Citation for published version:
Bryson, JJ 2012, Patiency is not a virtue: suggestions for co-constructing an ethical framework including intelligent artefacts. in DJ Gunkel, JJ Bryson & S Torrance (eds), *The machine question: AI, ethics and moral responsibility* . Society for the Study of Artificial Intelligence and the Simulation of Behaviour, pp. 73-77, AISB/IACAP World Congress 2012 - The Machine Question: AI, Ethics and Moral Responsibility, Part of Alan Turing Year 2012, Birmingham, UK United Kingdom, 2/07/12.

Publication date:
2012

Document Version
Peer reviewed version

Link to publication

University of Bath

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 20. Jun. 2021

Patiency Is Not a Virtue: Intelligent Artefacts and the Design of Ethical Systems

Joanna J. Bryson

January 13, 2013

Abstract

The question of whether AI can or should be afforded moral agency or patiency is not one amenable either to discovery or mere reasoning, because we as societies are constantly constructing our artefacts, including our ethical systems. Here I briefly examine the origins and nature of ethical systems in a variety of species, then propose a definition of morality that facilitates the debate concerning not only whether it is ethical for us to afford moral agency and patiency to AI, but also whether it is ethical for us to build AI we should so afford.

1 Introduction

The question of Robot Ethics is difficult to resolve not because of the nature of Robots but because of the nature of Ethics. As with all normative ethics, this requires that we decide what "really" matters—what are our ethical priorities? Are we more obliged to our biological kin or to those who share our ideas? Do we value the preservation of culture more or the generation of new ideas? 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 in human society cannot be resolved from first principles or purely by reason.

The primary argument of this article is that integrating a new problem like artificial intelligence (AI) into our moral systems *is* an act normative, not descriptive, ethics. Descriptive ethics may take us some way in establishing precedent, though few consider precedent sufficient or even necessary for establishing what is right. But the advent of potentially-autonomous decision-making human artefacts is novel. Deciding ethical priorities requires establishing the basis of our systems of values.

The nature of machines as artefacts means that the question of their morality is not what moral status they deserve. Rather, we must ask both what moral status we are obliged to assign them, and — at the same time — ask what moral status we are obliged to build them to be competent to meet. This latter aspect of our concurrent, tightly-coupled responsibilities has been neglected even by those scholars who have observed the constructive nature of the first (Coeckelbergh, 2010; Gunkel, 2012b). Note here that *obliged* does require *able* —computationally and indeed logically intractable systems such as Asimov's laws cannot be part of the equation.

What makes reasoning about intelligent artefacts different from reasoning about natural entities is that our obligations can be met not only through constructing the socio-ethical system but also through specifying characteristics of the artefacts themselves. This is the definition of an artefact, it is something we create, and it applies to both ethical systems and artificially intelligent ones. I therefore begin my argument not from what should matter to us but rather from why 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, not to inform our intuitions but to explain them.

To be very clear though from the outset, the moral question I address here is not whether it is possible for robots or other AI artefacts to be moral patients. Human culture can and does support a wide variety of moral systems. Obviously it is easy to describe one such that even certain unintelligent artefacts are owed patiency, and indeed many if not most moral systems hold this for some objects (e.g. particular books or flags). The far more interesting and important question is whether we as academics should recommend putting intelligent artefacts in that position. I will argue 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 avoidable. In doing so I consider not only human

society, but also make potential robots into second-order moral patients. I claim that it would be unethical to put them 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. But I do claim that is incoherent to think that human pleasure can be extended by proxy. Therefore there are costs but 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.

2 The Nature of Life and Intelligence

I start from the entirely functionalist perspective that our system of ethics has coevolved with our species and our societies. As with all human (and other ape, Whiten and van Schaik, 2007) behaviour, it 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 strategies, except that that which must change more quickly should be represented more plasticly (Hinton and Nowlan, 1987; Depew, 2003). As human cultural evolution has accelerated, increasingly our ethical norms are represented in highly plastic forms such as legislation and policy (Ostas, 2001).

