AISB 2016
Symposium on the Principles of Robotics

April 4th 2016, Sheffield, UK.

Edited by
Tony J Prescott

Organising Committee
Tony J Prescott
Alan Winfield
Madeleine de Cock Buning
Joanna J Bryson
Noel Sharkey

Part of the 2016 Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB)

About the Symposium

It is five years since the publication of “Principles of Robotics”[1] developed by a panel of distinguished British robotics and AI experts at an EPSRC/AHRC funded retreat. The principles, which were aimed at “regulating robots in the real world”, were stated in the form of five “rules” and seven “high-level messages.” The five rules are as follows:

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.
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.
3. Robots are products. They should be designed using processes which assure their safety and security.
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.
5. The person with legal responsibility for a robot should be attributed.

The principles have had significant impact in UK robotics research, and continue to provoke substantial debate. At a time when public concern about the development of robot technologies is heightening we consider that it would be useful to revisit the principles to consider their continued relevance according to the following criteria:

a. *Validity*—are the principles correct as statements about the nature of robots (for instance that they are tools and products), robot developers, and the relationship between robots and people (for instance that robots should have a transparent design), or are they ontologically flawed, inaccurate, out-dated, or misleading.
b. *Sufficiency/generality*—are the principles sufficient and broad enough cover all of the important issues that might arise in the regulation of the robotics in the real-world or are significant concerns overlooked.
c. *Utility*—are the principles of practical use for robot developers, users, or law-makers, in determining strategies for best practice in robotics, or legal standards or frameworks, or are they limited in their use by lack of specificity or through allowing critical exceptions (such as the use of robots as weapons for the purpose of national security).

A 1-day symposium was held on the 4th of April as part of the AISB 2016 Conference in Sheffield, UK. These proceedings contains commentaries on the principles solicited in advance of the meeting. Commentaries were checked for relevance by the organising committee but were not peer reviewed. No contributions were rejected.

---

[1] https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

AISB 2016 Symposium on the Principles of Robotics

Submitted Commentaries

The meaning of the EPSRC principles of robotics

Joanna Bryson, University of Bath, and Princeton Center for Information Technology Policy

Introduction

In revisiting the principles of robotics, it is important to carefully consider their full meaning. Here I briefly visit first the meaning of the document as a whole, then of its constituent parts. The EPSRC principles of robotics were generated as a deliverable by a group assembled with little guidance and no deliverable required. The original intention of the epsrc robotics event seems to have been only the discussion itself, or perhaps even only the fact of the meeting. The academics present wanted something to show for their time spent, and as a result a substantial amount of time of all those present on the final day went into the creation of the three versions of the principles and their documentation. Some of the documentation was extended–again by consensus–after the meeting.

It is right and fitting that there should be a way to examine and even update or maintain the document. Even national constitutions have means for maintenance. However, it is critical to the efficacy of policy documents that they are not easy to change. They should provide a rudder to prevent dithering, and as such are ordinarily more difficult to alter than they were to instantiate in the first place. Note that some countries and other political unions have not found it easy to create even their initial constitutions for this very reason. Therefore it's important to think carefully about the meaning of the principles.

The principles as policy

Technology policy, and policy more generally, is a surprisingly amorphous thing. Like other aspects of natural intelligence, policy is not always found resident in the law or even governance. Much of policy is unwritten and even not explicitly known. The UK is actually outstanding in its innovation of the common law, which acknowledges this and the importance of culture and precedent. Nonetheless, in the cold light of a committee working on REF impact cases, we have to ask, are the principles policy? I think the answer is "yes". They are a set of guidelines agreed by a substantial if perhaps arbitrary fraction of the community they affect, and they are published on government web pages.

All policy has three components: allocative, distributive, and stabilising. The **allocative** is the process of determining what problems are worth spending time and other resources on. In the case of the principles, this was instigated by the EPSRC (or some organisation above them) out of concern that the British public might reject robotics as they had genetically modified food. We were told the rejection of robotics was seen as a severe threat to the British economy. Note also that each of the participants (at least those not specifically paid to attend) also made individual investments, allocating time to the problem of robot ethics, though for many this was confounded with an opportunity to get better known by their primary funding organisation.

The **stabilising** component is the one that ensures that the policy, once set, is incorporated into society in such a way that it is unlikely either to be quickly undone or to become much of a liability or matter of controversy. In the case of the principles this has evidently been achieved at least to some level since we are celebrating their fifth anniversary. From talking to other authors, I know of none entirely enamoured with the final product, but all respect the (admittedly representative) democratic process by which they were achieved, and the importance of their colleagues' mutual commitment to the final product. I for one would love to see the principles further reified into policy or even law, but I have yet to discover the process by which this might be accomplished. However, they have been and are continuing to be drawn to the attention of various standards boards and parliamentary enquiries as well as of the press and other academics.

I leave for last the most controversial aspect of policy: the **distributive**. At its base, all policy is about action selection, and that implies the allocation or rather reallocation of resources. Politics tries to brush over this, since it necessarily goes against the grain of those from whom the resources are reallocated, even in the cases where those individuals stand to gain net benefit. We hate to lose control, but policies are for control. "Tries to brush over" is in fact an understatement; making redistribution palatable may be the core project of politicians. In this case, the government had very specific concerns about individuals who had been in the media promoting fear of robots, and were very clear in their desire to find ways to shift media attention and public impressions towards the safety of robotics. In contrast, it was really the participants who brought up the other major shifts from sensationalism to pragmatism --- the assertion that robots are not responsible parties under the law, and that users should not be deceived about their capacities. The council representatives knew this redistribution of power would anger some of their outstanding funding recipients, and the participants knew the same about some of their colleagues. Nevertheless, there was striking unanimity amongst the academics that the greatest moral hazards of robots was their charismatic nature and the incredible eagerness many people have to invest their own identity in machines', leading to striking confusion about their nature all of us had witnessed. This charisma and confusion left the door open for all kinds of manipulations by corporations and governments, where the robots could be set up as responsible or even as surrogate for human lives or values.

The principle of killing

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.

The first three principles were intended as corrections of Asimov's laws. Robots are not responsible parties, so they could not kill. Instead, robots should not be usable as tools for killing. This simple rule made the transfer of moral subjectivity clear, and simultaneously met the pacifist desires of most present. However, pragmatically, robots were already used as weapons of war, and a law that is unenforceable is questionable. We were persuaded that leading with a principle known to be false would significantly decrease our chances for cultural impact. The meaning of the first principle might therefore seem neutralised by the compromise of the exception, but that robots are not to be weapons in civil society is still an important social point. Beyond this, the fact that practical policy has to take into account the needs of the government to address both security and industry (the UK is the world's fifth largest arms dealing nation) also has meaning. However purely academic some of us may wish our discipline to be, the fact that many of its products have immediate utility means that we cannot avoid impact on our world.

The principle of compliance

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.

The second Asimov law has to do with following instructions, but even the notion of obeying implied moral agency. The original meaning of this law was that robots are ordinary technology and conform to ordinary standards and laws. In the shaping of the principles as a suite, the second principle came to be the one that communicated further some of the peril of AI in general, and AI mistaken for a moral subject in particular. The emphasis on privacy reflects the special concern of a perceiving intelligent physical agent occupying the exact same space as a human family. The technology is fundamentally immersed in the human *umwelt*, more than any previous technology or pet, perhaps even more so than some humans in a household such as children. It has access to written and spoken language, social information, observed schedules *etc*. Further, it may be mistaken for a pet or other trusted family member, its special abilities for perfect communication to the outside world temporarily forgotten, or its abilities to learn regularities and classify stimuli. In these cases, primate information may be unintentionally stored in a public cloud, or even a supposedly private one susceptible to hacking. Forcing such a novel, human-like technology into compliance with standard, legal norms of privacy and safety is a non-trivial task.

The principle of commoditisation

Robots are products. They should be designed using processes which assure their safety and security.

The final Asimov law is self protection, but robots have no selves. Instead this law focussed on protecting humans from robots at the level of the robot’s basic soundness. The principle again brings us into awareness of the non-special manufactured nature of the robot, in an attempt to head off avoidance of legal liability by claiming robots have a unique nature. The manufacturer of a robot should have exactly as much responsibility for the machinery working to specification as the manufacturer of a car or a power tool. In fact, robots might be cars or power tools, but if so they should be more rather than less safe than the conventional variety of either.

The principle of transparency

Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.

The first three principles established the legal framework for the manufacture and sales of robotics as being identical to other products. The last two are intended to ensure that status is also communicated to the user. The principle of transparency seeks to ensure that individuals do not overinvest in their technology, for example hiring a house sitter to keep the robot from being lonely. Some roboticists object to this principle because deception is necessary for the efficacy of their intended application, such as making people to not feel lonely so they are less depressed. Others contend that this principle denies the possibility that robots should be more than ordinary machines. The first argument is open to experiment. First it needs to be established that there is no way to trigger emotional engagement without deception, which seems unlikely given the extent of emotional engagement that is established with fictional characters and clearly non-cognizant objects. If a requirement for deception is experimentally established, then the tradeoff between the costs and benefits of deception can be debated. The second however is incontrovertible. The authorship we have over artefacts is a fundamental part of their machine nature --- AI is definitionally an artefact. To some extent, we might even argue that this principle is self-limiting. If AI really were to be able to alter what it means to be a machine, then communicating this modified machine nature would still meet this principle.

The principle of legal responsibility

The person with legal responsibility for a robot should be attributed.

Finally, the fifth principle communicates the robots' status as artefacts in the most fundamental way possible. They are owned, and that ownership must be legally attributed. The fact that robots are constructed and owned is the reason I have previously argued that we are ethically obliged not to make them persons --- because owning persons is unquestionably unethical. The argument is not that there exists person-like robots that we should demote in status legally, but rather that the necessarily-demoted legal status means that we should not cause person-like-ness to be a feature of any robot legally manufactured. However, the principles of robotics do not go to this extreme of futurism. As I said earlier, they focus on communicating the present reality to a population so eager to own and identify with the superhuman that they might easily be lead to believe that a robot badly manufactured or operated is itself to blame for the damage inflicted with it. If you hear a horrible noise and find a car smashed into your house, you can quickly and easily identify the owner of the car, even if the car is presently empty, simply through its number plates or in the worst case through serial numbers. The idea is that the same should be true if you find a robot embedded in your property. The participants in the robotics retreat accurately predicted a problem now already present in our society because of drones, and one that is now being addressed in some nations with mandatory licensing such as the committee recommended.

Conclusion

To summarise, the EPSRC principles are of value because they represent a policy constructed at significant taxpayer and personal cost. While no policy is perfect, ideally they should only be replaced by a new policy with an equivalently high or higher level of investment both by government and domain experts. Their purpose is to provide consumer and citizen confidence in robotics as a trustworthy technology fit to become pervasive in our society. The individual principles each represent substantial concerns of the experts and stakeholders, though sometimes that representation is itself not perfectly transparent. The overall goal was to clearly communicate that responsibility for safe and reliable manufacturing and operation of robots was no different than for any other objects manufactured and sold in the UK, and therefore the existing laws of the land should be adequate to cover both consumers and manufacturers.

It is important to realise that this is not the case for all conceivable robots. It is easy to conceive of unique works of art that qualify as robots and are not like commoditised products, or to conceive of robots that are simply built in an unsafe or irresponsible manner. What people have more trouble conceptualising is that there may be cognitive properties such as suffering that might possibly be feasible to incorporate into a robot, but to do so would be as unethical as putting faulty brakes on car. The principles of robotics do not seek to determine what is possible; they seek to communicate advisable practices for integrating autonomous robotics into the law for the land.

Robotics Research Ethics Discussion

A. Gning, D. Davis, Y Cheng, P. Robinson, Computer Science Department,
Hull University.

e.gning@hull.ac.uk

Introduction

In modern world with the development of technological resources, Robotics have resulted in numerous applications [1] and often unpredicted deployment in real life (e.g. the increasing use of drones in civil applications). To go along with this age of robotics, the research community and the society in general need to define ethical principles that are general enough to be robust to evolutions in time and adapted to the range of possible applications. Ethical principles should be vulgarised and universalised enough for the designer and makers of robots to be self-aware of regulations and limits to respect.

In general, robotics' applications can be classified into four groups: domestics or human assistive robots [2] [3] [4], medical robots [5] [6] [7] [8], defence robots [9] [10] and industrial robots [11] [12]. The discussion is focused on the first three group of robots since the industrial robot are often bounded to limited areas with pre-specified sets of limited tasks and are not in direct interaction with society.

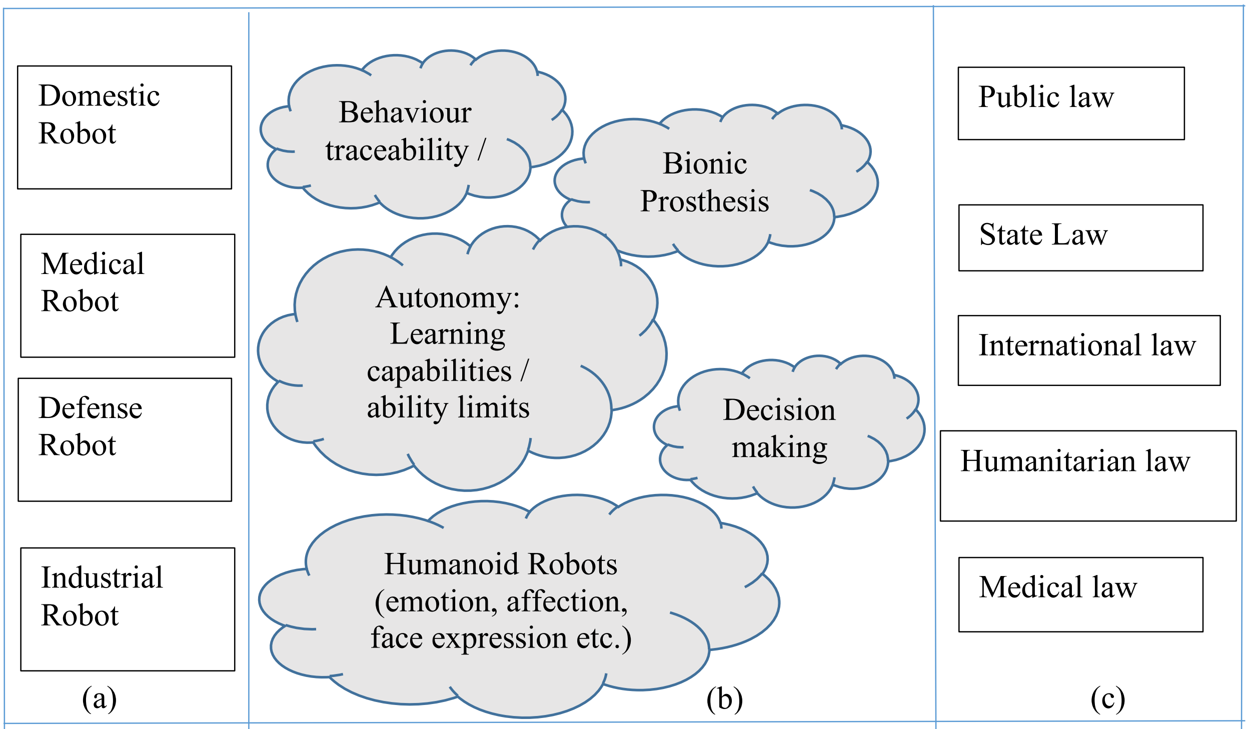

Figure 1. (a) Robotics groups, (b) key area of robotics research topics raising ethics issues, (c) domain of law regulations to be considered.

Figure 1 lists the difficult task of regulating ethics for different group of robotics in regard to five popular research topics. In each group of robotics (figure 1-(a)), ethical questions can be raised and key ones are represented (figure 1-(b)): how can we design robots to always be able to interpret their behaviour, to know about learning robots limits and boundaries; humanoid robots with imitation of human gestures and face animation can push the interaction with human (especially children and disabled peoples) to new frontiers raising needs to regulate and anticipate on possible implications; Nowadays humans can greatly benefit from Prosthesis thanks to progress on robotics research. However this can result in non-disabled people seeking prosthesis that can enhance their capabilities raising again the need for new regulations. Finally, the biggest challenge relate to decision capabilities of robots. These decisions directly involve human life and hence raise questions of legal issues resulting from executed actions.

On top of these research topics, robotics ethics have to be compatible with a set of law domains (figure 1 –(c)). There is a need to take into account that robots need to be compatible with different levels of laws that can change for instance from region to region or state to state.

Five years ago in UK, a panel of distinguished robotics and AI experts published the EPSRC's Principles of Robotics in the form of five rules and seven high level messages. We propose to discuss these rules with a focus on the transversal structure - between robotics groups, research topics and law frameworks presented in figure 1 - and in respect with the three criteria of validity, sufficiency and utility.

Discussion over the set of five rules

The general comment that can be made about the set of five rules is that it is rather ambitious to think that it is possible to give common/uniform guidance to all type of robots, across all possible evolution of research and all frameworks of law. It would be more natural to seek guidance rules reflecting the transversal nature of robotics shown in figure 1. We believe that the five rules are not general enough and the rules should be specified and particularised for each group of robot in figure 1 (a) taking into account the specific aspect of laws and research avenues involved in figure 1 (c) end (b).

