*(4), 369–391.
Moore, W. (1947). *Greener than you think*. New York: Random House.
Myers, C. B. (2010). Ethical robotics and why we really fear bad robots. *TNW News*.
Ostas, D. T. (2001). Deconstructing corporate social responsibility: Insights from legal and economic theory. *American Business Law Journal*, *38*(2), 261–299.
Parthemore, J., & Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness*, *06*(02), 141–161.
Pinker, S. (2012). *The better angels of our nature: The decline of violence in history and its causes*. London: Penguin.

Plotkin, H. (1995). Non-genetic transmission of information: Candidate cognitive processes and the evolution of culture. *Behavioural Processes*, *35*(1), 207–213.

Powers, S. T., Penn, A. S., & Watson, R. A. (2011). The concurrent evolution of cooperation and the population structures that support it. *Evolution*, *65*(6), 1527–1543.

Prescott, T. J. (2017). Robots are not just tools. *Connection Science*, *29*(2), 142–149.

Price, M. E., Pound, N., & Scott, I. M. (2014). Female economic dependence and the morality of promiscuity. *Archives of Sexual Behavior*, *43*(7), 1289–1301.

Rankin, D. J., Rocha, E. P. C., & Brown, S. P. (2010). What traits are carried on mobile genetic elements, and why? *Heredity*, *106*(1), 1–10.

Riedl, M. & Harrison, B. (2016). Using stories to teach human values to artificial agents. In *AI, Ethics, and Society: Workshop at the Thirtieth AAAI Conference on Artificial Intelligence*.

Rohlfshagen, P., & Bryson, J. J. (2010). Flexible latching: A biologically-inspired mechanism for improving the management of homeostatic goals. *Cognitive Computation*, *2*(3), 230–241.

Rosas, A. (2012). The holy will of ethical machines. In Gunkel, D. J., Bryson, J. J., & Torrance, S. (Eds.). *The Machine Question: AI, Ethics and Moral Responsibility*, AISB/IACAP World Congress (pp. 29–32), Birmingham, UK. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Roughgarden, J., Oishi, M., & Akçay, E. (2006). Reproductive social behavior: Cooperative games to replace sexual selection. *Science*, *311*(5763), 965–969.

Sharkey, N., & Sharkey, A. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies*, *11*(2), 161–313.

Siponen, M. (2004). A pragmatic evaluation of the theory of information ethics. *Ethics and Information Technology*, *6*(4), 279–290.

Stern, R. (2004). Does 'ought' imply 'can'? and did Kant think it does? *Utilitas*, *16*(1), 42–61.

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, *29*(3), 230–241.

Tonkens, R. (2009). A challenge for machine ethics. *Minds and Machines*, *19*(3), 421–438.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*(1), 35–57.

Ullán, A. M., Belver, M. H., Fernández, E., Lorente, F., Badía, M., & Fernández, B. (2014). The effect of a program to promote play to reduce children's post-surgical pain: With plush toys, it hurts less. *Pain Management Nursing*, *15*(1), 273–282.

de Waal, F. (2007). *Chimpanzee politics: Power and sex among apes* (25th anniversary ed.). Baltimore, MA: Johns Hopkins University.

de Waal, F. B. M. (1996). *Good Natured: The origins of right and wrong in humans and other animals*. Cambridge, MA: Harvard University.

Whiten, A., & van Schaik, C. P. (2007). The evolution of animal 'cultures' and social intelligence. *Philosophical Transactions of the Royal Society B–Biology*, *362*(1480), 603–620.

Wilson, A. G., Kim, B., & Herlands, W. (Eds). (2016). *Proceedings of the NIPS Workshop on Interpretable Machine Learning for Complex Systems*, Barcelona.

Winston, P. H. (1984). *Artificial Intelligence*. Boston, MA: Addison-Wesley.

Wortham, R. H., & Theodorou, A. (2017). Robot transparency, trust and utility. *Connection Science*, *29*(3), 242–248.

Youk, H., & Lim, W. A. (2014). Secreting and sensing the same molecule allows cells to achieve versatile social behaviors. *Science*, *343*(6171), 1242782.

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(3), 689–722.