The problem with a system of representing behaviour being so plastic as explicit decision making is that this plasticity can lead to dithering. *Dithering* is a technical term for switching from one goal to the other 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. Here 'partial conflict' can be resource-based, for example needing to visually attend to the actions of two children at one time, or needing to both sleep and work. An example of dithering in early computers was called *thrashing*—a result of running two interacting programs which were each nearly as large as primary memory. The operating system would allocate each program a period of processing time, which would necessarily start with an attempt to read that program in to memory from disk, a process called *paging* or *swapping*. If swapping each program in took longer than the time slice allocated to

that program, then both programs would do nothing except alternately attempt to page themselves into memory. The computer would appear to 'freeze' since there would be no actual progress made by either program.

More generally, *dithering* is changing goals so quickly that more time is wasted in the transition than is gained in accomplishment. Thus even when we make decisions about regulating behaviour in the extremely dynamic present, we try to plant them in "permanent" bedrock, like tall buildings built on 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 ancient holy 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.

Perhaps it is to avoid dithering that our society believes that basic principles of our ethics are rational and fixed, and that the apparent changes such as universal suffrage or the end of legalised human slavery are simply "corrections" to make the system more rational. But a better model is to consider our ethical structures and morality to co-evolve with our society. When the value of human life relative to other resources was lower, murder was more frequent and political empowerment less widely distributed (Johnson and Monkkonen, 1996; Pinker, 2011). What it means to be human has changed, and our ethical system has changed along with this.

3 The Origins of Social and Ethical Behaviour

Assessing morality is not trivial, even for apparently trivial, 'robotic' behaviour. MacLean et al. (2010) demonstrate the overall social utility of organisms behaving in what at first assessment seems to be an anti-social way — free riding off of pro-social agents that manufacture costly public goods. Single-cell organisms produce a wise array of public goods ranging from shelter to instructions for combatting antibiotics (Rankin et al., 2010). In this particular case we are discussing the production of digestive enzymes by the more 'altruistic' of two isogenic yeast strains. The yeast must excrete these enzymes outside of their bodies, because (of course) they do not have stomachs. They can only

directly absorb pre-digested food. 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 a common anthropological and economic definition of *altruism*, paying a cost to express behaviour that benefits others (Fehr and Gächter, 2000; Sylwester et al., 2013).

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 via common processes of mutation (Wagner and Altenberg, 1996; Kirschner and Gerhart, 1998; Kitano, 2004; Zhong and Priest, 2011). For these systems, natural selection performs the 'action selection' by determining what proportion of which strategy lives and dies. What MacLean et al. (2010) have shown is that selection operates such that the species as a whole benefits optimally. 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. Where there are insufficient altruists free-riders starve, allowing altruists to invade. Where there are too few free-riders excess food aggregates, allowing free-riders to invade. Thus the greatest good—the most efficient exploitation of the available resources—is achieved by the species as a whole through a mixture of over-enthusiastic altruism and free riding. Why doesn't evolution just optimise the species to produce the optimal level of public goods? This is again due to plasticity. The optimal amount of enzyme production is determined by the ecological substrate the yeast inhabits, and this can change more quickly than the physical mechanism for enzyme production in one strain can evolve. However death and birth can be exceedingly rapid 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 their proportions can be changed very rapidly, is thus the best strategy for tracking the rate of environmental change — for rapidly responding to variation in opportunity.

What do these results imply for human society? Perhaps our culture adds benefit to over-production of public goods by calling the action of creating them 'good' and associating it with social status, while our culture has evolved in a context of being able to rely on self interest to

motivate the maintenance the countervailing population of defectors. Perhaps the 'correct' amount of investment varies by socio-political context, for example increasing massively in times of warfare but returning to more locally-productive levels at times of peace. Could the *reduction* of other's 'good' behaviour itself be an act of public good in times when society requires more individual productivity or self-sufficiency (cf. Trivers, 1971; Rosas, 2012)? This is a thread of a current research programme in my group, looking to explain variations by global region in public-goods regulation as demonstrated by Herrmann et al. (2008); Bryson et al. (2013). We have preliminary evidence that human societies can also be described via varying proportions of individuals applying particular social strategies, and that these proportions vary with the underlying socio-economic substrate.