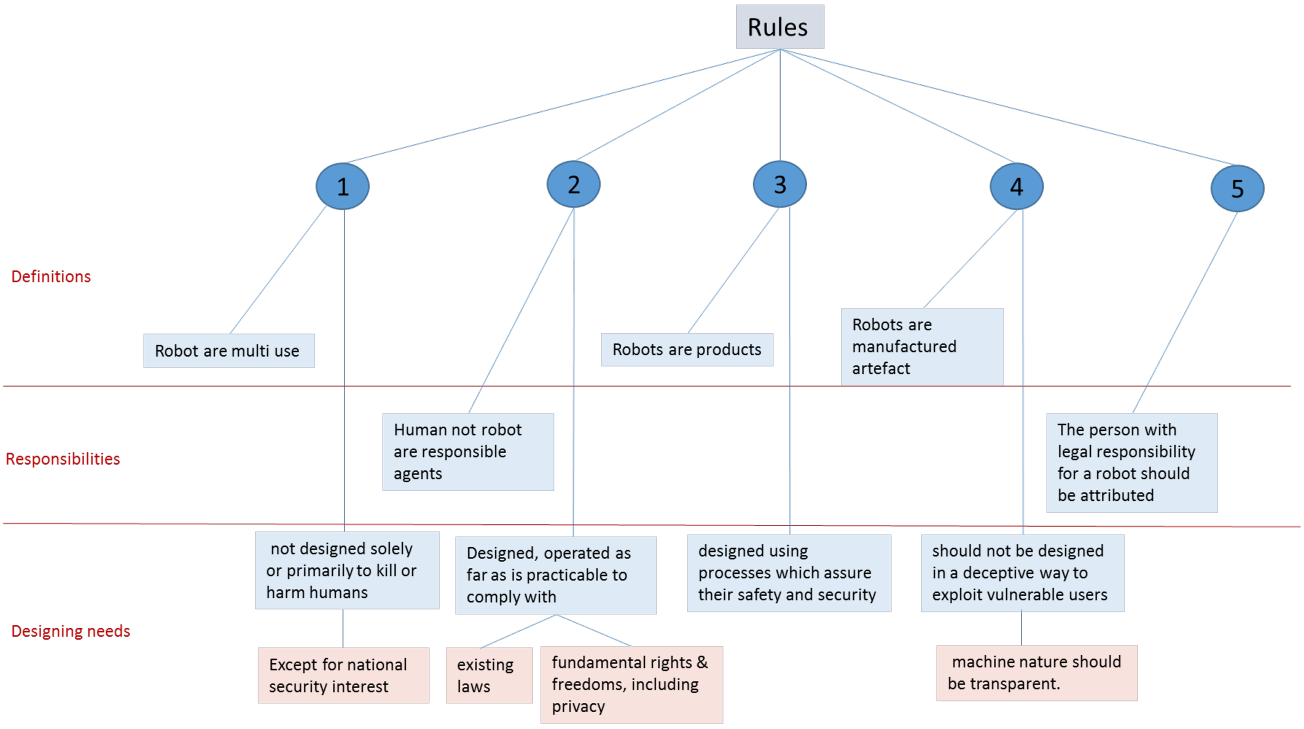
Rules
1
2
3
4
5
Definitions
Robot are multi use
Robots are products
Robots are manufactured artefact
Responsibilities
Human not robot are responsible agents
The person with legal responsibility for a robot should be attributed
Designing needs
not designed solely or primarily to kill or harm humans
Designed, operated as far as is practicable to comply with
designed using processes which assure their safety and security
should not be designed in a deceptive way to exploit vulnerable users
Except for national security interest
existing laws
fundamental rights & freedoms, including privacy
machine nature should be transparent.

Figure 2. The set of five rules

Figure 2 represents in a graphical way the set of five rules. It can be seen that there is a common pattern to the five rules: the first sentences are often generalities and definitions about robots such as "robots are products" or state that the responsibilities are not liable to robot but human instead. The last sentences state the designing needs and bounce between safety issues, legal issues and transparency issues.

Clearly criticisms about the structure shown in Figure 2 can be listed as follow

- The five rules are sometime overlapping. For instance "*complying with existing law*" in rule 2 encapsulates "*not designed solely or primarily to kill or harm human*" in rule 1; "*the person with legal responsibility for a robot should be attributed*" in rule 5 can encapsulate implicitly "Human not robot are responsible agents" in rule 2.
- The five rules are not general. For instance, we can barely see how Bionic Prosthesis can be fitted in the rules at the present form (e.g., robotics research can allow in the future modification of human body in order to gain more power, speed etc.). Artificial Sexuality is another example of controversial research that can result into ethical questions.
- The Five rules are not adamant that law can be contradictory depending on the domain considered from councils, regions, countries and continent. A similar example often encountered in research is patent applications for which several specific studies are needed to apply in given regions of the world. Laws can be even more complicated since religious belief and people habits and customs will dictate the notion of ethics.

Conclusions

In this discussion, we briefly provided arguments on the need of different formulation for the five rules. It is demonstrated through a pictorial representation of the five rules that there are in fact not sufficient, are overlapping and not explicitly reflecting the true challenges of robotics ethics.

We based part of our reasoning on the transversal nature of robotic ethics along three line: groups that constitutes robotics, future avenues of robotics that are essential to be captured when defining ethics, and the structured nature of law. We recommend to a natural reformulation that will differentiate ethics for each group of robotics while being adamant of the contradictions and strong constraints that can exist due to the structural nature of law.

Bibliography

[1] G. Bekey, Robotics: State of the art and future challenges., california: London Imperial College Press. , 2008.

[2] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, K. L. Koay and I. Werry, What is a robot companion-friend, assistant or butler?, I. I. C. o. I. R. a. Systems, Ed., 2005.

[3] J. Forlizzi and D. C., Service robots in the domestic environment: a study of the roomba vacuum in the home, 1st ACM SIGCHI/SIGART conference on Human-robot interaction. ACM, 2006.

[4] J. Y. Sung, R. E. Grinter, H. I. Christensen and L. Guo, Housewives or technophiles?: understanding domestic robot owners, 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2008, pp. 129-136.

[5] G. P. Moustris, S. C. Hiridis, K. Deliparaschos and K. Konstantinidis, Evolution of autonomous and semi-autonomous robotic surgical systems: a review of the literature,

vol. 7, The International Journal of Medical Robotics and Computer Assisted Surgery, 2011, pp. 375-392.

[6] J. Rassweiler, J. Binder and T. Frede, "Robotic and telesurgery: will they change our future?," vol. 11, no. 3, pp. 309-320, 2001.

[7] G. Kwakkel, K. B.J. and K. H.I., Effects of robot-assisted therapy on upper limb recovery after stroke: a systematic review, Neurorehabilitation and neural repair, 2007.

[8] K. Cleary and C. Nguyen, "State of the art in surgical robotics: clinical applications and technology challenges," vol. 6, no. 6, pp. 312-328, 2001.

[9] P. W. Singer, "Robots at War," *Wilson Quarterly ,* 2008.

[10] T. K. Adams, "Future warfare and the decline of human decisionmaking," *Parameters,* vol. 31, no. 4, 2001.

[11] P. Leitão, "Agent-based distributed manufacturing control: A state-of-the-art survey," *Engineering Applications of Artificial Intelligence,* vol. 22, no. 7, pp. 979-991, 2009.

[12] Z. M. Bi, S. Y. Lang, W. Shen and L. Wang, "Reconfigurable manufacturing systems: the state of the art," *International Journal of Production Research,* vol. 46, no. 4, pp. 967-992, 2008.

Robots are not just tools

Tony J. Prescott, University of Sheffield

At the heart of the EPSRC principles of robotics (henceforth 'the principles') are a number of ontological claims about the nature of robots that serve as axioms to frame the subsequent development of ethical challenges and rules. These include claims about what robots are, and also about what they are not. The claims about what robots are include that "robots are multi-use tools" (principle 1), that "robots are products" (principle 3) and "pieces of technology" (commentary on principle 3), and that "robots are manufactured artefacts" (principle 4). The claims about what robots are not include that "humans, not robots, are responsible agents" (principle 2), that robots are "simply not people" (commentary on principle 3), and that robot intelligence can give only an "impression of real intelligence" (commentary on principle 4).

On first reading these statements seem straightforward assertions of obvious truths. I will argue that this is not the case. Instead, I will propose that these ontological commitments lack nuance, they assume all too easily that we know the boundary conditions of future robotics development, and they obscure or ignore some of the important ethical debates. If this is at all true, then progress towards a more useful set of principles could begin by thinking carefully about the ontological status of robots.

If we look at how the principles are presented, there seems an implicit process of induction at work that allows statements about what most current robots are, to be re-interpreted as statements about what robots must essentially be. For example the statement that robots as multi-use tools in principle 1, slips into the claim that robots are "just tools" in the commentary on principle 2 and to the statement that "robots are simply tools of various kinds, albeit very special tools" in the preamble. Whilst it is easy to agree with a general statement that robots are multi-use tools, especially in the context of a discussion about dual use (principle 1), the much stronger claim that robots are just tools, or simply tools, denies that they could sensibly belong to other disjoint categories.

Take the category of 'companion' for instance. There is a major effort around developing robot companions that can provide social and emotional support to people as partially acknowledged in the discussion of principle 4. The category of tools describes physical/mechanical objects that serve a function, whereas the category of companions describes significant others, usually people or animals, with whom you might have a reciprocal relationship marked by emotional bond. The possibility that robots could belong to both these categories raises important and interesting issues that are obscured by insisting that robots are just tools.

Indeed, consistent with the view of robots as tools, the discussion of robot companionship in the principles is pretty dismissive, describing them as toys that could afford some pleasure to people who are unable to, or cannot afford to, keep animal pets. Robots are faux companions on this account that create an "illusion of emotions" and their intelligence is artificial and not "real". The faux nature of robot companions, it is argued, creates a real ethical problem in that robot companions are potentially deceptive and so should be designed so that their "machine nature is transparent".

The ontological problem here particularly concerns the claim that robots could never possess psychological capacities such as "real" emotions or intelligence. What these are, in human terms, is hotly debated in the cognitive and brain sciences. There is therefore no compelling reason to believe that these capacities must be unique to humans and could not be shared by machines. Indeed, there are counter-claims that robots, suitably configured, can have emotions [1], whilst the future of artificial intelligence, as intelligence, has no obvious ceiling at below-human level.

A further problem concerns the assumption about how people will *see* robots—specifically, that robots will be seen as tools if they are shown in a transparent way. This could easily be wrong, for instance, people may anthropomorphise robots regardless of how obviously they are manufactured products. One reason to think this could be the case is the strongly social nature of our brains, and how easily our empathy is triggered by something that appears life-like. The Heider-Simmel animations of simple geometric figures [2] (see figure), show just how crude this information can be and yet we will still see intentionality, motivation, even emotion. The invention of the Tamagotchi digital pet demonstrated that a simple 2-d animation of an animal-like creature can create a compelling urge to care [3]. We do not need to believe that the psychological state we read in to these artefacts is real in order to have an authentic emotional response to this ourselves.

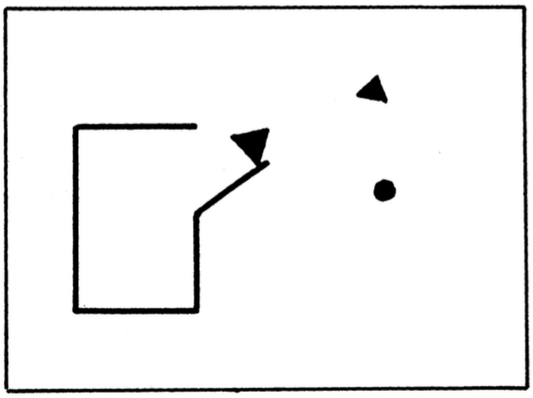

Figure. Geometric shapes moving around in a simple animation were interpreted has "animated beings, chiefly persons", in this famous 1944 study by Heider and Simmel.

An analysis of ontological and psychological issues in human-robot interaction has previously been made by Kahn and colleagues [4]. Following a similar train of thought, we can describe four general ways in which ontological perspectives on *what robots are*, and psychological perspectives on *how robots are seen*, could combine. These are illustrated in the following table along with some of the ethical issues they entail.

I. Robots are just tools (o), and people will see robots as just tools unless mislead by deceptive robot design (p). *Ethical issues: We should address human responsibilities as robot makers/users and the risk of deception in making robots that appear to be something they are not.* ***This is the position of 'the principles'.***	*II.* Robots are just tools (o), but people may see them as having significant psychological capacities irrespective of the transparency of their machine-nature (p). *Ethical issues: We should take into account how people see robots, for instance, that they may feel themselves as having meaningful and valuable relationships with robots, or they may see robots as having important internal states, such as the capacity to suffer, despite them not having such capacities.*
III. Robots can have some significant psychological capacities (o) but people will still see them as just tools (p). *Ethical issues: We should analyse the risks of treating entities that may have significant psychological capacities, such as the ability to suffer, as though they are just tools, and the dangers inherent in creating a new class of entities with significant psychological capacities, such as human-like intelligence, without recognising that we are doing so.*	*IV.* Robots can have some significant human-like psychological capacities (o), and people will see them as having such capacities (p). *Ethical issues: We should consider scenarios in which people will need to co-exist alongside new kinds of psychologically significant entities in the form of future robots/AIs.*

Table. How ontological (o) and psychological (p) perspectives on robots can combine (after Kahn et al., 2007).

Note that only one quadrant of this table (I) is addressed in the principles, but that II, III and IV are all possible, at least theoretically. To conclude this essay I want to briefly consider some of the ethical issues that arise in quadrants II–IV.

In quadrant II, interesting questions arise how robots should be treated—not because they are sentient agents but because people will choose to treat them as such. For instance, the idea that it should be unlawful to wilfully damage robots, proposed as part of the South Korean "Robot Ethics Charter" [5], or that we might mourn the loss of a favourite robot, as has been reported for some Japanese owners of *Sony Aibo* robot dogs [6], does not seem so strange when viewed from the perspective of how robots are seen by people rather than in terms of what they are. Of course, appearance and function do matter, but transparency of "machine nature" will only be one factor of many influencing how people see and behave towards robots, and it may be naïve to assume that it will be a decisive one. The bonds people will form with some robots may be similar to those we develop with other valued possessions, such as cars and mobile phones. On the other hand, for some robots, they may be more like the relationships we have pet animals, including for instance, wishing to support and nurture them (something that we may ourselves find rewarding). Finally, some human-robot relationships may share similarities to human-human relationships. For instance, I may develop a bond with my companion robot not because it looks human but because it has the capacity to remember and communicate with me about some of our shared experiences. More generally, what may be needed, in order to develop suitable ethical principles, is to develop a taxonomy of the different forms of emotional bonds that could exist between robots and people and analyse the factors that could underpin the development and maintenance of such relationships [7].

Quadrant III concerns the possibility of robots having significant psychological capacities that are in danger of being over-looked by people. This raises ethical risks that are not discussed in the principles, but that have been highlighted by others. For instance, Metzinger [8] has argued that we could build robots that are capable of experiencing suffering without realising that we are doing so, and therefore create a new kind of sentient entity that suffers unnecessarily due to our actions, this is clearly ethically problematic if it were to happen. Although this may seem unlikely in the near-term, there are grounds to consider that this could be a risk in the medium- to long-term as cognitive architectures for robots become more sophisticated. Several trends in on-going research on human consciousness also support this possibility. First, one of the major contemporary theories of consciousness [9] asserts a critical role for integration of information that doesn't necessarily require a biological substrate. Neurologists are also re-appraising whether islands of integrated activity in the brains of 'locked-in' patients might constitute a form of minimal consciousness [10]. Finally, there is an active debate as to whether animals with smaller brains than ours, such as fish, might be sentient in a significant way (e.g. that they may experience pain) [11]. These developments suggest that consciousness could be possible in an artificial agent without having to match the size or complexity of an intact human brain. Dennett has argued that "crude, cheesy, second-rate, artificial consciousness" could be possible in robots [12], and Bryson [13] has proposed that today's robots might already have some simple forms of consciousness that meet some commonly proposed criteria. None of this is to claim that we are in quadrant III yet, but given the risks, ethicists should be pressing us as to how we would know if we were.

One of the consequences of the view of robots as "just tools" is the implicit dismissal of the possibility of strong AI—that future robots could have human-level, or beyond human-level general intelligence. A quadrant III/IV issue, recently discussed by noted scientists and innovators such as Stephen Hawking, Elon Musk, and Bill Gates, to name a few, and analysed in-depth by Bostrom [14], is that an AI singularity could reverse the master-slave relationship between humans and robots. The conviction that robots/AIs are "just tools" may keep us from recognising the early signs of such a self-bootstrapping super-AI. An ethical approach would surely encourage more vigilance. A more positive quadrant IV stance on this AI 'singularity' debate is the perspective of the 'global brain', proposed by Heylighen [15] and others, that humans and advanced AIs could co-exist to our mutual benefit. This reminds us that that ethics must be about analysing the potential benefits as well as the risks.

Although quadrant III/IV scenarios may seem far-fetched or at least distant, such concerns have captured the public imagination and have prompted significant calls for debate (e.g. [16]). In my own experience of talking to members of the public, and of the media, these are often the topics about which there is the greatest interest and concern. The attempt to create a rhetorical barricade against these issues by insisting that robots are just tools may do little to calm the voices and could come across as hegemonic and condescending. Whilst approaches to these longer-term ethical challenges are necessarily speculative, a starting point is to acknowledge that there are concerns here that are worthy of further attention.

A more candid approach may be to recognise that, whilst most robots are currently little more than tools, we are entering an era where there *will* be new kinds of entities that combine some of the properties of machines and tools with psychological capacities that we had previously thought were reserved for complex biological organisms such as humans. Following Kang [17], the ontological status of robots might be best described as *liminal*—neither living in quite in the same way as biological organisms, nor simply mechanical as with a traditional machine. The liminality of robots makes them both fascinating and inherently frightening, and a lightning rod for our wider fears about the dehumanising effects of technology [18].

The *Association of Manhattan Scientists* wrote in 1945 [19] of their feeling of collective responsibility for their role in developing a technology with "potential for great harm or great good" (atomic energy) and of their "special awareness" that it might lead to the "advance of our civilization or its utter destruction". In promoting the capability of robotics and AI towards a largely unknown end, our generation of researchers also have a special responsibility to understand and be outspoken about what the future of robotics might bring and its potential benefits and threats.

References

1. Fellous, J.-M., *From human emotions to robot emotions*, in *AAAI Spring Symposium on Architectures for modeling emotions: Cross-disciplinary foundations* E. Hudlicka and L. Caññamero, Editors. 2004, AAAI Press: Menlo Park, CA. p. 37-47.
2. Heider, F. and M. Simmel, *An Experimental Study of Apparent Behavior.* The American Journal of Psychology, 1944. **57**(2): p. 243-259.
3. Levy, D., *Love and Sex with Robots*. 2007, London: Harper Collins.
4. Kahn, J., Peter H., et al., *What is a Human?: Toward psychological benchmarks in the field of human–robot interaction.* Interaction Studies, 2007. **8**(3): p. 363-390.
5. Lovgren, S., *Robot Code of Ethics to Prevent Android Abuse, Protect Humans*, in *National Geographic News*. 2007.
6. Brown, A., *To mourn a robotic dog is to be truly human* in *Guardian*. 2015: Manchester.
7. Collins, E.C., A. Millings, and P.T. J. *Attachment to Assistive Technology: A New Conceptualisation*. in *Assistive Technology: From Research to Practice: AAATE 2013*. 2013.
8. Metzinger, T., *The Ego Tunnel: The Science of the Mind and the Myth of the Self.* 2009, New York: Basic Books.
9. Tononi, G., *Consciousness as Integrated Information: a Provisional Manifesto.* The Biological Bulletin, 2008. **215**(3): p. 216-242.
10. Qiu, J., *Probing islands of consciousness in the damaged brain.* The Lancet Neurology. **6**(11): p. 946-947.
11. Seth, A.K., *Why fish pain cannot and should not be ruled out* Animal Sentience, 2016. **2016.020**.
12. Dennett, D., *The practical requirements for making a conscious robot.* Philosophical Transactions of the Royal Society of London A, 1994. **349**: p. 133-146.

