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

Joanna J. Bryson
University of Bath
BA2 7AY, United Kingdom

Abstract

The question of whether AI can or should be afforded moral agency or patiency is not one amenable either to discovery or simple reasoning, because we as societies are constantly constructing our artefacts, including our ethical systems. Consequently, the place of AI in society requires normative, not descriptive reasoning. Here I review the basis of social and ethical behaviour, then propose a definition of morality that facilitates the consideration of AI moral subjectivity. I argue that we are unlikely to construct a coherent ethics such that it is ethical to afford AI moral subjectivity. We are therefore obliged not to build AI we are obliged to.

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 considerations, robot 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 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 of normative, not descriptive, ethics. Descriptive ethics may take us some way in establishing precedent, but few consider precedent sufficient or even necessary for establishing what is right. *Is* is not *ought*. Here I will rather assume just two axioms for constructing a moral system:

1. The moral stance should be coherent, under the same principle as that unenforceable laws are not useful (McNeilly, 1968).
2. Where possible there should be minimal restructuring of existing norms, so that introduction of new norms will be less likely to create social disruption or long-term instability. This is based on the example of Common Law (Mahoney, 2001).

The nature of machines as artefacts means that the question of their morality is not simply what moral status they deserve (Miller, 2015). Rather, at the same time we ask both what moral status we ought to assign them, we must also ask what moral status we ought to build them 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 (Coeckelbergh, 2010; Gunkel, 2014). Here *ought* does require *able*—computationally and indeed logically intractable systems such as Asimov's laws are excluded (Myers, 2010).

What makes moral reasoning about intelligent artefacts different from moral reasoning about natural entities is that our obligations can be met not only through constructing the socio-ethical system but also through specifications of the artefacts.

This is the definition of an artefact. Yet empirically this point defies the intuition of many who cannot conceive of intelligence in non-human contexts. Or to be more precise, the historical correlation of language and reasoning with the prototypical moral subjects, humans, is taken as necessarily 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. Therefore the next section of this paper discusses not what *should* matter to us, but rather why things do.

To be very clear, the moral question I address here is not whether it is possible for robots or other artefacts to be moral patients. 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. The more interesting and important question is whether we as AI experts should recommend putting intelligent artefacts in that position, and if so, who or what would benefit.

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, 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 strategies, except that what must change more quickly should be represented more plasticly (Depew, 2003). 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 the other so rapidly that little or no progress is made on either. 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 that each require access to the majority of main memory—alternating so rapidly that each spends the majority of its allocated time swapping into memory from disk, and neither achieves its real function. 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 prefer to regulate social behaviour even in an extremely dynamic present by planting norms in a "permanent" bedrock past, 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 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 and fixed, and that the apparent changes such as universal suffrage or the end of legalised human slavery are simply "corrections". 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 less sanctioned, and political empowerment was less widely distributed (Johnson and Monkkonen, 1996; Pinker, 2012). What it means to be human has changed, and our ethical systems have accommodated that change.

Fundamental Social 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 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 combatting antibiotics (Rankin, Rocha, and Brown, 2010). MacLean et al. (2010) focus on the production of digestive enzymes. Having no stomachs, yeast must excrete such enzymes outside of their bodies. This 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 nearby. The production of these enzymes thus meets the common 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 via common mutations (Kitano, 2004). For these systems, natural selection performs the 'action selection' 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 both strategies (cf. Akçay and Van 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. 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. Why can't the altruistic strain evolve to produce the right level of public goods? This is again due to 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 their proportions can be changed very rapidly, is thus an agile solution.

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 self interest and individual learning are sufficient to motivate and maintain the countervailing population of underproducers. Perhaps the 'correct' amount of investment varies by socio-political context, for example with national military investment making sense in times of war, but local business being more advantageous at other times. This implies that the *reduction* of other's 'good' behaviour can itself be an act of public good in times when society benefits from more individual productivity or self-sufficiency (cf. Trivers, 1971; Rosas, 2012; Bryson et al., 2014). If so, the implications would be that it is easier for human institutions as well to change their collective assessment of ideal public-goods investment than to change their exact level of output or detect the ideal level of effort when investing.