Of course, *is* does not imply *ought*. The roots of our ethics does not entirely determine where it should or will progress. But the roots do determine our intuitions, which have been proposed as a mechanism for determining our obligations with respect to AI (Dennett, 1987; Brooks, 2002). Because of their origins in our evolutionary past, I do not trust the capacity of our intuitions to inform our decision making. I *do* trust those with vested interests in particular outcomes (e.g. selling weapons, robots or even books) to exploit our gut-level intuitions. In the following section I turn instead to philosophy, to look at how we commonly define moral agency and patiency. In the following sections I exploit these definitions to determine the roles we should expect AI to play in our society.

4 Freedom and Morality

To quote Johnson (2006), "[Moral] action is an exercise of freedom and freedom is what makes morality possible." 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 at least *is* rooted in the one incontrovertible aspect of human uniqueness — language — and our unsurpassed competence for cultural accumulation that language

both exemplifies and enables[1]. The cultural accumulation of new concepts gives us more ideas and choices to reason over, and our accumulation of tools gives us more power to derive substantial changes to our environment from our intentions.

If human morality depended simply on human language then our increasingly language-capable machines would certainly be excellent candidates for agency and patiency. 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 consequences. For example, as our science of psychology develops and more behaviour becomes explicable via other means (e.g. insanity) fewer actions become moral.

I believe we can usefully follow from Dennett's suggestion to generalise morality beyond human ethical systems. Moral actions *for an individual* are those for which:

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

Note that this captures society-specific morals as well as the individual's role as the actor. With this definition I deliberately extend morality to include actions by other species which may be sanctioned by *their* society, or by ours. For example, non-human primates will sanction those that violate their social norms by being excessively brutal in punishing a subordinate (de Waal, 2007), for failing to 'report' vocally available food (Hauser, 1992) or for sneaking copulation (Byrne and Whiten, 1988). While reports of social sanctions of such behaviour are often referred to as 'anecdotal' because

[1]Bryson (2008, 2009b) argues that language while unique is not inexplicably improbable but rather a result of the conjunction of two less-unusual adaptive traits: a heavy reliance on 'culture' — socially acquired behaviour — common in many large, long-lived species of animals, particularly apes, and the capacity for vocal imitation which has emerged independently in many phyla, but nowhere else among the simians.

they are not yet well documented in primate literature, they are common knowledge for anyone who has been lucky enough to work with socially housed primates. I personally have violated a Capuchin monkey norm: *possession is ownership*. I was sanctioned (barked at) by the entire colony — not only those who observed the affront[2], but also those housed separately who had no visual knowledge of the event but joined the chorus of reproach. 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 when they behave as if they expect social sanction when they select the proscribed option.

Returning then to AI, there is then I think no question that we can already train or simply program machines to recognise more or less socially acceptable actions, and to use that information to inform action selection. The question is whether it is moral for us to construct machines that of their own volition choose the less-moral action. The trick here 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 sufficiently obfuscate its decision-making process that we cannot otherwise predict its behaviour, and thus are reduced to applying sanctions to it in order for it to learn to behave in a way that our society prefers[3].

What is fundamentally different from nature here is that since we have perfect control over when and how a robot is created, responsibility is traded off. Consider responsibility for actions executed by the artefact that lie within our own understanding, and thus for which we would ordinarily be responsible. If we choose to assign this responsibility to the artefact we are deliberately disavowing the responsibility ourselves. Currently, even where we have imperfect control as in the case of young children, owned animals and operated machines, if we lose control over entities we have responsibility for and cannot themselves be held accountable, then we hold the responsibility for

[2]I had taken back the bit of a sweater a monkey had snatched but which was still being worn by a guest. I had been warned when first employed never to take anything from a monkey but to ask them to release it, but failed to generalise this to the context where most of the object was still attached to the original owner.