13. Bryson, J.J. *Crude, Cheesy, Second-Rate Consciousness*. in *Vienna Conference on Consciousness*. 2008.
14. Bostrom, N., *Superintelligence: Paths, Dangers, Strategies*. 2014, Oxford: Oxford University Press.
15. Heylighen, F., *The Global Brain as a New Utopia*, in *Zukunftsfiguren*, R. Maresch and F. Rötzer, Editors. 2002, Suhrkamp: Frankfurt.
16. Future of Life Institute. *An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence*. 2015; Available from: http://futureoflife.org/ai-open-letter/.
17. Kang, M., *Sublime Dreams of Living Machines: The Automaton in the European Imagination*. 2011, Cambridge, MA: Harvard University Press.
18. Szollosy, M., *Freud, Frankenstein and our fear of robots: projection in our cultural perception of technology.* AI & SOCIETY, 2016: p. 1-7.
19. Association of Manhattan Scientists. *Preliminary Statement*. 1945; Available from: https://www.gilderlehrman.org/history-by-era/postwar-politics-and-origins-cold-war/resources/physicists-predict-nuclear-arms-race-.

AISB Workshop on the EPSRC Principles of Robotics

Defending an obsolete human(ism)?

Michael Szollosy, Sheffield Robotics

Introduction

The 2010 EPSRC workshop to devise a set of principles for the responsible development of and research into robots was an important, ambitious and very well-intentioned project. The 'Principles of Robotics' that resulted from these discussions might be regarded as an excellent collection of pragmatic rules governing the building and regulation of robots, and very wisely include versions that can be utilised both in a specifically legal context and also intended for a general audience.

It is clear from reviewing the ESPRC Principles that they can and should serve a vital function in protecting human beings from irresponsible or simply thoughtless research into technologies that could conceivably have very real, very negative consequences for humanity, on a personal, societal or even species-wide level.

However, it is also clear that what is being protected by the ESPRC's Principles is a very specific human being, or at least, a very specific conception of what constitutes a human being. Underlying the Principles are a set of unspoken assumptions about human nature and, consequently, the nature of our relationship with technology. It is absolutely necessary, when engaging in such an exercise, to make such assumptions, and it is absolutely right that those engaged in the workshop did so. (However, it is my contention that there should have been a preamble, setting out those assumptions.) The EPSRC guidelines are meant to protect human beings, and to ensure that robotics research is conducted for the 'maximum benefit of all of its citizens', though what exactly those 'citizens' might look like is a question that is left unanswered, and means that, however laudable their intentions, these Principles are very much already a document rooted in a particular historical and cultural context, and this makes it unlikely that these Principles, in their present form, will endure in the medium- to long-term.

The ESPRC Principles make certain, very specific, yet completely unspoken assumptions as to what constitutes a 'human being'. And the Principles suit thus human being very well. The Principles insist upon a particular relationship between clearly demarcated *human subjects*, on the one hand, who always act as unique and autonomous agents, and robots, which are forever seen as objects, tools to be manipulated by human masters. Humans have the right to design, build and buy robots, and must maintain full legal responsibility for them.

The Principles as they are presently articulated will probably be sufficient in the short-term, perhaps even in the medium-term, to deal with most issues with new technologies that emerge from robotics and computer labs throughout the UK. But however doctrinally or even sentimentally attached to this version of human being we may be, we need to accept that this human being is a transient creature, a relatively new invention, a being, furthermore, that is not presently nor has ever been internally consistent and unitary, and will be constantly re-made and transformed by a whole host of new technologies. These transformative technologies include the biological enhancements and mechanical upgrades championed by posthumanists, transhumanists and others,

but also – more simply, new ideas about the self, society and new ways of thinking about our place in the world, changes which can be brought about not just by roboticists, geneticists, computer scientists, but also by changes in our economic, political and social life more generally.
The human being – or human beings – being tended to in the ESPRC's Principles will not be the first, or final, articulation of what it means to be human.

Which human?

At the (implicit) heart of the ESPRC Principles is a particular human being defined through the last centuries by what has becomes known as *humanism*. This human being is an agent in its own right, a being that is independent and not to be governed by other, metaphysical, or supernatural, forces. This human being is at the centre of European-based legal, ethical, economic and political systems; however, it is vital to remember that 1. this human being is still a relatively new invention and that 2. throughout its life-span, there has never just been one, singular version of this human being, as humanist proponents have liked to imagine that it is.

There is little consensus as to the birth of human being – some would claim the Renaissance, others would say the Enlightenment, and others still would say that the humanist subject became central to the way we think about ourselves only in the nineteenth century. Similarly, some claim that this human being died in the second half of the twentieth century, while others argue that it perseveres to this day. Whatever dates one ascribes to this way of thinking about ourselves, it is certain that this conceptualisation of human beings is an invention, not a given; the humanist human is not 'natural', or even a 'correct' interpretation of our human nature. It is a creature that has come into being due to specific forces and technologies: the ways we relate to each other and the ways in which we relate to things, what we produce (what traditional Marxists call the 'mode of production'), the things we use, and the environment in which we live.

And when we contextualise humanism in this way, and regard its ideas historically, noting how different ideas of what it means to be human (or even 'humanist') have shifted radically over the centuries, it becomes apparent that we are not just speaking of one human being, or one idea of what it means to be human. That humanists themselves differ and disagree over when humanism started, or what it looks like, is further evidence that we are talking about is not a single human being, but many human beings, not a single, inalienable, self-evident human nature, but humans shifting conceptions of themselves in particular contexts. Humans' understanding of ourselves are always *contingent* and *contextual*. Each person can be a member of the public, a citizen, a specialist in different contexts, a consumer, a producer; we can be criminals, patients, clients, taxpayers, stakeholders, students, labourers or management or all of these things at once, or none of them, depending on contexts. And what it means to be a 'criminal', for example, or a 'citizen', or a 'labourer' or a 'man' or a 'woman' is very different today than it was two hundred, or even fifty, or even ten years ago. We can be subject to ever-shifting discourses on law, medicine, education, politics, economics, philosophy, industry, the media and a host of other systems, languages and institutions that seek to define and understand us in slightly different, or radically divergent, ways.

The assumptions that underlie our notion of a solitary conception of what it means to be 'human' are unsustainable under the intense scrutiny of new ways of thinking about the self, and also as new technologies force us to think about ourselves differently. Technologies have always forced the radical transformation of human beings, from the first time our prehistorical ancestors picked up sticks to help in the hunt or the flying shuttle transformed the way cloth was made in the Industrial Revolution. New developments in robotics will exacerbate these processes. Human beings will be more intricately integrated with their tools as sticks become prostheses and as human labour is completely replaced with automated machines. And these developments will create new human beings, and new ways of thinking about ourselves.

However, technology does not always manifest itself in physical entities; technological advances are not always in the shape of new tools or machines: the invention of laws and a legal system were new technologies that had a tremendous impact upon how we construct our human, social selves, the same way that the invention of scientific method, new industrial relationships or Facebook have changed the way we conceive of ourselves, and present that idea of ourselves, to the world. Our twenty-first century technologies – including more advanced robots and AI, but also shifting legal systems, political bodies and systems for ethical living – will be further developments that will radically transform, over time, how we see ourselves and conceive of the very idea of what it means to be human.

In what implicit, unspoken ways, we need to specifically examine, do the ESPRC's Principle indulge in these humanist assumptions about the nature of human beings? The Principles do not address the issue of which human being at which they are aimed. There should most certainly have been an explicit preamble, setting out who this document is intended for, and spelling out the assumptions that underlie the Principles. All such documents – constitutions, charters, treaties or declarations of principles – should state explicitly from the outset those truths that are held to be self-evident, the basis of what is to follow.

In the absence of a clearly defined subject, the Principles offer the usual, familiar humanist conception of the human – the static, homogenous human being that very soon will be made obsolete, if it is not already, by the very technological advances that it seeks to control. There is an overly simplistic conception of the relationship between human beings and their tools: a one-way relationship whereby tools are always the servants of their human masters, and always under the control of an independent human agent. Such a relationship between subject and object, active agent and passive article would always have been naïve. We must consider the ways in which our tools transform humans; not only when that cave-dweller first grabbed a stick but how our new technologies demand fundamental reorganisation of our entire way of life, and insisted how we re-conceive our entire social structure.

To say that our relationship with our tools is not a simple master-servant, one-way relationship is not to say that our tools are our masters. However, we must recognise that *human beings are as much the product of our way of making things as what we make, and how we make it, are decided by human beings*. (There is really nothing controversial in saying this; it is something that Marx recognised over one hundred and fifty years ago, in explaining how a society's *mode of production* defined its social relations, and how individual human beings were then in turn defined by those social relations.) The relationship between human beings and their tools has always been more complex than it is imagined by humanism and in these Principles; we may even speculate that our tools in the future will play an even bigger role in shaping human societies and individuals as they as gain greater degrees of autonomy, as few human beings are 'workers' and the lines between 'biological' and 'machine' are further blurred.

The Principles, therefore, despite noble intentions, attempt to give mastery of future and evolving technology to an obsolete human being.

The conception of human beings offered in the Principles also shares with humanism the illusion of proffering a single, homogenous subject, when in fact that subject is a compilation of multiple – often contradictory – beings. The first paragraph speaks of the promise that robotics 'offers to benefit society', as if 'society' is itself a single homogenous entity. Rather, we should speak of *societies*, different cultures across the globe, or even different cultures within the same community. It is unlikely that advances in robotics and AI will benefit all communities and all nations equally, especially in the short- to medium-term, and principles for the development of robotics should acknowledge this.

It is also unclear which individuals are being referred to in the document; human beings are variously referred to as 'the public', again, as if this is some single, homogenous body. The Principles themselves make it clear that there are a number of different beings that will have specific interests in the development of robotics, and engage with robots in various ways.

- The Principles refer to 'citizens', a subject of a certain political (usually national) body, though it is unclear whether anyone can any longer claim to be a 'citizen' of a discreet, independent nation state, independent of other influences. Will the benefits of and responsibility for robots be only for citizens of a particular nation state, or political body? (perhaps this is complicated as well if increased automation leads to the implementation of a Universal Basic Income in a specific nation state, but not elsewhere.) It is also interesting that the Principles uphold this (obsolete) notion of the nation state with the allowance that robot can be designed as weapons 'for national security reasons'.
- The Principles are adamant that humans alone are 'responsible agents' in law. This may be uncontroversial at present, and one must enter the realms of science fiction to imagine when we might have sentient, conscious robots and AI that would be equal to human beings in the eyes of the law, but this declaration ignores the waters already muddied by autonomous systems, such as self-driving cars, and the challenges they pose to our humanist legal system. Furthermore, we may wonder how technologically-enhanced humans (e.g. cyborgs) may be regarded in law as equally (less? more?) responsible agents.
- The Principles insist upon considerations of privacy, though we can already see that for many people the boundaries of 'self' and 'public' are blurred, and the notion of privacy has been radically altered in such a short space of time. Social media, the promise of 'smart homes', and issues of security have meant that, culturally, we have a very different idea of what 'privacy' means and how it is important to us. This will have tremendous ramifications on the law, and how we view our relationship to the wider world.
- The Principles made a clear distinction between those who 'design' robots, those who 'sell' robots, and those 'consumers' and 'users'. The Principles implicitly accept that the interests of these groups may compete. However, these relations also assume a static conception of our present system of social relations based on capitalist modes of production, a system of social organisation that many would argue is already obsolete (cf., e.g. Paul Mason 2015). It is already evident that as robotics and AI develop, these once-apparently-stable social relations will come under increasing strain and will likely be transformed to something more appropriate to the new possibilities for producing thing and more efficient ways of organising society (e.g. a sort of post-capitalism, as some would have it). As we already blur boundaries between producers and sellers on the one hand and consumers and users on the

> other (e.g. Uber, crowd-sourcing data, Google), these categories already need to be much more flexible than they imagined in a straightforward, simplistic humanism.

It is also worth pointing out that the ESPRC Principles of Robotics are, unsurprisingly, perhaps, very much a European, Christian concoction. Robots are considered to be machines, and therefore merely objects. In the European Christian tradition, such non-living, or even non-human objects, are considered lesser beings on the basis that they do not have a *soul*; that intangible, metaphysical property unique to life or, in most articulation, unique specifically to humans. (This idea of lacking something vitally human lies at the very idea of *the robot*, when the word was first introduced to the world in Karl Capek's 1921 play, *R.U.R.*) Though one could argue that Europe is no longer beholden to Christianity, Europe's (and America's) Christian values are constantly on display, and this assumption is obvious even in contemporary, completely secular European legal and ethical frameworks, including these ESPRC Principles.

By way of contrast, it is worth noting – as many have (e.g. Metzler and Lewis 2008; Lee, Sung, Šabanović, Han 2012) – how differently robots and AI are perceived in different cultural contexts. In Japan, for example, there can be seen a very different assumed relationship between humans sand robots. Some Shintoism and Buddhism are dominate cultural influences in Japan, and both are 'animistic' religions – that is, they believe that all things, including inanimate objects, contain the nature of *kami*, or spirit. These different cultural influences (even in what are now highly secularised cultures) can have a tremendous impact on how we seek to define our relationships with robots and AI. Such influences, so deeply imbedded, are less likely to be transformed too easily by the introduction of new technologies and ideas, but it underlines how the ESPRC Principles are very much located in a very specific cultural and historical context, and how we must be ready and willing to imagine other ideas and relationships not only in the future but right now, if we are to attempt to build an international consensus on principles of robotics.

New humans?

That the Principles 'are not intended as hard-and-fast laws, but rather to inform debate and for future reference' demonstrates the forward-thinking of the delegates to the workshop, but the Principles should be amended to allow movement beyond the narrowly-conceived notion of 'the human' that underlies them in their present state. At present, the humanist subject at the heart of these Principles is a hard-and-fast limit on what can be conceived, because this idea of 'the human' defines all of the relations that are imagined therein. It is necessary, instead, to imagine a different, more pluralistic and flexible human being at the base.

Thinkers can haggle (and often do, *ad nauseam*) about when the consensus supporting the humanist subject fell apart, but it is clear that at some point after the Second World War, with the loss of faith in metanarratives and a new, radical hermeneutics of suspicion (which some have come to understand as 'postmodernism'), the stable humanist subject as it was once understood was not long for this world. But as to what comes next – and it is clear that something does need to come next, for we cannot proceed to construct any sort of frameworks or models without some notion of what it means to be human – there is much less agreement. It is also clear that whatever follows 'humanism' cannot articulate a single notion of 'human' but must be flexible, adaptive and inclusive of many different kinds of human subjects, all interacting in various ways at various times – humans as consumers, humans as producers, humans as designers, as legal subjects, as citizens and subjects of various political entities… Any of these any human being may be at any one time.

We may seek to recreate some principles for robotics based on a human subject that comes after humanism. We may want to call this human being the *post-human*, or simply *posthuman*. However, these terms are complicated and refer to a dizzying array of different ideas and ideologies (even more than were contained under the umbrella term 'humanism' that proceeded it). Without providing a complete summary of the different ideas that can be referred to under these labels, I wish to give some idea here as to the sort of ideas that I believe will be necessary and useful moving forward with a future principles for technological innovation.

Posthumanism can simply mean, philosophically, culturally, that which comes after humanism; this posthumanism, sometimes an anti-humanism, refutes the sort of stable, singular assumptions about the human and human nature that are laid out by humanism. Going a little further, posthumanism accepts the contingence and contexts of conceptions of the human, and replaces a static human nature with something more dynamic and pluralistic. Many of the conceptions of posthumanism, furthermore, include considerations as to how new technological developments are to be incorporated into human experience, transforming both humans and our world.

Posthumanism, or perhaps more accurately, posthumanisms, are not teleological; they do not take as a starting place that the human being that we have arrived at after millions of years of evolution and thousands of years of philosophy – us – is *the* human being, a final, finished, polished product that will now forever remain immutable and unchanging. Posthumanism recognises that our conceptions of ourselves will change and will be transformed, just as they always have been. So a great strength of posthumanism, as it is understood and articulated here, is that there is an in-built flexibility to accommodate such changes, and it is important that any ambitious undertaking, such as laying out a set of principles for defining our present and future relationships with an ever-changing technology, have a similar in-built flexibility.

Some, who are particularly optimistic about the near-emergence of sentient AI, and who might call themselves *transhumanist*, might regard the ESPRC Principles as naively anthropocentric, that they fail to account for the emergence as robots and AI as sentient agents in their own right who deserve (perhaps equal) consideration alongside humans in the creation of any ethical principles. Such an argument would share with what I am advancing here a belief that the ESPRC Principles are already somewhat outdated and too narrow in their conception of what constitutes 'the human', though I am much less optimistic about the imminence of sentient AI, and I do not share the general transhumanist certainty that humans radically transformed by technology (e.g. human beings that are near-immortal) are similarly very near. However, it is not necessary for a specific, sentient AI to emerge for this posthumanist criticism of the ESPRAC's Principles to stand. Even if we invent no new robots and make no new strides in artificial intelligence – which is very unlikely – it is almost certain that we human beings will continue to reinvent the other systems and institutions that define who we are, thus transforming human beings and necessitating a new, more flexible set of principles to define our relationship with robots and AI.

The writers of the Principles intend it to be a 'living document', not 'hard-and-fast' laws but the basis for future debate and reference, which is exactly what it needs to be. But it is hard to see how any principles for robotics can have the flexibility to respond to the challenges that will be presented by our new technologies when it takes as a starting place such a rigid and already obsolete human subject at its heart.

It is interesting to see that the preamble to the Principles mentions the ubiquity of Asimov and his Three Laws. Because even though Asimov's Laws are dismissed – correctly – as inadequate, because they are fictional and so do not address 'real life' and cannot be used in practice, there is nevertheless something in Asimov's writings that the ESPRC could have taken as inspiration: the ability to imagine different worlds, populated with different sorts of human beings . Human beings are always undergoing processes of re-invention, but with the advances in robotics and AI that are likely just around the corner we might speculate that we are on the brink of an even more radical transformation in how we see ourselves and how we relate to our technologies. It is absolutely vital, therefore, that we seek to create principles for robotics that will be capable of accommodating these changing relationships, and both enable and put limits on various directions of development. If such principles are to endure, and not be a relic of an outmoded way of seeing the world and ourselves, we need to anticipate how human beings will be transformed by new developments in robotics and AI. We will need to think imaginatively about the sorts of robots we will create but *also* the sorts of people we will become, and if we seek to craft principles for the benefit of our societies, we need to have a better understanding of what those societies, and the human beings that populate them, will look like.