Is does not imply *ought*. The roots of our ethics do not entirely determine where we should or will progress. But roots do determine our intuitions. Our intuitions have been proposed as a mechanism for determining our obligations with respect to robots and AI (Dennett, 1987; Brooks, 2002). 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. I do however trust those with vested interests—such as selling weapons, robots, or even books—to exploit such intuitions. In the next section I turn as an alternative to philosophy, to look at how we rigorously define moral agency and patiency. In the following sections I exploit these definitions to propose a more coherent, minimally disruptive path to constructing robot ethics, as promised in the Introduction.

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 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 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 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 consequences. For example, as our science develops and our behaviour becomes more explicable via other means (e.g. insanity) fewer actions are moral.

I believe we can usefully follow from Dennett to generalise morality beyond human ethics. Moral actions are those for which:

1. a particular behavioural context affords more than one possible action for an 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 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 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 individuals that violate their social norms, e.g. for 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)[1]. Similarly, this definition allows us to say pets 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.

[1]While reports of social sanctions of such behaviour are often referred to as 'anecdotal' they are common knowledge for anyone 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, but all in hearing range of the sanction.

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. The question is whether it is moral for us to construct machines that would of their own volition choose the less-moral action. The key 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. I do not consider training action selection via reinforcement learning or neural networks to be obfuscated in this sense. Even if 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 and sufficient for moral clarity. 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 courts of law.

What is fundamentally different from nature here is that 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, if we lose control over entities we have responsibility for and cannot themselves be held accountable, then we are held responsible for 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 can and should be held accountable for negligence and manslaughter (Liao, 2015). Why—or in what circumstances—should this be different for a robot?

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 way that helped it compete with other species. Behaviour possibly disadvantageous very local to free riders was less-locally advantageous to the species. The definition of morality introduced above depends on social benefit. Considering whether a robot should be a moral subject 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 designating moral agency and patiency to 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 principal 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).
- *The perspective of AI well being*. Although this argument has been overlooked by some critics (notably Gunkel, 2012), Bryson (2010, 2009) makes AI into second-order moral patients by arguing that we should not put it 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 suffer. 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 '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, and therefore the dysphoric aspect of subjugation may also be universal. But in designed artefacts we can safely eliminate this dysphoric aspect of subservience. 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.

The Introduction suggested criteria for ethical systems of coherence and lack of social disruption. In this context, 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, and I can think of no coherent reason to 'pass the baton' to machines made to share and compete for these preferences. 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?

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 mental prosthetic to our own needs and desires.

The best argument 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 UK's *Principles*

of Robotics specifically address this problem in its fourth principle, and in two ways (Boden et al., 2011, cf. Appendix A). 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 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, Wolf, and Grodzinsky, 2015). Because of this consideration, the 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), we should be able to both experience emotional engagement and maintain explicit knowledge of their lack of moral subjectivity.

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 evidence for 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, regulation, and social norms on what is the appropriate amount of agency 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 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 systems 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 secondary argument which is admittedly still open to debate, is that avoidance would be the 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.

References

Akçay, E., and 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).

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).

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

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*, 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.; Powers, S. T.; and Sylwester, K. 2014. Explaining cultural variation in public goods games. In Gibson, M. A., and Lawson, D. W., eds., *Applied Evolutionary Anthropology: Darwinian Approaches to Contem-*

porary World Issues. Heidelberg: Springer. 201–222.

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*, 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 Wilks, Y., ed., *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. Amsterdam: John Benjamins. 63–74.

Byrne, R. W., and Whiten, A., eds. 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*. Massachusetts: The MIT Press.

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

Depew, D. J. 2003. Baldwin and his many effects. In Weber, B. H., and Depew, D. J., eds., *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. 2012. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. The MIT Press.

Gunkel, D. J. 2014. A vindication of the rights of machines. *Philosophy & Technology* 27(1):113–132.

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.

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.

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.

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.; and Gudelj, I. 2010. A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biol* 8(9):e1000486.

Mahoney, P. G. 2001. The common law and economic growth: Hayek might be right. *The Journal of Legal Studies* 30(2):pp. 503–525.

McNeilly, F. S. 1968. The enforceability of law. *Noûs* 2(1):47–64.

Miller, K.; Wolf, M. J.; and 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* 1–23. in press, online first.

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., and 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.

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.

Rosas, A. 2012. The holy will of ethical machines. In Gunkel, D. J.; Bryson, J. J.; and Torrance, S., eds., *The Machine Question: AI, Ethics and Moral Responsibility*, AISB/IACAP World Congress, 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.

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.

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.

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).

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 (Internet) access to documentation about how its 'mind' works, a sufficient faction of the population would stay informed prevent gross exploitation.
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.