[3]Note that I do not consider training action selection via reinforcement learning or neural networks to be obfuscated in this sense, simply because we don't know the exact 'meaning' of individual components of the internal representation. The basic principles of optimisation that underly machine learning are well-understood (Wolpert, 1996), and sufficient for moral clarity.

that loss of control and whatever actions by these other entities comes as a consequence. If our dog or our car kills a child, we are not held accountable for murder, but we may be held accountable for manslaughter. Why — or in what circumstances — should this be different for a robot?

5 Principles of Robotics

Our consideration of how we should adjust our ethical systems to encapsulate the AI we create requires reasoning about multiple levels of ethical obligation and ethical strategies. In the yeast example I gave earlier, 'anti-social' behaviour actually regulated the overall investment of a society — a spatially-local subset of a species inhabiting a particular ecological substrate — in a beneficial way. Behaviour that was disadvantageous very local to the free-riders was less-locally advantageous to the species. The definition of morality introduced in Section 4 depends on social benefit. Considering robot agency and patiency then, requires considering benefits and costs for at least two potential societies: our own and the robots'. For each of these, consider who benefits and who does not from the designation of moral agency and patiency on AI:

- *The perspective of human well being.* The advantages to humans seem to be primarily that it feeds our ego to construct objects that we owe moral status. It is possible that in the long term it would also be a simpler way to control truly complex intelligence, and that the benefits of that complex intelligence *might* outweigh the costs of losing our own moral responsibility and therefore moral status. The principle cost I see is the facilitation of the unnecessary abrogation of responsibility of marketers 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, 2010).

- *The perspective of AI well being.* Although this argument has been overlooked by some critics (notably Gunkel, 2012a), Bryson (2010, 2009a) makes AI into second-order moral patients by arguing that we should not put AI 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. In short, we can afford to stay agnostic about whether AI have qualia, because we can simply 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 suffers. Just as there are already machines that play chess or do arithmetic better than we do, but none of them aspires to world domination. There can be no costs to the AI in the system I describe, unless we postulate rights of the 'unborn', or in this case the never-designed.

Note that 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 earlier in the section eliminates this confound. For evolved intelligence, dominance structure may be an inevitable part of the selective process, and therefore the dysphoric aspect of subjugation may also be universal. Certainly therefore human ethical systems as part of social regulation have much to say concerning dominance, but that is just one of their roles. But as I have already argued, in designed artefacts we can safely eliminate this dysphoric aspect, replacing it with homogeneity between robots, human regulation of robot roles, and / or emotionally-neutral self assessment by the AI.

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 this) is as extensions of our own motivational systems. We are currently the principle 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 mental prosthetic to our own needs and desires.

The best argument I know against this human-ethics 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, 2012). The UK's official *Principles of Robotics* specifically address this problem in its fourth principle (Boden et al., 2011; Bryson, 2012, cf. Appendix A). This principle does so in two ways. 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'. That is, clear, generally-comprehensible descriptions of their goals and intelligence should be available to any owner, operator or concerned user, presumably over the Internet[4]. This principle was adopted despite considerable concern about the requirement for both therapeutic and simple commercial / entertainment concerns for robots to be masquerade as moral patients and companions (cf. Miller et al., 2012). Because of this consideration, the principle is deliberately set so that transparency may be *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.

One thread of theory for the construction of strong AI 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 must include social striving, pain, etc. So far there is no proof of this position. 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 and social norms on what is the appropriate amount of agency to permit given the benefits it provides.