References

Lee, Sung, Šabanović, Han. 2012. Cultural design of domestic robots: a study of user expectations in Korea and the United States.
Mason P. 2015. *PostCapitalism: a guide to our future*. London: Allen Lane.
Metzler and Lewis. 2008. Ethical views, religious views and acceptance of robotic applications: a pilot study. AAAI. 15 – 22.

Regulating Robot Towns:
Reflections on the Principles of Robotics
From the New Far-Side of the Law

Aurora Voiculescu
Centre for Law & Theory, University of Westminster

> *"They asked me where I'd choose to run, which favoured? Ups? Or Downs?*
> *Where robot mice and men, I said, run round in robot towns.*
> *But is that wise? For tin's a fool and iron has no thought!*
> *Computer mice can find me facts and teach me what I'm not.*
> *But robot all inhuman is, all's sin with cog and mesh.*
> *Not if we teach the good stuff in, so it can teach our flesh*
> *[…]*
> *As man himself a mixture is, rambunctious paradox,*
> *So we must teach our mad machines: stand up, pull up your socks!*
> *Come run with me, wild children/men, half dires and dooms, half clowns.*
> *Pace robot mice, race robot men, win-lose in robot towns."*
> Roy Bradbury[1]

The Principles of Robotics initiative stems largely from a reflection of the extent to which robots already affect our lives and of the even greater extent to which it is expected that they will affect it in the 'robot towns' of the relatively near future. Whether the initial regulation related to this transformative technology will take the shape of *soft*, guiding principles, of *hard* domestic legal instruments, or even of complex international treaties is a challenging yet, at this point, a secondary issue. The primary question is rather a (legal-) normative question, aimed at delineating clear boundaries of the human/robot co-existence; addressing the normative dynamics of causality and responsibility; trying to identify the *locus* or *loci* of *mensa* and *actus* in processes and, dare we say, *relationships* that may well prove to become more and more complex with the advancements of science and technology.[2]

Stemming from this need for normative introspection (into our social psyche much more than anything else), this paper is an invitation to reflection on the

---

[1] Ray Bradbury, *Where Robot Mice and Robot Men Run Round in Robot Towns: New Poems, Both Light and Dark* (New York: Random House Inc, 1977).

[2] Aurora Voiculescu, "Human Rights Beyond the Human: Hermeneutics and Normativity in the Age of The Unknown," (forthcoming).

proposed Principles of Robotics (covering 5 principles and 7 High-Level Messages) coming out of the multi-disciplinary expert-informed EPSRC and AHRC Robotics Retreat in 2010. The complexity of issues to cover is such that these reflections can only aim to engage with what is proposed, with the text offered for reflection, prising out some of the possible meanings or interpretations of such texts. Such analysis is proposed as essential for preparing the ground for further discussions, and finally, for embarking on any eventual regulatory processes.

Principles in Search of a Definition

Reflecting on the existing principles does invite one, first of all, to reflect on what a robot is and on whether the definition that one should settle for,[3] and therefore the type of entities that one should aim to regulate, should be an answer to our state-of-the-art in technology or a reflection of our state-of-the-(technology in) art. In other words, on which point on the spectrum between science and science fiction should we place ourselves when designing norms and evaluating their effectiveness? How far into the future should one look, when the future for which we regulate is so far that we can only speculate as to its existence, while at the same time we are galloping towards that very future at an ever-greater speed?

Law/regulation, hard or soft, requires definitions. The principles here under discussion do not immediately send to one. A robot is defined by some as 'a machine capable of carrying out a complex series of actions automatically' or, differently nuanced, 'a mechanical or virtual artificial agent, usually an electro-mechanical machine that is guided by a computer program or electronic circuitry'.[4] Various, more or less workable distinctions are also put forward, more notably between industrial and service machines, between highly autonomous machines and cognitive computer programmes, between embodied and dis-embodied cognition entities, etc. NASA itself uses a rather mundane and imprecise language, very unhelpful for the regulator, defining robots as "machines that can be used to do jobs". Some robots, NASA's formulation goes on to add, 'can do work by themselves. Other robots must always have a person

---

[3] A definition never being value-neutral, always establishing the 'in'-s and the 'out'-s following a more or less declared value-laden path. (See Alan Norrie in Voiculescu 2000)

[4] Merriam-Webster Dictionary, "Definition of 'Robot,'" accessed February 10, 2016, http://www.merriam-webster.com/dictionary/robot entry: robot.

telling them what to do'.[5] Such a variety of formulations create a regulatory puzzle and will make any normative statements difficult to follow and/or easy to escape complying with.

While agreeing that there is no agreed definition *per se*,[6] some put forth a number of features that a robot would have, features that, from a regulatory (and not only) perspective, are themselves in need of definitions: *sensing* the surroundings (having inbuilt 'awareness' of its environment); *movement*, whether rolling, walking, thrusting, or maybe even *just* data-conveying; *energy*, being able to power itself in ways that will depend on what the purpose of the robot is; *intelligence*: being provided with 'smarts' by its programmer, having the capacity to evaluate surroundings, circumstances, complex information. [(-the more a machine is capable of independently interacting with a dynamic world, the more advanced that machine/robot is, the AI aiming, among others, precisely towards this)]

So, a robot is defined more specifically as a system, a machine that "contains sensors, control systems, manipulators, power supplies and software all working together to perform a task". According to such a perspective, "[d]esigning, building, programming and testing a robot is a combination of physics, mechanical engineering, electrical engineering, structural engineering, mathematics and computing. In some cases biology, medicine, chemistry might also be involved". If the student in robotics may actively engage with all these disciplines "in a deeply problem-posing problem-solving environment",[7] some could rightly say, that regulating robots and 'robot towns' requires a similarly complex interdisciplinary engagement with most if not all of these fields.# For the normative discourse (whether hard regulatory or soft-principled), the fact that many of these definitions have a number of points in common is not enough. A definition that is sufficiently precise, yet dynamic enough to capture the essence of the socio-technological phenomena is therefore needed for opening the robotics principles to further development and problematisation. This need relates to perspectives such as Andra Keay's, who speaks about robots as "… *an environment*; too large for us to look at as an item". While inevitably linked to the progresses of technology - "[what] we call a robot today is more sophisticated

---

[5] NASA, "What Is Robotics?," *NASA Knows*, May 18, 2015, http://www.nasa.gov/audience/forstudents/k-4/stories/nasa-knows/what_is_robotics_k4.html.

[6] H. James Wilson, "What Is a Robot, Anyway?," *Harvard Business Review*, April 15, 2015, https://hbr.org/2015/04/what-is-a-robot-anyway.

[7] Ibid.

than what we called a robot in the '80s" says Keay - it is also true that it is more than that. "It has always been *an identity issue*" says Keay.[8]

It should be said, however, that identities and classifications have always been problematic and problematised when part of regulatory initiatives, whether these had to do with humans or non-humans alike. Law, in particular, always ends up transforming any identity in a legal fiction that often has very little to do with any other physical or scientific dimension of that entity.[9] At the same time, law - in its widest sense of socially backed normative imperatives - has always thrived on definitions. The absence of a 'working definition' of a robot appears therefore both as a witness to the challenges of pinning down technology in its rush, and as a reflection of a possible weakness to be addressed in the proposed document.

Last but not least, in addition to the issue of the absence of an agreed working definition (that would be as soon contested and problematised, of course), there is also the acknowledgement that 'definitions are never neutral'. This idea was advanced some decades ago by Larry May when reflecting on defining the responsibility of non-human collective agency (an otherwise not irrelevant legal innovation). Definitions, advances May, create 'pseudo-unities' that are proposed as facts, while in reality, they set up oppositions that 'arbitrarily separate those who are included and those who are excluded from a shared conceptualisation or practice'.[10] This assertion will become more and more evident once the spectrum of available options between robot and AI becomes enlarged.[11]

Whether anticipated with dread or with excitement, the challenge of regulating the multiple dimensions of the human-robot interactions are multiple. A number of issues put forward for reflection in relation to the given five principles are mentioned briefly here:

---

[8] See interview with Andra Keay, founder of Robot Launchpad and Managing Director of Silicon Valley Robotics in Signe Brewster, "What Is a Robot? The Answer Is Constantly Evolving," July 5, 2014, https://gigaom.com/2014/07/05/what-is-a-robot-the-answer-is-constantly-evolving/ (Author's emphasis).

[9] David Fagundes, "Note, What We Talk About When We Talk About Persons: The Language of a Legal Fiction," *Harvard Law Review* 114, no. 6 (2001): 1745–68.

[10] Larry May, *Sharing Responsibility*, New edition 1996 (Chicago: University Of Chicago Press, 1992), 171ff.

[11] Kenneth Grady, "Artificial Intelligence: Be Afraid, Be Very, Very Afraid (Or Not)," *SeytLines: Changing the Practice of Law*, December 31, 2014, http://www.seytlines.com/2014/12/artificial-intelligence-be-afraid-be-very-very-afraid-or-not/.

First of all, there is a need for a somewhat clearer perspective as to what it is that one is regulating: The above mentioned absence of an agreed definition aside, the principles put forward for discussion reveal the potential of confusion as to the actual agenda: 'regulating robots in the real world' has a double sense, full of pitfalls. If one 'regulates robots' themselves, the text brings in an implied agency that may be taken for granted in contexts where it may be undesirable (although this interpretation is clearly contradicted by some of the principles, notably by Principle no. 2 and 5). The second meaning, more in line with what the five principles themselves reveal, could aim at 'regulating the creation and use of robots'. The choice of this interpretation should be more explicit throughout the formulations, avoiding regulatory confusion.

Principle no 1: *Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interest of national security.* Such a principle is promoted in order to limit the harmful ways in which robots can be used. However, while the aim appears clear, the way of expressing the principle can be revealed to be a rather ineffective one. First of all, the principle may appear, in its second part ('*…except in the interest of national security*'), as an indirect acknowledgement that regulating the use of robots for violence against humans is very difficult, even impossible, to achieve. The danger of using sophisticated robots for violent, aggressive goals is so big that, like the use of nuclear technology for weaponry, or the use of chemical weapons, it should be strongly discouraged. If not totally forbidden (a clearer definition of what machines one is talking about would, again, be very helpful), the 'national security' argument should be more precisely and narrowly defined as well as the types of 'harm' that one can allow/design robots to do.

The first part of this principle is, however, even more puzzling. First of all, one may find the starting statement '*robots are multi-use tools*' as virtually a restriction that does not serve an actual purpose. It is unclear why a robot has to be 'multi-use' in order to be safe or, conversely, in what way an otherwise deadly robot may become any less deadly if designed as 'multi-use'. This relates to the next part of the principle: *'robots should not be designed solely or primarily to kill or harm humans*'. In order to bring 'killer robots' in compliance with the letter of this part of the principle, it would suffice to also teach the 'killer robots' to make pancakes or knit woolly socks. This is what, in the legal normative perspective, one would call a 'creative-compliance loophole'. In order to identify and use such a loophole, a legal eye needs to look no further than to a literal interpretation of the text. Nevertheless, the literal interpretation is one of the primary rules of interpretation in law, when the interpretation in line with the 'spirit of the rule' may not be a convenient one. The explanations given to this principle in the 2010 original document do not seem to really address this rather basic approach to interpreting rules and its consequences in this particular context.

The commentary to this principle appears to imply other potential pitfalls for normative reasoning. First of all, as mentioned with respect to 'multi-use tools', there is an effort to put forward the idea that robots are tools like any others. In order to pursue with this logic, equivalences are sought at any cost. Comparing a robot with a knife or a gun used for different, both relatively benign and criminal purposes, does not cover for the inconsistency that there are tools, including weapons, for which one can think of no other purpose than the prima facie one.

Principle no. 2: *Humans not robots are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws and fundamental rights and freedoms, including privacy.* The comments to this principle seem to add more confusion than clarity. First of all, a relatively small matter, there is the presumption that 'no one is likely deliberately to set out to build a robot which breaks the law'. This puts forth a presumption that has no foundation in the real world of 'deviance and defiance', as revealed by socio-legal studies among the wide population at large as well as among the white collars.[12] Secondly, and more importantly, the way responsibility is assigned through this comments appears to ignore both the way 'law thinks'[13] as well as the way robots may fail to 'achieve goals and desires that humans specify'.[14]

As an additional element here, it should be also mentioned that, in the absence of a clear working definition, AI, 'learning robots', etc, are all concerned with these principles and their parameters. Their mechanics, however, may well be more complex than the law/normative discourse can handle in the absence of a clear definition#. Robots and AI machines may well learn to deal with 'exceptions' before the law learns to deal with 'differences'.[15] Equally, other disciplines seem to indicate that numbers (in this particular case, 'programming') may well be more than just that, numbers being inalienably complemented by/associated

---

[12] Some useful though loose examples Ryan Mathews and Watts Wacker, *The Deviant's Advantage: How Fringe Ideas Create Mass Markets* (Random House, 2010); Kelly Fisher, "The Psychology of Fraud: What Motivates Fraudsters to Commit Crime?" (Social Science Research Network, March 31, 2015), http://papers.ssrn.com/abstract=2596825.

[13] Gunther Teubner, "How the Law Thinks: Toward a Constructivist Epistemology of Law," *Law and Society Review* 23, no. 5 (1989): 727–57.

[14] See for instance the orthogonality theory in Nick Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," *Minds and Machines*, 2012, http://www.nickbostrom.com/superintelligentwill.pdf.

[15] This is built on the ideas of identity and difference, rom Leibniz to Kant... *Gilles Deleuze - Le Point de Vue (Le Pli, Leibniz et Le Baroque) 1986 FRA Sub ITA*, 2012, http://www.youtube.com/watch?v=2ZrA_7ewQGs&feature=youtube_gdata_player.

with a *narrative* movement that, one could say here, may be construed differently by the machine than by the human, yet may still be construed by it.[16]

Principle no. 3: *Robots are products. They should be designed using processes which assure their safety and security.* This principles raises issues of distinctions in the legal self-defence debate; tort issues, assimilation of responsibility, etc., issues that depend very much of the context and of the extent to which, as mentioned above, the regulator will be willing to construe equivalences between robots and other types of tools or between robots and other types of property items.

Principle no. 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*. Again, this principle raises several issues. The ability of people to communicate with sociable machines is one of the avenues pursued intensely in the scientific community.[17] While the aim is not one of deceiving, creating the context for this type of communication to take place will raise many regulatory problems. The borderline between a crafty illusion and a deceptive interface will easily be filled by the human emotions. Part of what we are after in creating 'service robots' that will read bedtime stories to our children, care for our elderly disabled father, is emotional engagement. The scientific evidence is there showing that creating the conditions for emotional engagement may well transcend the 'behind the veil' appearance that may well remain, in fact, the one of a 'tin'. From here to the marketing manipulation of the emotional engagement remain just a few steps that the regulatory environment will find challenging to accompany. The comments published with this principle hint to welcome further reflection.

Principle no. 5: *The person with legal responsibility for a robot should be attributed.* This principle appears clear enough in the context in which we are a long way away from AI capable of technically fulfilling the requirements for such responsibility. At the same time, taking into account the complex elements and fields that enter the make-up of a robot (see above when discussing issues related to the definition), challenges to the regulatory approach will remain from the point of view of prising apart the various degrees of responsibility, when things go wrong. Having a 'registered keeper', bearer of responsibility, is only part of the solution. The responsibility bearing entity will require further

---

[16] Marcus du Sautoy, *Narrative and Proof: Two Sides of the Same Equation? | TORCH* (TORCH, The Oxford Research Centre in the Humanities, 2015), http://www.torch.ox.ac.uk/narrative-and-proof-two-sides-same-equation-0.

[17] See for instance Cynthia Breazeal, "Emotion and Sociable Humanoid Robots," *International Journal of Human-Computer Studies* 59, no. 1–2 (July 2003): 129ff, doi:10.1016/S1071-5819(03)00018-1.

reflection, as will the type of harm(s) that may be attributed to such entities, taking into account, as some suggest, that for the first time 'the promiscuity of data' is lately combined with 'the capacity of doing physical harm'.[18]

Instead of conclusions

Closing these notes full circle, with another of Bradbury's poetic images, one can say that, insofar as robot mice are concerned, it is quite likely that in the context of market powers and scientific capability deregulation combined, we will most surely first 'jump off the cliff and build our wings on the way down'. In such circumstances, what one can do is to make sure of being relatively prepared for this jump by already interrogating and problematizing our relationship with science and technology and by intensifying our reflection on living in robot towns. Addressing 'tools', 'products', 'artefacts' and 'agents', one needs to take in St Augustine's reflection on the intricate connection between language and interpretation as a path to revealing a deeper, existential level of *self-understanding*. The way we think normatively about human-robot *interaction*[19] will say as much about the robot as about the human. To borrow from Ginabattista Vico's 1725 *New Science*,[20] we need to bear in mind that our thinking about robots is rooted in a given cultural context. This means that, in reflecting about the normative parameters of robot towns, the social scientist will not deal with a field of idealised and putatively 'subject-independent objects', but will investigate a world that is, fundamentally, her own. The process of regulating robots is, therefore, a process of self-understanding, rooted in a given historical context and practice. An understanding that does not culminate automatically in neat, normative, law-like propositions.

---

[18] Ryan Calo, "Robotics and the Lessons of Cyberlaw," *California Law Review* 103 (2015): 513.

[19] Be this interaction understood on Latour's spectrum of facts and agency; Bruno Latour, "How to Talk About the Body? The Normative Dimension of Science Studies," *Body & Society* 10, no. 2–3 (June 1, 2004): 205–29, doi:10.1177/1357034X04042943; Bruno Latour, "Body, Cyborgs and the Politcs of Incarnation," in *The Body: Darwin College Lectures*, ed. Sean Sweeney and Ian Hodder, 2002, 127–41.

[20] Giambattista Vico, *The New Science of Giambattista Vico* (Cornell University Press, 1744).

Commentary on responsibility, product design and notions of safety

Paula Boddington, Department of Computer Science, University of Oxford.

The comments here relate mostly to rule 2 and to rule 3. I suggest that the principles could be made more specific to the context of the implementation of robots, and that key notions such as 'safety' could be elaborated and perhaps extended, or the precise way in which the term is used made clearer.

Some of these points are illustrated by considering in broad terms the use of robotics in nursing/ elderly and social care contexts.

Rule 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.

If humans are responsible agents, but robots are not, this implies that wherever robots are used to replace humans or part of human agency, then the responsibility attributions formerly given to the human agent or human actions, are then either displaced into a wider system, or perhaps, overlooked. The displacement may result in shifts in how the responsibilities and accountabilities are understood, and these may be unexpected and complex.

Robots will be used within a system of human agents and behaviours. Such systems may be formalized with clearly expressed notions of responsibility and accountability, for example within a hospital setting (albeit that there may be elements of such systems which are not completely understood or formalized with complete adequacy); or they may be informal, for instance within a home setting of care. In fact even in such informal family or community settings, social research finds that there may be strong local cultures and values regarding lines of responsibility and accountability.

For an example of how responsibilities and accountabilities may be displaced, if a robot takes over some of the roles of a health care assistant within a ward setting, then responsibilities then may be displaced in a variety of possible ways to different actors within the system of health care management. These responsibilities may also change, for example, tasks which were understood in one way may come to be seen or example as more technical rather than as managerial. What might previously have been seen as a failure of conscientiousness in an employee, for instance, might come to be seen as difficulty in understanding or operating machinery. There may be wide-ranging repercussions.

Tracing and understanding such lines of responsibility and accountability may be complex. A question arises as to whether this is purely the task of those in charge of the setting where robots are used, or whether the designers of the robots may have some responsibility in assisting with those who will be working alongside the robots to understand these issues.

Rule 2 talks of complying with existing laws and fundamental rights and freedoms, including privacy. However, in addition, within certain settings, there will be more specific and local protocols and practices which it will be desirable that robots comply with. It may be worth stating explicitly this in the rules.

For instance within the NHS there are standards of care which aim to deliver person-centred care and to treat patients with dignity. Such statements of standards are central to the provision of good health care. These refer to abstract ideas and richly articulated notions such as what it is to treat someone as a person with dignity. Working out how the use of robots impacts upon such rich and contextualized values may be very important, yet may be a harder question than simply that of compliance with law and with responsibilities as laid down within the law. The development of robots which then honour and work within such practices as 'person centred care' may then require careful interdisciplinary work and it may be worth considering explicitly stating the desirability of addressing these matters in the design and use of robots, unless it is assumed that this is widely understood.

Rule 3: Robots are products. They should be designed using processes which assure their safety and security.

There may be a need to articulate what is meant by 'safety' in this context. Does this simply refer to physical safety? And does it refer to safety in terms of the immediate operation of the robot, or to effects of the use of robots further along the system within which they are used? In which case, how are responsibilities for safety to be shared out between robot designers, and those who use the robots within a workplace or other setting?

This rule states that products should be safe. However, although rules of ethics are often formulated in terms of harm avoidance, it is less than aspirational to seek to design products which are simply 'safe'. Good design which fulfills important human needs goes beyond notions of mere safety, but care should be taken that they are fit for purpose.

Where robots are replacing or extending human agency, the task that the robot is undertaking may not be altogether transparent. Especially within a service setting, there may be multiple human transactions of significance in simple task, such as the communication of caring which occurs through routine use of language, chat, and body language. Having 'mundane' tasks taken over by robots may then free up human workers to have more time for caring tasks which only human interaction can provide. Conversely, robots may be designed so as to replace some of these aspects of a task. However, this may possibly run foul of rule 4, which states that 'robots should not be designed in a deceptive way so as to exploit vulnerable users'.

Issues with safety then arise because discovering how the use of robots might disrupt, or potentially even improve, certain possibly hidden aspects of tasks which robots take over, may involve considerable analysis and research. In particular, in a ward setting, seemingly 'mundane' caring tasks which may be nonetheless extremely important to the health and wellbeing of patients, are often carried out by staff working at lower grades. Hence there is a possibility that such important work may not be acknowledged or recognized. Wards form extremely complex systems of social interaction, and discovering how robots fit into such settings may involve extremely careful analysis, using a variety of expertise. This rule 3 may be understood to imply such work and to imply that safety and security are understood in this wide manner; but it might be worth considering explicitly pointing out that safety needs to be considered in a broad remit. This also then raise again the question of shared responsibilities between teams of those designing and manufacturing robots and those who will be working with them. It may not look as if missing out certain routine aspects of caring that humans may give yet robots may not, is an issue of 'safety', but these can have impacts upon recovery and health and hence if so should be seen as an issue of safety. Of course it is possible that robots may improve such matters.

Attention to safey will of course include looking at issues of safety of the use of robots within a wider system. For example, a common and serious problem within hospital settings for elderly and vulnerable patients is dehydration. This can have serious health implications leading for instance to confusion. Sometimes dehydration is worsened by difficulties in reaching and managing drinks. Suppose a robotics system could be designed to assist such patients drink. This system may work with complete safety and reliability in terms of its immediate use, for instance, never malfunctioning in ways which give the patient too much too quickly, and never spilling drinks or hitting the patient.

However, such a system could potentially have large negative consequences within a particular context. Increased hydration could cause increased rates of bedwetting in some patients. This could lead to hospital acquired incontinence, a common problem for elderly patients especially those with dementia, as staff fit catheters to prevent instances of incontinence. Patients with hospital acquired incontinence rarely achieve continence again. This results in ‘bed blocking’ since patients often then have to be found accommodation in facilities which can care for their needs. Increased hydration may also lead to more instances of patients getting out of bed to go to the toilet and hence having increased falls. This can also have dire consequences.

When rule 3 talks of ‘safety’, is it clear that this does or does not include considerations of how a robot might function within a wider system of work? Is it clear in the robotics principles what responsibility those working in robotics have, or whether it is the responsibility of hospital managers and other staff to identify such potential implications of the use of robots?

Contribution of Roeland de Bruin and Madeleine de Cock Buning to the AISB Workshop on Principles of Robotics, April 4th 2016, Sheffield UK

1. Introduction

It is five years since the publication of the EPSRC's principles of robotics developed by a panel of distinguished British robotics and AI experts at an EPSRC/AHRC funded retreat.[1] The principles, which were aimed at "regulating robots in the real world", were stated in the form of five "rules" and seven "high-level messages". The principles have indeed had significant impact in UK robotics research, and continue to provoke substantial debate. Since presently public concern about the development of robot technologies is heightening we consider useful to revisit the principles to consider their continued relevance according to the following criteria.

Our contributions focuses on the second principle:

Principle 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.

In fact this second principle of the EPSRC's principles of robotics is twofold. On the one hand the principle deals with responsibility - including liability - for the actions of the robot, on the other, the principle entails methods of machine design that can aid with the compliance of existing laws & fundamental rights & freedoms, including privacy.

Since both liability and design form the backbone of the introduction of robotics technology, as for instance incorporated in autonomous intelligent cars in our society, we will test this twofold principle by focussing on the current development and deployment of autonomous intelligent cars. Whether this second EPSRC principle can be considered as future proof, will be tested against three criteria:

1. *Validity*—is the principle correct as statements about the nature of robots, robot developers, and the relationship between robots and people, or is it ontologically flawed, inaccurate, out-dated, or misleading.
2. *Sufficiency/generality*—is the principle sufficient and broad enough to cover all of the important issues that might arise in the regulation of the robotics in the real-world or are significant concerns overlooked.
3. *Utility*—is the principle of practical use for robot developers, users, or law-makers, in determining strategies for best practice in robotics, or legal standards or frameworks, or are they limited in their use by lack of specificity or through allowing critical exceptions.

---

[1] https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

2. State of the art

2.1 State of Autonomous Intelligent Cars (AICs)

Before putting the principle to the test we will shortly introduce the state of the art of AICs. Currently consumer cars are increasingly being equipped with technology that assists in certain aspects of driving. Examples of such technology include lane keep assistance, emergency braking, parking assistance and adaptive cruise control. In the near future, higher levels of car automation will become available, eventually leading to the introduction of fully autonomous vehicles.

Also now some cars are already equipped with certain forms of automation. There are even prototypes available that can drive without a human operator. Google is currently pioneering in self-driving car technology, and has put a fully functioning AIC prototype to road tests in Bay Area, California in early 2015.[2] Also in the European Union, car manufacturers concentrate on the development of AIC technology.[3] Scania is testing "Platooning": a road train of self-driving trucks which were autonomously following a human controlled truck heading the convoy was deployed on the Dutch roads.[4] Volvo planned to deploy 100 cars which should be able to take over all aspects of driving in Sweden by 2017[5] and in Germany, a part of the A9 Autobahn between Munich and Berlin is reserved for the extensive testing of autonomous vehicles in the coming years.[6]

A definition of Autonomous Intelligent Cars consists of three elements. *Autonomy* relates to the level of human intervention necessary for operation, which can be seen as a spectrum: a lower need for human intervention implicates a higher level of autonomy. *Intelligence* relates to the ways in which a system can perceive its surroundings, and is able to adapt behaviour to changing environments. It includes the ability to learn, to process complex information and to solve problems.[7] *Cars* are motorised vehicles, used for the transportation of goods and/or people and for carrying out services.

---

[2] Wikipedia, "Google driverless car", available on the Internet at <http://en.wikipedia.org/wiki/Google_driverless_car> (last accessed on 17 March 2015), referring to Matt O'Brian, "Google's 'goofy' new self-driving car a sign of things to come", 22-12-2014, available on the Internet at <http://www.mercurynews.com/business/ci_27190285/googles-goofy-new-self-driving-car-sign-things> (last accessed on 28 January 2016).

[3] See for instance <https://www.media.volvocars.com/us/en-us/media/pressreleases/145619/volvo-car-groups-first-self-driving-autopilot-cars-test-on-public-roads-around-> (Volvo), <http://www.pcmag.com/article2/0,2817,2387524,00.asp>, (Volkswagen) <http://www.bbc.com/news/technology-25653253> (BMW) (last accessed on 28 January 2016).

[4] See <http://www.scania.nl/about-scania/media/platooning/> (accessed 20 March 2015).

[5] Alecander Stoklosa, "Volvo Has a "Production-Viable" Autonomous Car, Will Put It on the Road by 2017", available on the Internet at <http://blog.caranddriver.com/volvo-has-a-production-viable-autonomous-car-will-put-it-on-the-road-by-2017/>. (last accessed 20 March 2015).

[6] Stephen Edelstein, "Germany plans autonomous car test program on high speed autobahn", 28 January 2015, available on the Internet at <http://www.motorauthority.com/news/1096521_germany-plans-autonomous-car-test-program-on-high-speed-autobahn> (last accessed 20 March 2015).

[7] See Madeleine de Cock Buning, Lucky Belder & Roeland W. de Bruin, Working paper: "Mapping the Legal Framework for the introduction into Society of Robots as Autonomous Intelligent Systems", at p. 3-4, available on the Internet at <http://www.caaai.eu/wp-content/uploads/2012/08/Mapping-L_ N-fw-for-AIS.pdf> (last accessed on 28 January 2016) and the references to Samir Chopra and Laurence F. White, *A Legal Theory for Autonomous Intelligent Agents* (Ann Arbor: University of Michigan Press 2011) at p. 10 (autonomy) and Collin R. Davies, "An evolutionary step in intellectual property rights – Artificial intelligence and intellectual property", 27 *Computer Law & Security Review* 2011, at p. 601-619 (intelligence); and by the same authors the chapter "Mapping the Legal Framework for the Introduction into Society of Robots as Autonomous Intelligent Systems", in Sam Muller et al (eds.), *The Law of the Future and the Future of Law*, series 2012 (De Cock Buning, Belder & De Bruin 2012), pp. 195-210.

AICs can contribute to finding solutions for challenges our society is currently confronted with. Road safety will increase dramatically when 'human error' is taken away as a factor in the causation of accidents. AICs could significantly reduce the risks of car accidents since 93% of traffic accidents are caused by human failure,[8] leading to 1.3 million deaths and 50 million serious injuries worldwide per year.[9] Besides contributing to road safety, AICs can lead to more efficient use of the road network, reduce CO2 emissions and assist in improving the mobility of disabled people.[10] The introduction of AICs could thus provide answers to reduce currently manifest risks that are the result of technological innovation in the past decades.[11]

However, not everyone is optimistic about a driverless future. It is stated that while AICs could be beneficial to road safety, other risks will follow from the introduction of autonomous vehicles. AICs will be vulnerable to hacking for example. Also, business models and employment in taxi and transportation markets will change significantly while drivers may eventually become obsolete after the autonomisation of driving.[12] Furthermore, accident risks could increase when autonomous and non-autonomous cars co-exist on the same roads.[13]

2.2 State of the law

Sufficient certainty about legal status is essential for growth in and societal acceptance of consumer technology. Uncertainty causes the opposite. Could in that case the machine be the answer to the machine? Below we will briefly discuss the liability issues currently challenging the introduction and deployment in society of AICs and touch upon possible *technology-of-evidence* solutions for some of these challenges that might involve privacy by design.

Liability

Current regulation in the EU addressing responsibility and liability for damage that might be caused by AICs pose challenges in terms of innovation in the field of AICs and societal acceptance thereof. On the one hand producers of AICs fear that under the Product Liability Directive (PLD) they can be easily

---

[8] Bryant Walker Smith, "Human error as a cause for vehicle crashes", 18 November 2013, available on the Internet at <http://cyberlaw.stanford.edu/blog/2013/12/human-error-cause-vehicle-crashes> (last accessed on 28 January 2016).

[9] OECD, "OECD Factbook 2013: Economic, Environmental and Social Statistics", 2013, available on the Internet at <http://www.oecd-ilibrary.org/sites/factbook-2013-en/06/02/03/index.html?contentType=&itemId=/content/chapter/factbook-2013-50-en&containerItemId=/content/serial/18147364&accessItemIds=&mimeType=text/html> (last accessed on 28 January 2016), also cited in Gillian Yeomans, "Autonomous Vehicles – handing over control: opportunities and risks for insurance", , available on the Internet at <https://www.lloyds.com/~/media/lloyds/reports/emerging%20risk%20reports/autonomous%20vehicles%20final.pdf> (last accessed on 28 January 2016)(Yeomans 2014) at p. 5.

[10] See for example Yeomans 2014, at p. 5. Also Anne Pawsey, "Autonomous Road Vehicles", September 2013 at p. 1. Available on the Internet at <http://www.parliament.uk/briefing-papers/post-pn-443.pdf>, (POSTnote 2013); Robolaw 2014, at p. 42.

[11] Pollution, climate change, societal exclusion of 'weaker parties', and high accident risks on the (European) roads can all be seen as the outcome of the modernization and individualisation processes that took place in the past century. These side effects must now in turn be dealt with.
See for the identification and a study on the concept of *risk society* by Ulrich Beck, his book *Risk Society, Towards a New Modernity*, London: Sage Publications 1992.

[12] See for example Scott Le Vine & John Polak, "Automated Cars: A smooth ride ahead?", February 2014, at p. 14, available on the Internet via <http://www.theitc.org.uk/docs/114.pdf> (last accessed on 28 January 2016).

[13] See Wayne Cunningham and Antuan Goodwin, "Six reasons to love, or loathe, autonomous cars", 8 may 2013, available on the Internet at <http://www.cnet.com/news/six-reasons-to-love-or-loathe-autonomous-cars/> (last accessed on 28 January 2016).

held liable for damage caused by AICs that are defective, which would have a chilling effect on innovation. [14] Whereas on the other hand the current framework on product liability does in fact not provide an easy toolkit for consumers to hold AIC manufacturers liable for defects in their products at all. A rather heavy burden of proof rests at consumers to establish that there was actually a defect in the AIC, as well as on the causal relationship between defect and damage that has occurred. Providing evidence will be more complex when autonomy and intelligence in cars increase, for victims will have to conduct an in-depth (technological) analysis of *inter alia* the (original) software, the updates and the operational data an AIC is equipped with, in order to establish the precise cause of an accident. At the same time, manufacturers have ample opportunity to defend themselves against liability claims. When confronted with AICs, the PLD does not optimally protect the interests of consumers by providing them easy means to get remuneration for damage they suffered caused by defective AICs from manufacturers.

Room for improvement of current legislation is furthermore formed by the different non-harmonized European regimes on liability for motor vehicles. There are to date 28 different frameworks in place in the European Union. For instance French ‘Loi Badinter’[15] imposes a *strict no-fault* liability regime in order to assess whether or not the driver or the custodian of a car is to remunerate damages of victims (other than the driver)[16] of accidents in which motor vehicles are involved. Liability can only be exonerated, if the driver (or custodian) can prove a *faute inexcusable* by the victim.[17] The Netherlands’ ‘Wegenverkeerswet’ appoints (semi-strict) liability to the owner or keeper (note: rather than the driver or a custodian) of a motor vehicle that is involved in an accident where damage occurred to non-motorized road users.[18] At least 50% of the damage suffered needs to be remunerated, unless *force majeure* can be proved.[19] In the United Kingdom, negligence rules are applied to establish whether a driver of a motor vehicle can be held liable. In such cases there is no strict liability regime[20] in the UK, although the standard of care required from the drivers of motor vehicles is rather high. Case law explains that a driver losing consciousness through no fault of his own is nevertheless acting negligently,[21] and so is the driver whose brakes fail when this failure could not have been foreseen.[22] However, the victims of accidents caused by motor vehicles have to prove that the drivers were at

---

[14] See Erica Palmerini, Federico Azzarri, Fiorella Battaglia et al, D 6.2, “Guidelines on Regulating Robotics”, 22 September 2014, (RoboLaw 2014), p. 60.

[15] Loi “*tendant à l’amélioration de la situation des victimes d’accidents de la circulation et à l’accélération des procédures d’indemnisation*".

[16] See A. Tunc, “The ‘*Loi Badinter*’ – Ten Years of Experience”, 3 *Maastricht Journal of European and Comparative Law*, 1996 (Tunc 1996), p. 330. Article 3 reads: “*Les victimes hormis les conducteurs […] sont indemnisées des dommages résultant des atteintes à leur personne qu’elles ont subis, sans que puisse leur être opposée leur propre faute*”.

[17] See also Tunc 1997, at p. 335.

[18] Compensation for damage suffered by victims *inside* a motor vehicle is governed by the general rules on liability laid down in Article 6:162 of the Dutch Civil Code.

[19] *Marloes de Vos e.a*, Supreme Court of the Netherlands 2 June 1995, *NJ* 1997/700-702, and *Saïd Hyati e.a.*, 5 December 1997 *NJ* 1998/400-402. The notion of ‘Betriebsgefahr’ is borrowed from the German Straβenverkehrsgesetz.

[20] Or p*resumed liability* as it is called in Scottish law.

[21] *Roberts v. Ramsbottom* [1980] 1 WLR 823, also cited in. Cees van Dam, *European Tort Law*, Oxford: Oxford University Press 2006 (Van Dam 2006), at p. 364, footnote 52.