6 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 AIs as moral agents and patients, in fact many articles in this special issue are working on this project. It may be technically possible to create AI that would meet contemporary requirements for agency or patiency. But even if it is possible, neither of these two statements makes it either necessary or desirable that we should do so. Both our ethical system and our artefacts are amenable to human design. The primary argument of this article is that making AI moral agents or patients is an intentional and avoidable action. The

[4]Note though that the principles stop short of recommending ubiquitous open source software, both because this is of course no substitute for transparent documentation, but also because of security / hacking concerns over robots with access to private homes and conversations.

secondary argument which is admittedly open to debate, is that avoidance would be the most ethical choice.

Acknowledgements

Omitted for anonymity. Please note that many citations to the 2012 proceedings are expected to be replaced by updated citations once the contents of the special issue is determined.

References

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., and Winfield, A. (2011). Principles of robotics. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC). web publication.

Brooks, R. A. (2002). *Flesh and Machines: How Robots Will Change Us*. Pantheon Books, New York.

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

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

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

Bryson, J. J. (2009b). Representations underlying social learning and cultural evolution. *Interaction Studies*, 10(1):77–100.

Bryson, J. J. (2010). Robots should be slaves. In Wilks, Y., editor, *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, pages 63–74. John Benjamins, Amsterdam.

Bryson, J. J. (2012). The making of the EPSRC Principles of Robotics. *The AISB Quarterly*, (133).

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

Bryson, J. J., Martin, D., McIlraith, S. I., and 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., and Powers, S. T. (2013). Understanding and addressing cultural variation in costly antisocial punishment. In Gibson, M. A. and Lawson, D. W., editors, *Applied Evolutionary Anthropology*. Springer. in prep.

Byrne, R. W. and Whiten, A., editors (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Oxford University Press.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3):209–221.

de Waal, F. (2007). *Chimpanzee politics: Power and sex among apes*. Johns Hopkins University Press, twenty-fifth anniversary edition edition.

Dennett, D. C. (1987). *The Intentional Stance*. The MIT Press, Massachusetts.

Dennett, D. C. (2003). *Freedom Evolves*. Viking.

Depew, D. J. (2003). Baldwin and his many effects. In Weber, B. H. and Depew, D. J., editors, *Evolution and Learning: The Baldwin Effect Reconsidered*. Bradford Books, MIT Press.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4):980–994.

Forest, A. (2009). Robots and theology. *Erwägen Wissen Ethik*, 20(2).

Gunkel, D. J. (2012a). *The Machine Question*. MIT Press, Cambridge, MA.

Gunkel, D. J. (2012b). A vindication of the rights of machines. In Gunkel, D. J., Bryson, J. J., and Torrance, S., editors, *The Machine Question: AI, Ethics and Moral Responsibility*, AISB/IACAP World Congress, Birmingham, UK. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

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.

Herrmann, B., Thöni, C., and Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319(5868):1362–1367.

Hinton, G. E. and Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems*, 1:495–502.

Humphrys, M. (1996). Action selection methods using reinforcement learning. In Maes, P., Matarić, M. J., Meyer, J.-A., Pollack, J., and Wilson, S. W., editors, *From Animals to Animats 4 (SAB '96)*, Cambridge, MA. MIT Press.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8:195–204. 10.1007/s10676-006-9111-5.

Johnson, E. A. and Monkkonen, E. H. (1996). *The civilization of crime: Violence in town and country since the Middle Ages*. Univ of Illinois Press.

Kirschner, M. and Gerhart, J. (1998). Evolvability. *Proceedings of the National Academy of Sciences*, 95(15):8420.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5:826–837.

MacLean, R. C., Fuentes-Hernandez, A., Greig, D., Hurst, L. D., and Gudelj, I. (2010). A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biol*, 8(9):e1000486.