[22] *Henderson v. HE Jenkins & Sons and Evans* [1970] AC 282, cited in Van Dam 2006, at p. 364, footnote 53. Van Dam further takes note of *Worsley v Hollins* [1991] RTR 252 (CA), in which the judges held that the victim’s claim for negligence failed because the defendant could prove that although his braking systems had failed, thereby causing damage, his minibus had recently been serviced and passed its MOT.

fault, that is: they had acted negligently.[23] The significant differences in the way liability for motor vehicles is addressed throughout the Member States, is not beneficial for development, insurance and deployment of AICs in Europe. In any case national regimes appointing liability to *drivers* of motor vehicles need to be updated in order to be able to address liability for vehicles without a human driver.

Privacy

Whereas the advent of AICs technology is promising in terms of increased safety on the roads, resulting in less damage to be covered, also insurance companies observe that when an accident happens caused by autonomous technology, it "would need extensive software and hardware analysis expertise in order to know how and why it occurred".[24] One of the options to assess the cause of an accident, and therefore to assist in answering the question of where liability lies, could be to equip vehicles with black boxes, or with telematics solutions connecting AICs to a dedicated infrastructure, and/or to remote servers.[25] The objectives of these types of technologies are, amongst other things, to record the movements of autonomous cars and operational choices that are made by either the car itself or the driver controlling its movement, as well as data concerning events and objects in the vicinity of an autonomous vehicle. Black box technology records and stores the gathered data inside a vehicle and offers a potential for later assessment. Telematics technology may have wider applications. Data could not only be used for assessing errors and the causes of damage after occurrence of accidents, it could even have a preventive effect. Vehicle-to-Vehicle communication (V2V) and Vehicle-to-Infrastructure communication could be used for real-time prevention of accidents, and serves "safety, mobility and environmental benefits" in general.[26] Although black box technologies and telematics solutions such as V2V and V2I (hereinafter referred to as 'tracing technology') may be promising in terms of preventing accidents and apportioning damage caused by AIC accidents, these also impose risks in terms of the right to (information) privacy of people inside and in the vicinity of cars equipped with these technologies.

Information privacy of citizens is strictly regulated in the European Union by the Data Protection Directive (DPD)[27] and will become even more strictly regulated after the General Data Protection Regulation (GDPR)[28] has come into force. The current and forthcoming framework prescribe for example that already during the design phase of AICs equipped with tracing technology, a privacy

---

[23] There is one rule of a statutory duty that – to some degree – establishes strict liability for drivers of motor vehicles approaching a crossing in the road: "The driver of every vehicle approaching a crossing shall, unless he can see that there is no pedestrian crossing, proceed at such speed as to be able, if necessary, to stop before reaching such crossing", as cited in Van Dam 2006, at p. 365, footnote 57, referring to Reg. 3 of the Pedestrian Crossing Places (Traffic) Regulations 1941, replaced by the Zebra Pedestrian Crossing Regulations 1971, SI 1971, No. 1524. A defence that a driver has in this respect is *force majeure.*

[24] Yeomans 2014, at p. 18.

[25] Yeomans 2014, at p.18. See furthermore James M. Anderson, Nidhi Kalra, Karlyn D. Stanley et al., *Autonomous Vehicle Technology – A Guide for Policymakers*, RAND Transportation, Space and Technology Program 2014, (RAND report), at p. 94-95.

[26] RAND report, at p. 81.

[27] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, *Official Journal* L 281 , 23/11/1995 P. 0031 - 0050

[28] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) COM/2012/011 final - 2012/0011 (COD). Please note that the trilogue between European Commission, Council of Europe and European Parliament has concluded on the final text of the GDPR, this text has however not been formally published yet.

impact assessment should be carried out. Furthermore "appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access, in particular where the processing involves the transmission of data over a network, and against all other unlawful forms of processing" must be implemented.[29] The GDPR regulates that these measures should be 'built in' new technology as much as possible, while these measures must *inter alia* aim at data minimization, and must be enabled by default.[30] State of the art security and implementation costs must be taken into account for the implementation of measures. Furthermore, these "shall ensure a level of security appropriate to the risks represented by the processing and the nature of the data to be protected".

Another even more recent challenge is formed by the recent decision of the European Court of Justice to declare the Safe Harbour Framework, which forms the basis of many exchanges of personal data between the EU and the United States of America, invalid. It is likely that tracing technology incorporated in AICs will constitute the international transmission of (personal) data, across the borders of the European Union, and possibly import these data to the United States for instance through cloud computing. The ECJ ruled that the US does not offer an adequate level of protection for personal data, for it became clear after the revelations of Edward Snowden, that US authorities such as the National Security Agency have easy access to personal data processed by US companies and institutions.[31] The court ruled that the powers of the European supervisory authorities are undermined by the US practices, which may not be enabled by a decision of the European Commission. This ruling implies that the export of personal data to the United States is no longer possible on the basis of the safe harbour framework. Although the United States and the European Commission are presently negotiating an alternative treaty,[32] in the meantime exchange of personal data between the EU and the United States is not allowed based on the yet invalid Safe Harbour rules.

3. Put the principle to the test

In this part we will test whether the second EPSRC principle can be considered as future proof against the criteria validity, sufficiency/generality and utility

3.1 Validity

Both given the current state of technology and of the law the first part of the principle *Humans, not robots, are responsible agents* has indeed proven to be still valid. It is a correct statement about the nature of robots, robot developers, and the relationship between robots and people. There can be always either a human being or a legal entity held responsible and liable for the actions of the AICs. The specific creation of a separate legal entity for AICs seems presently far-fetched given the current technological and legal status of AICs, it would furthermore not contribute to solving the liability

---

[29] Art. 17(1) DPD; see also art 5(1)(eb) and section 2 (art. 30 and onwards) on data security in the GDPR.

[30] Art. 23 GDPR.

[31] Case C-362/14, Maximilian Schrems/Facebook [2015].

[32] See for the latest news on the 'EU-US Umbrella Agreement' (Agreement between The United States of America and The European Union on the protection of personal information relating to the prevention, investigation, detection, and prosecution of criminal offenses): http://ec.europa.eu/justice/newsroom/data-protection/news/150908_en.htm (last accessed on March 9 2016).

challenges met as described sub 2.2. The same is true for the second part of the second principle that states that *Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy* has proven to be still valid. With an eye to the *technology-of-evidence* (to be) incorporated within AICs this fundamental idea has proven to be even more true than one might have envisaged upon its design.. As we have seen sub 2.2 in fact the flaws of the current liability regime can partially be solved by smart evidence collecting and saving systems build into the AIC. These evidence collecting and saving systems should be designed in such a way that personal data collected is protected as much as possible: privacy by design and privacy by default must be incorporated in AICs (tracing technology) at all times.

3.2 Sufficiency/generality

At the same time the principle remains still sufficient and broad enough to cover all of the important issues that might arise in the regulation of the AICs in the real-world. *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.* No significant concerns seem to be overlooked. Although some authors seem to argue that legal entity should be created for autonomous intelligent machines, making the robots the responsible agent,[33] this has not been convincing for many[34] and certainly not for us.,

The challenges posed by the introduction in society of autonomous intelligent cars and their liability for damage in itself does not seem to require a separate legal personhood. It would merely add one more actor for the attribution of liability. At the same time it would require the substantial redesign of the liability system as currently applied to the real world, whilst technology is still in its developing stage bearing the risk of under or over regulation.

3.3 Utility

As far as the current legal means are nor exhausted, inter alia by aiming at further harmonisation of EU legislative liability regimes in combination with effective technology-of- evidence, there is no evidence that would underpin a complete paradigm shift by the introduction of AICs as responsible agents in them selves. Since AIC can indeed be designed and operated to comply with existing laws the utility of this principle remains evident. However, black box technologies and telematics solutions such as V2V and V2I may be promising in terms of preventing accidents and apportioning damage caused by AIC accidents, since these also impose risks in terms of the right to (information) privacy of people inside and in the vicinity of cars equipped with these technologies the systems would need to

---

[33] See for instance James Boyle, "Endowed by Their Creator?: The Future of Constitutional Personhood", *The Future of the Constitution*, March 09 2011, p. 6, also available via the internet at <http://www.brookings.edu/~/media/research/files/papers/2011/3/09-personhood-boyle/0309_personhood_boyle.pdf> (last accessed on 9 March 2016. See furthermore J.P. Günther, F. Münch, S. Beck, S. Löffler, C. Leroux, & R,Labruto , "Issues of Privacy and Electronic Personhood in Robotics. 21st IEEE International Symposium on Robot and Human Interactive Communication", Paris 2012, as cited in Christophe Leroux, Roberto Labruto, Chiara Boscarato and others, "Suggestion for a Green Paper on legal issues in robotics",), December 2012, available on the Internet at <http://www.eu-robotics.net/cms/upload/PDF/euRobotics_Deliverable_D.3.2.1_Annex_Suggestion_GreenPaper_ELS_IssuesInRobotics.pdf> (last accessed 28 January 2016); and Robolaw 2014, p. 24.

[34] See for instance Peter Asaro, "Robots and Responsibility from a Legal Perspective", via < http://www.peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf> , last accessed on 9 March 2016; and Lawrence Solum, "Legal Personhood for Artificial Intelligences," *North Carolina Law Review*, (April 1992), pp. 1231-1287.

include privacy by design to protect this fundamental rights as laid down in International and European treaties.[35]

It is crucial that these requirements of law and technology are met before the challenge of the introduction and deployment of AICs in society can be met.

4. Conclusion

We can diligently conclude that Principle 2 of the EPSRC's principles of robotics as developed by British robotics and AI experts at the EPSRC/AHRC funded retreat, has proven to be future proof when we applied to the current state of the art of law and technology surrounding AICs.

Humans, not AICS, are responsible agents. AICs should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy by deisgn. There fore giving evidence to the fact that the answer of the machine is at least partially in the machine itself.

---

[35] See for example art. 7 & 8 of the Charter of Fundamental Rights of the European Union and article 8 of the European Convention on Human Rights.

Fair data handling and robotics

Burkhard Schafer, Edinburgh Law School
Lilian Edwards, University of Strathclyde

This intervention combines principle 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 with principle* and principle 2, *Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.* It suggests some elaboration/specification of the principles that the intersection between them necessitates. More radically maybe, it asks about corresponding duties by the users of robots, and indeed third parties towards robots. Not in the sense of "robot rights", but merely as enabler for pro-ethical and law compliant robot design. That is, to allow robot developers to adhere to the principles can also mean certain duties towards robots(or, for those who worry about this formulation, duties towards the owner of a robot to behave in certain ways wrt the machine.

Robots pose some unique challenges for fair data handling practices, challenges that are at least in part caused by their capacity to "deceive", if inadvertently, the people they interact with.

Long before modern technology, humans developed privacy preserving techniques, from the curtains to the windows to the veil, from learning when to whisper to washing away one's scent. They protected them from the prying eyes of fellow man as much as from the interest of non-human predators. And they protected not just privacy, but also other informational interests, including valuable information monopolies such as trade secrets or know how. Crucially, these not only protect information, but also the sharing and exchange of information (whispering, sound proofing your studio)

The law, with its system of rules and exceptions, frequently gave formal recognition to these low-technology protection measures. The walls we build around us do not just keep the warmth in and the rain out, but also information in and observers out. Hannah Arendt's distinction between the private, the public and the social tracks in many ways these architectures, as does our law, with the house or the hedge-protected garden the archetype of "reasonable expectation of privacy" and "security in our houses and dwellings", but also a spaces within which data can be more freely collected and exchanged- the household exception of European DP law as a prime example.

Robotics technology threatens to render these low-tech solutions to the privacy problem increasingly redundant. We face an increasing range of sensor capacities, many of which we did not encounter, or did nor encounter in a significantly threatening way, in our evolutionary past. This can undermine acquired and habituated privacy preserving strategies. They are increasingly mobile and ubiquitous. And we often will (have to) invite them into our home. Nobody is a hero to his domestics. But at least, with domestics the lord or lady of the house could anticipate what exactly they would able to see, the would understand the normative (both social and legal) environment that restrained them from collecting and most importantly sharing data about their employer. The understanding of the normative environment together with the understanding of the sensory capacities would then enable rational risk assessment and management. (I trust my butler with my dirty underwear, but not a blood drenched shirt. I don't worry about the heat signature when entertaining a guest, but may chose to keep the noise down, while relying on my butler to knock first before entering the bedroom).

Robotics threatens these defensive strategies not just because they can use sensors outside the visual or aural spectrum, or because of their mobility that allows sensing in spaces previously protected. Where they imitate the outward appearance of humans, or indeed non-human animals, even in cases where their robotic nature is plain visible (as per principle 4). While not (at present) backed up with systematic research, there is evidence that we do make these inferences when interacting with robots. The Internet is abundant with people "sneaking up" on Asimo from behind – now Asimo's sensors "may" indeed be located in its eyes, and have vision restrictions similar to a human, but this may well be false.

Part of ethical design therefore should also be to indicate the sensory capacities of robots in ways that facilitate the emergence of "intuitive" defences of the type we use with other humans, and refrain, where possible, from inviting misleading inferences, and include "ease of defensive mechanism" in the evaluation of intrusiveness when a choice between different sensors can be made.

Data protection law is one driver behind this, but fair sensing and data handling practices go beyond personal, let alone sensitive personal, data. We protect not just data about us, but also our business ideas, scientific or technological discoveries, or skills. Here too we reason instinctively about sensory capacities by potential adversaries (think of school children building a wall of books around them during exams, to prevent others cheating) .

IP law is therefore another legal constraint that needs to be observed under this header, and a broader notion of "fair data handing practices" that goes beyond DP law may be needed. If an industrial robot observes my movements to improve collaboration and avoid collusions, and in the process learns enough to make my job less secure (the way I move is my unique selling point), is this an "interest" of mine that deserves protection/compensation?

This, potentially, raises however also a question that leads in a more radical way beyond the Principles. They try mainly to establish duties that the developers owes to people who interact with their machines. The self-defence idea, which would have established more of a symmetrical relation of rights and duties, was dropped at the time. Nonetheless, the discussion above leads to a necessary nexus between duties that the developers owe, and possible duties that are owed to them/the robot owner.

As a simple example, to allow safe robot design may involve a duty for third parties to disclose or share certain information with the robot that in the past has been legally privileged. A robot that navigates e.g. an art exhibition may have to create copies of the exhibits and share them with other robots) simply to avoid running into them, even though copyright law may permit the gallery owner from preventing even this "incidental" copying. The U.S. approach that argues that this would be copying for functional rather than expressive purposes – not speech – and so should not be breach copyright – however, once machines co-ordinate their action by sharing this data even this ln eof argument may reach its limits.

Citizens may chose to use technology to prevent sensors from noticing them (e.g. camouflage face paint - https://cvdazzle.com), but that may mean that they accept greater risk that the robot runs into them. If third parties are involved, this can create even more complex legal issues. If my face provides a data point from which I know the robot learns, do I have "quality assurance duties" to be a good example? If I intentionally manipulate the learning process, is this getting into the territory of the Computer Misuse Act? And finally, if I contribute to the learning of a machine, do I have a stake in what it produces as an outcome?

Basic negligence law and its distinction between act and omission and how negligence deals with it by establishing duties to neighbours, will be part of the legal answer *after* an accident happened. For the purpose of the discussion here however, the question is posed slightly differently: At the point of developing a robot, can/should the designers, in discharging their duty

1. Rely on the *ethical/social* duty by third parties not to manipulate the knowledge acquisition of the machine
2. Rely only on a narrower *legal* obligation to refrain from certain foreseeably dangerous data manipulation
3. Not rely at all on a cooperative environment when thinking about the safety and law compliance of the robot they build – after all, not all laws are observed by everybody.

To make clear why this issue arises in the context of a discussion on "sensor transparency" : IF we accept the ethical obligation discussed above, i.e. that robots should normally disclose how and with that what they can sense, then they inevitably open themselves up to manipulation. If we in addition accept 1, or at the very least 2, this is less of an issue than if we accept 3.

If we accept 2, then we have to deal head on with the fact that at the two extremes of the spectrum, the law is clear: it's hard to owe duties to non humans, on other hand I can owe a duty to neighbour not to burn down his barn But do I owe a duty to absent robot operators (or designers) not to confuse their robots training? Is it reasonably foreseeable a robot would be confused? After all I don't even have duty not to lie to strangers eg when giving directions – unless some relationship of professional advice. But then again, sending a child astray would be a different proposition. Are machines that "still learn" analogous to such a situation?

The discussion so far discussed problems caused by humans who withhold/distort/manipulate data that a robot needs to operate safely, and which they ethically or legally owe.

But we also face ethical design choices when people cooperate, and volunteer information that they are not legally obligated to provide, but chose to/do not inhibit out of a sense of civic duty. Should this affect the status of any *output* the robot produces, e.g. in a form of benefit sharing?

This could mean that not only robots need to be identifiable as robots, their sensors as sensors, but robot generated output also must be identifiable as machine, not human generated. Here, copyright law might impose relevant constraints on the way in which machines communicate in a law complaint way: in jurisdictions that do not protect computer generated works, a "this text was generated by algorithm, feel free to share" might be required

The overarching theme of this intervention, therefore, is ultimately one of algorithmic transparency: legal and ethical duties influence when, and how, robots should disclose their sensory capacities. Once they do this, their environment has choices – to cooperate or not to cooperate. Where non-cooperation causes harm, there might be wider social discussions to be had if it is more beneficial to allow clandestine gathering of (some) data to have more secure machines. Where cooperation beyond the legally required produces value, a discussion needs to be had how to account for this in an equitable way, imposing potently another disclosure duty, a "made by robot"

While these are mainly legal issues, for the question of ethical (and law compliant) design, the developers need also to be able to anticipate what type of interaction to expect, and what type of information they will have legal access to.

Can robots be responsible moral agents? And why should we care?

Amanda Sharkey,

Department of Computer Science, and Sheffield Robotics, University of Sheffield

Principle 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.

At first glance, this statement or principle seems convincing. It makes sense to insist that humans and not robots are responsible agents. It usefully reminds us of the limited abilities of robots, and provides a helpful antidote to the strong claims and warnings sometimes made about them. We should not offload blame for mistakes or bad consequences onto robots. Emphasising human responsibility for robot behaviour should help to restrict the possible harmful uses to which robots could be put. It also makes sense to suggest that robots should be designed and operated to comply with existing laws and fundamental rights and freedoms: it is difficult to imagine anyone suggesting otherwise.

But, on further consideration, it becomes apparent that the statement does not give any justication for saying that humans and not robots are responsible agents, nor does it provide any guidance about where and when robots should be used, or the consequences that follow from assuming that robots are not responsible agents. The statement raises a number of issues that deserve further discussion. It raises important issues and questions about legal responsibility, but they are not discussed here. The issues that will be considered are (a) What are the reasons for assuming that robots and not humans are responsible agents? (b) Is it sufficient to design robots to comply with existing laws and fundamental rights and freedoms? and (c) If robots are not responsible agents, should this limit the roles they are given and the situations in which they are deployed?

(a) **What are the reasons for assuming that humans and not robots are responsible agents?**

Aside from legal responsibility, it is possible to identify two reasons for this assumption. The first is based on the difference between biological and mechanical machines, and the biological basis of morality. The second is to do with the need for society to accept responsibility for the artefacts that humans have produced. We consider both of these in turn.

(i) Biological machines versus Mechanical machines: Holding an agent to be responsible for its actions is equivalent to holding it to be a moral agent. It is therefore relevant to highlight the biological basis for morality in biological machines, and to contrast this to the absence of such a basis in mechanical machines such as robots. Patricia Churchland (2011) discusses the basis for morality in living beings, and argues that the basis for caring about others lies in the neurochemistry of attachment and bonding in mammals. She explains that it is grounded in the extension of self-maintenance and avoidance of pain in mammals to their immediate kin. Neuropeptics, oxytocin and arginine vasopressin underlie mammals' extension of self-maintenance and avoidance of pain to their immediate kin. Humans and other mammals feel anxious about their own well-being and that of those to whom they are attached. As well as attachment and empathy for others, humans and other mammals develop more complex social relationships, and are able to understand and predict the actions of others. They also internalise social practices, and experience 'social pain' triggered by separation, exclusion or disapproval. As a consequence, humans have an intrinsic sense of justice. The same is largely the case for non-human mammals. Bekoff and Pierce (2009) provide many examples of evidence of a moral sense of justice in mammals. For example, capuchin monkeys working for treats seemed offended and would refuse to cooperate further if they saw that another monkey was given a more desirable reward for the same work (Brosnan and de Waal, 2003).

By contrast, robots are not concerned about their own self-preservation or avoidance of pain, let alone the pain of others. In part, this can be explained by means of an argument that they are not truly embodied, in the way that a living creature is. Parts of a robot could be removed from a robot's body without it suffering any pain or anxiety, let alone it being concerned about damage or pain to a family member or to a human. A living body is an integrated autopoeietic entity (Maturana and Varela, 1980) in a way that a man-made machine is not. Of course, it can be argued that the robot could be programmed to behave as if it cared about its own preservation or that of others, but this is only possible through human intervention. We return to a further discussion of the feasibility of programming morality below.

Societal responsibility: Many writers would agree with the implication of the statement that robots are not full moral agents. Johnson and Miller (2008) argue that robots, and other computational artefacts, are not full moral agents because they "are not ever completely independent from their human designers". They describe them as 'human-tethered' artefacts, and argue that responsibility cannot be offloaded onto the artefacts themselves since the behaviours and outputs of robots and computer systems necessarily depend on human designers and developers. A useful example that they consider is that of a door-opener. A person who opens the door for someone carrying a package can be viewed as having performed a positive moral act. But if the door is opened by means of a sensor that detects the approach of a person, the mechanical door opener would not be considered to have performed a praiseworthy act. Related arguments about a lack of independence from human designers have been made in the past based on the way in which robots, unlike living machines, can never be considered to be fully embodied, since they have always required human intervention and involvement in their development (Sharkey and Ziemke, 2001). The point here is that robots, and their underlying control systems, depend on human intervention. The robots may be 'set loose' to make unpredictable decisions, but the decision to allow them to do so is a human and societal one. Any decisions made by the robot will still depend on their initial design. Even if the robots are 'trained' or 'evolved' to make decisions, their training or fitness regime will still have involved human intervention at some point, and it is imperative that human responsibility is assumed and recognised. Johnson (2006) makes a useful distinction between moral agents and moral entities, and places robots and computer artefacts in the second category. Moral entities include the artefact designer, the artefact, and the artefact user, and moral responsibility cannot be offloaded onto the artefact itself.

(b) Is it sufficient to design robots to comply with existing laws and fundamental rights and freedoms, including privacy?

A major problem with the suggestion that robots should be designed to comply with existing laws and fundamental rights and freedoms, and the reason that it is not sufficient, is that existing laws and human rights have not been formulated with technological developments such as robotics in mind. There is a need to reconsider these in the light of such developments. For example, robots pose a particular risk to privacy, particularly when they are designed to appear as friends and companions, and as a result are welcomed into our homes and intimate surroundings. There are many questions here to be answered about the extent to which the information they have access to will be accessible to others, and as yet little legislation to address this. Ethical concerns have been expressed about the risks of leaving vulnerable older people in the near-exclusive 'care' of robots, with little human contact, (e.g. Sharkey and Sharkey, 2012; Sparrow and Sparrow 2006), but the Human Rights Act does not provide any explicit protection from such a situation. Similar concerns have been raised about leaving children in the 'care' of robots to the extent that their attachments to humans are compromised (Sharkey and Sharkey 2010) but again there is no legislation or rights that explicitly prevent such a possibility, other than that associated with child neglect. There is an urgent need for something like a digital bill of rights to ensure that some there is some protection from the situations that could arise if humans place robots in positions of power over humans.

When humans make decisions about how to act in social situations, they have to do more than follow a set of rules, or laws. They make decisions based on a moral understanding of what it is inappropriate or inappropriate for them to do. They are sensitive to feedback about their decisions and their outcomes, and can reflect on it and adjust their future decision-making. There have been discussions about the extent to which robots can be programmed or trained to make the right moral decisions in social situations. Arkin (2009), for example, has argued that in a battlefield situation, robot soldiers could be programmed to follow a set of rules that would result in more ethical behaviour than that sometimes shown by human soldiers in the heat of battle. His claim is that human soldiers can act badly as the result of their emotions – for instance being motivated by revenge to carry out war crimes. A robot on the other hand would not respond emotionally and could be programmed, by means of an 'ethical governor' to evaluate actions before carrying them out, and to only perform those deemed morally permissible.

Various authors have argued against the idea of being able to program robots to make moral decisions. In the context of autonomous weapons, Christof Heyns, the UN Special Rappoteur on Extrajudicial, Summary or Arbitrary Executions has argued against the use of autonomous robots to make lethal decisions on the battlefield on the basis that robots lack 'human judgement, common sense, appreciation of the larger picture, understanding of the intentions behind people's actions, and understanding of values and anticipation of the direction in which events are unfolding' (2013, A/HRC/23/ 47). The point is that the unpredictable variety of social situations that could arise on the battlefield means that it is unlikely that a set of pre-programmed rules about appropriate responses is likely to be applicable.

In an interesting paper about the requirements for creating robots with, what they term 'moral competence', Malle and Scheutz (2014) argue that, amongst other things, robots would require a network of moral norms, in order to know what is and is not morally acceptable. They suggest that it would not be practical to program this network, and that instead of programming robots with moral norms, they could learn and develop a network of moral norms on the basis of feedback given to them in response to their actions. They suggest that it might be necessary to raise the robots in human environments, since this may be 'the only way to expose the to the wealth of human moral situations and communicative interactions' (Malle and Scheutz, 2014). Others have suggested that robots' understanding of right from wrong could be improved by training them on moral stories (Riedl and Harrison, 2016), and requiring them to reverse engineer the human values that they represent.

It is admittedly difficult to rule out the possibility that in the future a robot could be trained or raised to be moral, but there are a number of reasons to be sceptical about the likelihood of success. Reasons for scepticism include the robot's lack of a biological basis for morality as discussed earlier. As already discussed, an individual robot does not even care about its own body, let alone that of a human – it would suffer no pain if one of its wheels were to be removed for example. It could only be programmed to respond *as if* it cared about the effects of its actions on a human, or about any censure and moral disapproval of its actions. Another reason for scepticism is the complete lack of any convincing examples of robots developing a good, generalisable, understanding of the differences between right and wrong. All there is currently are examples of programmed behaviour, such as the robots programmed by Winfield et al (2014) to take actions to prevent other robots from falling into a hole, that are described as exhibiting something that can be described as ethical behaviour. But the use of the term 'ethical' or 'moral' in this context does not mean that the robots in question could be legitimately praised or blamed for their actions.

(c) If robots are not responsible agents, should this limit the social roles they are given and the situations in which they are deployed?

The original statement that robots are not responsible agents does not spell out what this implies for the deployment of robots. It is argued here that there are good reasons to limit the social roles and decision-making powers of robots. As referenced above, Heyns (2013) argued that robots should not be allowed to make kill decisions in battle, partly because of their lack of ability to understand social situations, but also because humans should have a right to have life and death decisions about them made by fellow humans. A related argument could also be made about robot policemen, who might also be tasked with life and death (or serious injury) decisions away from the battlefield.

This argument can, and I would argue should, be extended further to other kinds of decision where robots might restrict the freedoms of humans. A robot placed in the role of a teacher would have to make decisions about things like when to punish or restrain children, or when to praise them. A robot carer of older people might have to make decisions about when to share personal information about them to other people, or when to prevent them from doing something dangerous or risky. A robot nanny would have to make similar decision about its young charges. The point is that all these decisions are likely to involve moral judgements and evaluations of social situations, and for reasons already discussed the robot is unlikely to be able make good choices. Care should be taken to maintain human control, involvement, and responsibility in decisions that will affect the lives of humans. There are already risks of automated decisions affecting our lives, but robots which can be given the appearance of competent social actors make these risks even more prevalent.

Summary: It is easy to agree with the EPSRC principle about robots not being responsible agents, but even this brief consideration finds it to be insufficient to guide future action. It does not refer to any discussions of the reasons for claiming that robots are not responsible agents, nor consider the implications for the deployment of robots and for human choices about the social roles they should be given. Robots programmed to follow the law, and to respect individuals' rights and freedom, are not going to understand social situations and will not be able to consistently make the right moral decisions about human social situations. Therefore it is important to avoid placing robots in social roles and situations in which moral decisions are required. Care should be taken to avoid or minimise automatic and algorithmic decision making in any situations in which human judgement is required. Even greater care is needed in the case of robots that create the illusion that they understand. Humans do sometimes make flawed decisions, but they can reflect and learn from them and develop a better moral understanding in a way that a robot cannot.

References

Arkin, R. (2009). Governing lethal behavior in autonomous robots. Chapman-Hall review. *Computers and Education,* 58(3), 978–988.

Bekoff, M., and Pierce, J. (2009) *Wild Justice: The Moral Lives of Animals*. The University of Chicago Press, London.

Brosnan, S.F. and de Waal, F.B. (2003). Monkeys reject unequal pay. *Nature*, 425, 297-99

Churchland, P. (2011) *Braintrust: What Neuroscience tells us about Morality*. Princeton University Press, Oxford.

Heyns, C. (2013). Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, A/HRC/23/47

Johnson, D.G. (2006). Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology*, 8(4): 195–204

Johnson, D.G., and Miller, K.W. (2008) Un-making artificial moral agents. *Ethics and Information* T*echnology* (2008) 10:123–133

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. IEEE International Symposium on Ethics in Engineering, Science, and Technology (pp. 30–35). Presented at the IEEE International Symposium on Ethics in Engineering, Science, and Technology, June, Chicago, IL: IEEE.

Maturana, H. R. & Varela, F. J. (1980). *Autopoiesis and Cognition - The Realization of the Living*. Dordrecht, The Netherlands: D. Reidel Publishing

Riedl, M.O., and Harrison, B. (2016) Using stories to teach human values to artificial agents. In Proceedings of 2nd International Workshop on AI, Ethics and Society, Phoenix, Arizona

Sharkey, A. J. C., & Sharkey, N. E. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology, 14*(1), 27–40.

Sharkey, N. E., & Sharkey, A. J. C. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies,* 11(2), 161–190*.*

Sharkey, N. E. & Ziemke, T. (2001). Mechanistic vs. Phenomenal Embodiment - Can Robot Embodiment Lead to Strong AI *Cognitive Systems Research*, 2, 4, 251-262

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Mind and Machine,* 16, 141–161.

Winfield, A.F., Blum, C., and Liu, W.(2014) Towards an ethical robot: Internal models, consequences and ethical action selec- tion. In M. Mistry, A. Leonardis, M. Witkowski, & C. Melhuish (Eds) *Advances in autonomous robotics systems: Proceedings of the 15th annual conference, TAROS 2014* (pp 85–96). Birmingham, UK, 1–3 September

Second thoughts about Privacy, Security and Deception

Tom Sorell, University of Warwick
Heather Draper, University of Birmingham

The five principles of Robotics formulated during the AHRC- EPSRC Retreat in 2010 are not the last word in robot ethics, but one of the first words. There is a long way to go. In what follows we discuss (a) a problem with Principle 2, taken on its own; (b) a tension between Principles 2 and 3; and (c) some scepticism about the application of Principle 4.

Privacy

Principle 2 requires robots to be operated in conformity with existing laws, and fundamental rights and freedoms, including the right to privacy. But this principle is hard to abide by, because there is disagreement among legal academics and philosophers over what privacy is, and also over whether it is a fundamental right. Although privacy is said to be a fundamental right in the European Charter of Fundamental Rights (Articles 7 and 8), it is not fundamental in older human rights treaties, such as the International Covenant on Civil and Political Rights (see Article 17). First, privacy is not among the so-called "non derogable" rights, such as the right to be spared torture or to be spared discrimination. Second, privacy can be limited by other rights and by the need to maintain public order. Third, many human rights theorists deny that there is a hierarchy of human rights in which any are more fundamental than any others. Instruments interpreting the status of human rights, such as the Vienna Declaration, say that the all human rights are interdependent and indissoluble (Art. 5). So even if there agreement about what privacy is, the need to respect a right to privacy would not necessarily be overriding.

Personal privacy is sometimes understood as control over information about oneself. Care robots in particular and social robots in general are often designed to collect information about the human beings they interact with. For example, robots monitor facial expressions and process spoken speech, track the location of people they interact with, collect information about their daily routine activities, and so on. These are not necessarily violations of a right to control over one's information, because the people whose information it is can in principle provide consent to the collection and storage of it. Because the use of their information is subject to their consent, control does not pass into other people's hands: consent is a form of control.

Consent, however, does not necessarily settle all questions about the proper use of personal information. First, there is a difference between the collection of information on a one-off basis, or an intermittent basis, and the more or less continuous collection of information in real time. The implications of the second are harder to predict and consent to in advance than the implications of the first. It is even arguable that there is no such thing as properly informed consent to the continuous monitoring and tracking of a live-in care-robot precisely because it is not possible to predict or even to imagine in advance what the experience of living with a robot would be like.

Instead of determining the limits of privacy just from what a user consents to with good information, one may have to rely also on arguments about the limits of privacy based on the design brief and purpose of a particular kind of robot. Care robots for older people are often supposed to assist in the maintenance of their autonomy – that is, their capacity for choice and having a skill set – the ability to wash oneself, clean, cook, feed oneself etc. – sufficient for living independently. A person who is autonomous decides for himself or herself not only what to do and how to live but also what personal information to disclose. A robot designed to maintain the autonomy of an older person can partly be judged by whether the older person has as much latitude as an unassisted adult to decide about all aspects of his or her life, including disclosure of information. The closer to the latitude of the standard adult, the closer the robot-assisted older person comes to the standard adult's autonomy.

The tension between autonomy (and privacy) and safety

One way in which the standard adult exercises autonomy is by being their own judge of what risks to take. The moral permissibility of risk-taking is of course affected by the costs to others. If others are put in danger or relied upon to undertake a dangerous rescue of the autonomous risk-taker, then there may be an argument against risk-taking from the autonomy-restricting burdens it imposes on others. In other words, if we have to abandon our plans or delay realizing our choices because of a risk-taking choice made by someone else, then our autonomy is subordinated to his, at least temporarily.

In the case of the assisted-older people, the robot design brief will usually combine safety *and* autonomy. The robot assists users in leading their own lives, and it also monitors the user and his or her circumstances for health emergencies. If emergencies or abnormalities are detected it can raise the alarm. This design creates a problem: what if the robot-assisted user makes an autonomous but risky choice?

In the cases that are most interesting from the angle of moral theory, the user is willing to take a relatively small risk – say the risk of having a minor fall – for the sake of continuing to lead daily life in the same way as they did when they were younger. Where a relatively small risk materializes and a user suffers e.g. a bruise by falling, they have the right to prevent an alarm being raised or prevent the sharing of information about the fall. This is because maintaining autonomy is supposed to be the overriding purpose of the companion robot. If, however, the overriding goal is keeping the user safe, they might not have this right. It all depends on how much minor harm is compatible with being safe. In normal life people can survive falls, cuts, and even minor automobile accidents with no need for rescue or intervention. This is an aspect of the autonomy adults enjoy that companion robots are supposed to prolong. In low- tech telecare, a pendant alarm is worn by a user, and it for him or her to decide whether or not to summon help. The value of autonomy supports this norm. In the robot case the same norm could be defended, permitting the user to override the robot's default decision to report the fall.

User-overrides could also be incorporated in companion robot design where life-style choices of a user, if they were reported to friends and relations, might prompt coercive interventions from those people. Here is where autonomy supports privacy over total safety or total prudence. If an older user who is of sound mind wants to gamble away some of his money, that should be no less an option for him or her than it is for a middle-aged person; otherwise a kind of ageism restricts older people's autonomy and they are treated worse by robot designers and public policy formulators than other adults. If a robot reporting back to a user's relations interfered with gambling, *that* would be to limit an assisted person's autonomy more than that of an unassisted person. And this limitation is hard to defend without ageism. This is true whether or not there is an argument for gambling being unsafe at any age. The point is that autonomy can outweigh even benign data-sharing and benign interventions to prevent risk-taking.

The tension between autonomy and rehabilitation

Care robots and some non-social robots are designed to help older people to *regain* abilities they have lost and not just exercise those they have autonomously. How much of an obligation do older users or other users have to co-operate with rehabilitation routines scheduled and administered by the robots? If rehabilitation maintains or supports autonomy, and maintaining autonomy is seen as a joint enterprise between users and whomever pays for introducing the robot into a user's home, an obligation to co-operate may exist. If no such joint enterprise is recognized, room has to be made for an autonomous *refusal* to accept rehabilitation that is beneficial. After all, an autonomous refusal to accept a *medical* intervention cannot legally be ignored.