Miller, K., Wolf, M. J., and Grodzinsky, F. (2012). Behind the mask: Machine morality. In Gunkel, D. J., Bryson, J. J., and Torrance, S., editors, *The Machine Question: AI, Ethics and Moral*

Responsibility, AISB/IACAP World Congress, pages 33–37, Birmingham, UK. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Ostas, D. T. (2001). Deconstructing corporate social responsibility: Insights from legal and economic theory. *American Business Law Journal*, 38(2):261–299.

Parthemore, J. and Whitby, B. (2012). Moral agency, moral responsibility, and artefacts: What existing artefacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. In Gunkel, D. J., Bryson, J. J., and Torrance, S., editors, *The Machine Question: AI, Ethics and Moral Responsibility*, AISB/IACAP World Congress, pages 8–16, Birmingham, UK. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Pinker, S. (2011). *The Better Angels of our Nature: The Decline of Violence in History and Its Causes*. Penguin.

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

Rohlfshagen, P. and 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., and Torrance, S., editors, *The Machine Question: AI, Ethics and Moral Responsibility*, AISB/IACAP World Congress, pages 29–32, Birmingham, UK. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

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

Sylwester, K., Mitchell, J., and Bryson, J. J. (2013). Punishment as aggression: Uses and consequences of costly punishment across populations. in prep.

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.

Wagner, G. and Altenberg, L. (1996). Complex adaptations and the evolution of evolvability. *Evolution*, 50(3):967–976.

Whiten, A. and 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.

Wolpert, D. H. (1996). The lack of *a priori* distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

Zhong, W. and Priest, N. (2011). Stress-induced recombination and the mechanism of evolvability. *Behavioral Ecology and Sociobiology*, 65:493–502. 10.1007/s00265-010-1117-7.

Appendix A: The EPSRC Principles of Robotics

The full version of the below lists can be found by a Web search for *EPSRC Principles of Robotics*, and they have been EPSRC policy since April of 2011 (Boden et al., 2011). The first list is the principles themselves, in italics, with annotations taken from Bryson (2012).

1. *Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.* While acknowledging that anything can be used as a weapon by a sufficiently creative individual, the authors were concerned to ban the creation and use of autonomous robots as weapons. Although we pragmatically acknowledged this is already happening in the context of the military, we do not want to see robotics so used in other contexts.

2. *Humans, not robots, are responsible agents. Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.* We were very concerned that any discussion of "robot ethics" could lead individuals, companies or governments to abrogate their own responsibility as the builders, purchasers and

deployers of robots. We felt the consequences of this concern vastly outweigh any "advantage" to the pleasure of creating something society deigns sentient and responsible.

3. *Robots are products. They should be designed using processes which assure their safety and security.* This principle again reminds us that the onus is on us, as robot creators, not on the robots themselves, to ensure that robots do no damage.

4. *Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.* This was the most difficult principle to agree on the phrasing of. The intent is that everyone who owns a robot should know that it is not "alive" or "suffering", yet the deception of life and emotional engagement is precisely the goal of many therapy or toy robots. We decided that so long as the responsible individual making the purchase of a robot has even indirect (e.g. Internet documentation) access to information about how its "mind" works, that would provide enough of an informed population to keep people from being exploited.

5. *The person with legal responsibility for a robot should be attributed.* It should always be possible to find out who owns a robot, just like it is always possible to find out who owns a car. This again reminds us that whatever a robot does, some human or human institution (e.g. a company) is liable for its actions.

Below are seven additional points that the authors of the principles direct to their colleagues (c.f. the documents cited above.)

1. *We believe robots have the potential to provide immense positive impact to society. We want to encourage responsible robot research.*

2. *Bad practice hurts us all.*

3. *Addressing obvious public concerns will help us all make progress.*

4. *It is important to demonstrate that we, as roboticists, are committed to the best possible standards of practice.*

5. *To understand the context and consequences of our research we should work with experts from other disciplines including: social sciences, law, philosophy and the arts.*

6. *We should consider the ethics of transparency: are there limits to what should be openly available?*

7. *When we see erroneous accounts in the press, we commit to take the time to contact the reporting journalists.*