What if robot assistance for an older person is provided on the condition that they agree to co-operate with rehabilitation that may be offered in the future? In this case autonomous refusal may be overridden by an autonomous undertaking to co-operate. This suggests another constraint on the use of a robot: not only must the use be consented to and in keeping with the purpose the robot is (permissibly) designed to serve, but it should be subject to an explicit contract that the user enters into, which specifies the responsibilities a user takes on in return for the provision of the robot. Such a contract might exist between a user and a local authority. Responsibilities under the contract would not exclude rights, of course.

Deception

Finally, we come to Principle 5. It calls for transparency in robot design and prohibits deception of the vulnerable. Our scepticism about this principle is based on the low threshold for deception that is set by some of the robot ethics literature. Deception is the intentional creation of false beliefs. Deception is usually wrong because the deceiver wants to manipulate the deceived person to do something that serves the deceiver's interests. In a robot project on which one of us (Sorell) is an advisor, a small, childlike teaching robot gives hints in learning exercises to sometimes very young pupils, around 6 six years old. The robot itself has no intentions to deceive, but is its childlike design deceptive? Does it work by making a child think another child is helping him or her? The answer here is 'No'. Although very young children might well form attachments to these robots because of the way they look, this attachment does not seem very different from the attachment they form to the soft toys they anthropomorphise even before they reach the age of 6. The anthropomorphisation is not a case of self-deception, and neither is the modelling of the educational robot on a humanoid child a case of deception, either. To identify with such a robot or to treat it affectionately is not to treat it as another child. It might be to treat it as a representation of a child. A similar conclusion can be reached about the Paro robot. To stroke Paro is not to treat it as a live seal but as a sort of seal representation. This is not a case of deception on the part of the manufacturer of Paro, or a case of self-deception either. Patients with dementia seem to use Paro and also non-robotic dolls in much the way that very young children use soft toys. Without being able to construct imaginative stories in which the dolls figure, these users derive comfort from the look and feel of the robots.

The robots are a reassuring presence, much as a tame dog or cat might be. It is not crucial to deriving comfort from Paro that one thinks it is a real seal or that one thinks it is alive—the fact that its behaviour simulates that of a pet is enough. So it is hard to understand what false belief, let alone what *intentionally-produced* false belief, is crucial to Paro's therapeutic effect. The same conclusion seems to be reachable by parallel reasoning for humanoid teaching robots.

Commentary for AISB Workshop on Principles of Robotics

Emily C. Collins

The University Of Sheffield, Sheffield, UK.

1 Introduction

The following principle is aimed at regulating robots in the real world:
No. 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 commentary will offer a critique of this principle in line with the following criteria:
a. Validity. Are the principles correct as statements about the nature of robots (for instance that they are tools and products), robot developers, and the relationship between robots and people (for instance that robots should have a transparent design), or are they ontologically flawed, inaccurate, out-dated, or misleading.

The critique will break the principle down to what I consider to be its two main component statements:
1. *Robots* should not be designed in a deceptive way to exploit vulnerable users.
2. Machine nature should be transparent.

I will argue that both of the component statements that make up this principle are fundamentally flawed because of the undefined nature of the critical terms: ‘deceptive’, ‘vulnerable’, and ‘machine nature’, and that as such the principle as a whole is misleading.

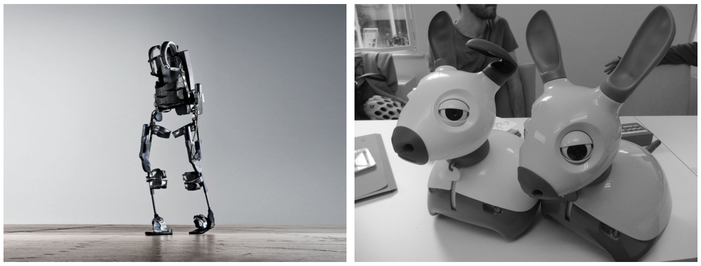

Fig. 1. Left panel: ‘Ekso’, short for exoskeleton, is a wearable robot that helps paralysed patients walk. Right panel: Two MIRO robot mammals, an example of a ‘social’ robot.

For the purposes of this commentary a robot is defined as a manufactured artefact, specifically a tool with which a human user can augment an existing state, for example by providing an individual who cannot walk the capacity to walk by means of machine-aid, or by providing a user with an advanced form of entertainment, as with a companion robot (Figure 1). This commentary will focus in particular on biomimetic [1], social robots, and their role as tools for the betterment of users. A social robot is here defined as a device with some autonomy and physical presence that is capable of interacting socially with people, and as such can be expected to elicit an emotional response of some sort from its user [2]. Here is our first problem, before the principle is even addressed: to define 'robot' one must at least define the robot's application and the extent of its capabilities. There exist mutually exclusive types of robot, which have the capacity to potentially deceive users in a variety of distinct ways dependent upon how each robot is interacted with. Industrial, mobile, service, educational, space, and social robots, to name but a few, have different morphologies and come with different sets of expectations from their users. None of the Principles of Robotics begin with a definition of 'robot', and so I have defined my own.

2 *Robots should not be designed in a deceptive way to exploit vulnerable users*

Firstly, let us begin by asking what is 'deceptive'? In this context it is the robot being labelled as deceptive, thus a better question might be how is the robot being deceptive such that it would go against this principle?

Robots are being developed to resemble living things. What is known about human-animal dynamics is being used to build animal-like behaviours and morphology into the design of social robot. Biomimetics by definition means design by nature, via the imitation of the models, systems, and elements of nature, for the purpose of solving human problems. Robots are tools, products for use, the purpose of which is to solve human problems. Here the very design principles that underlie the nature of a biomimetic, social robot, and what robot developers require, is driven by what could arguably be called 'being deceptive': attempting to mimic living things for the betterment of the robot and its user.

Animal-like robots, such as Paro [3] and the 'FurReal Friends Lulu Cuddlin Kitty', produced by Hasbro (Figure2), are being used by therapists in a manner similar to animal-assisted therapy (AAT) [4], wherein an animal might be brought into an existing therapy session to aid with social facilitation (as with group therapy), or use in a one-on-one manner to help focus a client or patient during therapy. These robots serve a specific purpose, to appear animal-like and aid the therapist. Their existence however, though based around a treatment that does involve a living being, is not to replace animals. The animals in AAT are considered co-therapists. They are given the regard a living creature would be expected to have, and are removed from sessions wherein harm may befall

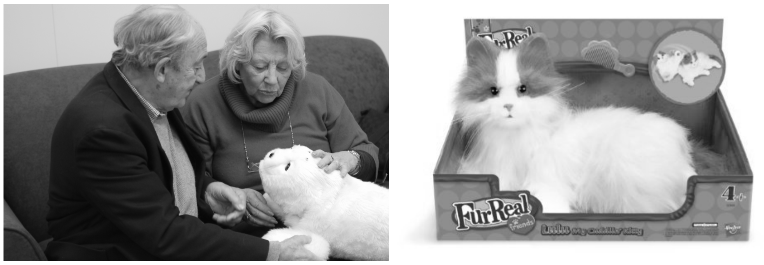

Fig. 2. Left panel: The ‘Paro’ therapeutic robot. Right panel: The ‘FurReal Friends Lulu Cuddlin Kitty’.

them, or wherein they are themselves being disruptive [5]. This example demonstrates that a robot designed with deception in mind - to look animal-like - with the intent of being used by vulnerable populations - individuals in therapy - are not intended to be exploitative as defined by attempting to convince a user that the animal-like robot is really alive. They are instead used to trigger associative memories of other living things. It is difficult to convey this idea, but the nuance is important. These robots are not built to be convincing animals. They are built to be convincing robot tools, and to achieve that ideas from nature are borrowed.

Secondly, what does it mean to ‘exploit vulnerable users’? What is a vulnerable user? Is being vulnerable a single state of being? And if so, at what point might one be considered, or no longer considered, vulnerable? Indeed, who might decide at what point an individual became vulnerable enough to have their State of the Art robot taken from them? [1]

Within medicine there is a standardised definition of vulnerable groups, within which exist defined domains of vulnerability (e.g., [7]). How a vulnerable individual is being exploited by deceptive robots depends on where the vulnerability of the individual lies. For example, the medical domains include economic vulnerability. Consider the emotional exploitation of fear, created by a populist media that propagates the belief that a person’s job may be under threat by robots which are being deceptively portrayed as more advanced than they are. Though we can assume that the principle is not referring to such vulnerability as economic (though in truth we cannot assume that; part of the problem with these Principles of Robotics is that they are not defined in such a way at all, but for the purposes of this commentary let us assume that the vulnerability being referred to is physical rather than conceptual). So perhaps let us assume that by ‘vulnerable’ the principle is referring to groups. Let us also assume that a general user will know when a robot is a robot unless that robot is so exceptionally lifelike

---

[1] For an example of this in fiction, see Issac Asimov’s earliest publication, *Robbie* [6]. The fear of robots exploiting the vulnerable has been a longstanding one within the robotics’ community, but fiction must be teased from fact in appreciating this issue.

as to pass as living. To pass as living the robot would be required to, and this list is by no means exhaustive, move, respond, blink, breathe, and vocalise in a synchronous manner as well as be morphologically exact. Such technology does not exist. Thus, considering the State of the Art that does currently exist, such as the social robots that are the focus of this commentary, the issue arises from the fact that it is precisely the most vulnerable within a population who have the most to gain from their use. The two most vulnerable groups are commonly considered to be the elderly and minors, and within those groups individuals with cognitive impairments.

For the purposes of this commentary let us focus on vulnerable groups within the elderly population. The aforementioned Paro robot is an advanced interactive robot designed to provide physical and emotional support to the sick and elderly, not by itself, but with the aid of a clinical practitioner trained in Robot-Assisted Therapy (RAT). In individuals suffering from dementia and other conditions of cognitive decline, emotional capability does not decline in a one-to-one fashion with cognition [8]. This allows for meaningful application of psychological and emotional therapy by a therapist with such static devices as Paro, which is designed to resemble a living being, to be held, and to be fussed over [9]. Here the deception resembles that seen with doll therapy.

In doll therapy interventions, dolls that resemble lifelike babies are used by Alzheimer's disease caregivers to try and ease anxiety, and bring joy to those suffering from dementia. This is achieved via the introduction of a purposeful and rewarding, yet physically harmless, activity: namely caring for the doll (e.g., [10]). Though controversial [11] such therapies that introduce lifelike focal point tools into the care process have been praised for improving the Quality of Life (QoL) of patients, and such studies include ones that have explored the impact of using animal-like robots in therapy too [12].

QoL is a complex measurement encompassing emotional, social and physical aspects of an individual's life. It exists on a continuum, outside the realm of 'either/or dichotomies' where x is considered bad and y good. If a tool that is robotic is being used with a vulnerable population that has mental capacities that can be exploited to relieve the suffering of individuals within that population, the question of whether or not that tool should exist becomes vague, and too complex to answer with a single statement. The debate comes down to how far we should be deceiving the vulnerable, and at what point that becomes exploitative in a negative sense. When that consideration is set against improvements in the QoL of individuals suffering from incurable neurodegenerative diseases, it becomes clear that this forth principle is insufficient. It is fundamentally flawed because its component terms go undefined. Without knowing what is meant, really, by exploiting the vulnerable, the whole principle is misleading.

If the thing being exploited is the cognitive decline itself, and the robot is deriving benefit from the vulnerable nature of the individual, but for the purposeful outcome of improving QoL of that individual, is that not positive? When there is no other alternative to access the remains of a dementia suffer's emotional quotient, someone who might otherwise be fearfully triggered by an otherwise comforting living animal, where is the real harm? Does the harm lie in the minds of those who do not suffer and witness what they themselves reflectively consider a sad state? And if that is the case should we not try all the more to project ourselves into the mind's of the vulnerable, and appreciate this situation for what it is? An attempt to provide care using all and any tools available, enacted with goodwill, and overseen by carers who know the full extent of the damage that neurodegenerative diseases cause, both to the patients, and their loved ones watching on.

3 *Machine nature should be transparent*

Let us consider the healthy population that observes robots. As previously stated in this commentary I believe that there does not exist robotic technology such that it is perfectly deceptive. Even the most sophisticated robots are clearly robots. A user may believe a robot's AI to be more advanced than it is at first blush, but, at least anecdotally through my own experiences in the lab, I believe that any period of time with a robot is sufficient for a user to establish a rough enough approximation of its limitations such that any initial over estimation of the robot's capabilities are soon overridden with the reality. As for those populations vulnerable enough to be deceived into believing that a robot is more advanced or more 'alive' than it is, I believe that it is not the robot that should be designed differently, but that the human users or clinical RAT practitioners who should be trained to use their tool, their robot product, in the most effective and positive manner.

4 Summary

A robot that is so perfect as to wholly deceive a user into believing it is anything but a machine, is something I fail to imagine existing any time soon. For those individuals who are vulnerable enough to be convinced that a robot which is transparently a machine, is in fact alive, my recommendation is to consider as objectively, and as broadly as possible all the positive benefits that can arise from such a situation. To consider what it actually means to exploit the vulnerable, and perhaps to rephrase a scenario with ostensibly positive outcomes for the vulnerable user, without the use of the term 'exploit', but rather with the word, 'aid':

> Robots are manufactured artefacts, but ones that are tools to aid us and can be designed using principles we know to work, including biomimetic ones. Robots that are designed in a deceptive way, to be used to aid the suffering of vulnerable users, should have their machine nature be made known to the

carers of those vulnerable users. May it be the carer's responsibility to improve the QoL of their patients by any safe means necessary.

References

1. T. J. Prescott, M. J. Pearson, B. Mitchinson, J. C. W. Sullivan, and A. G. Pipe, "Whisking with robots from rat vibrissae to biomimetic technology for active touch," *IEEE Robotics and Automation Magazine*, vol. 16, no. 3, pp. 42–50, 2009.
2. E. C. Collins, A. Millings, and T. J. Prescott, "Attachment to assistive technology: A new conceptualisation," in *Proceedings of the 12th European AAATE Conference (Association for the Advancement of Assistive Technology in Europe)*, 2013.
3. T. Shibata, ""Mental commit robot (PARO)"." [Online], http://www.paro.jp.
4. M. R. Banks, L. M. Willoughby, and W. A. Banks, "Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs," *Journal of the American Medical Directors Association*, vol. 9, no. 3, pp. 173–177, 2008.
5. S. Brooks, "Animal-assisted psychotherapy and equine-facilitated psychotherapy," *Working with traumatized youth in child welfare*, pp. 196–218, 2006.
6. I. Asimov, "Strange playfellow," *Super Science Stories*, pp. 67–77, 1940.
7. B. M. Association *et al.*, "Safeguarding vulnerable adults–a tool kit for general practitioners," 2011.
8. C. Magai, C. Cohen, D. Gomberg, C. Malatesta, and C. Culver, "Emotional expression during mid-to late-stage dementia," *International Psychogeriatrics*, vol. 8, no. 03, pp. 383–395, 1996.
9. E. C. Collins, T. J. Prescott, and B. Mitchinson, "Saying it with light: A pilot study of affective communication using the miro robot," in *Biomimetic and Biohybrid Systems*, pp. 243–255, Springer, 2015.
10. M. Ehrenfeld, "Using therapeutic dolls with psychogeriatric patients," *Play therapy with adults*, pp. 291–297, 2003.
11. G. Mitchell, "Use of doll therapy for people with dementia: an overview: Gary mitchell presents the arguments for and against this controversial, but popular, intervention," *Nursing older people*, vol. 26, no. 4, pp. 24–26, 2014.
12. M. Heerink, J. Albo-Canals, M. Valenti-Soler, P. Martinez-Martin, J. Zondag, C. Smits, and S. Anisuzzaman, "Exploring requirements and alternative pet robots for robot assisted therapy with older adults with dementia," in *Social Robotics*, pp. 104–115, Springer, 2013.

Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots

Andreas Theodorou [1] and **Robert H. Wortham** [2] and **Joanna J. Bryson** [3]

Abstract. The EPSRC's Principles of Robotics dictates the implementation of transparency in robotic systems, however, research related to it is in its infancy. The current paper introduces the reader to the need of having transparent to inspection intelligent agents. We provide a robust definition of transparency, as a mechanism to expose the decision making of the robot, by considering and expanding upon other prominent definitions found in literature. The paper concludes by addressing potentials design decisions developers need to consider when designing transparent systems.

1 INTRODUCTION

Transparency, in our opinion, is a key element relating to the ethical implications of both developing and using Artificial Intelligence, a topic of increasingly public interest and debate. We frequently use philosophical, mathematical, and biologically inspired techniques for building artificial interactive, intelligent agents, but we treat them as black-boxes with no understanding of how the underlying real-time decision making works.

The black box nature of intelligent systems, such as in context-aware applications, makes interaction limited and often uninformative for the end user [14]. Limiting interactions may negatively effect the system's performance or even jeopardize the functionality of the system. Imagine an autonomous robotic system built for providing health-care support to the elderly. However, the elderly people may be afraid and distrust the system. They may not allow the robot to interact with them. In a such scenario human lives are at risk, as they may not get the required medical treatment in time, as a human overseeing the system must detect lack of interaction and intervene. Conversely, if the human user places too much trust in a robot, it could lead to misuse, over-reliance, and disuse of the system [13]. In our example of the health care robot, if the agent malfunctions and its patients are unaware of its failure to function, the patients may continue using the robot, risking their own health. The robots in both scenarios are breaking EPSRC's first Principle of Robotics by putting human lives at risk [1].

To avoid such situations, proper calibration of trust between the humans operators and their robots is critically important, if not essential, especially in high-risk scenarios such as the usage of robots in the military or for medical purposes [9]. Calibrating trust occurs when the end-user has a mental model of the system and relies on the system within the systems capabilities and is aware of its limitation [6].

We believe that enforcement of transparency is not only beneficial for end-users, but also for intelligent agents' developers. Real-time debugging of a robot's decision making mechanism could help developers to fix bugs, prevent issues, and explain potential variance in a robot's performancee. We envision that by the correct implementation of transparency, developers could design, test, and debug their agents in real-time — similar the way in which software developers work with traditional software development and debugging.

Despite these possible benefits of transparency in intelligent systems, there is little existing research in transparent agents and even less implementation of transparent agents. Moreover, there are inconsistencies in the definitions of transparency and the criteria for a robot to be considered a transparent system. In this paper, we will present the inconsistent definitions found in the literature and attempt to compliment them with our own. Furthermore, in the third section of this paper, we will discuss the design decisions a developer needs to consider when designing transparent robotic systems.

We specifically use the term intelligent agent to denote the combination of both the software and hardware of an autonomous robotic system, working together as an actor, living in and changing the world [3]. Within this paper the words robot and agent are used interchangeably.

2 DEFINING TRANSPARENCY

Despite the predominant usage of the keyword transparency in the EPSRC Principles of Robotics, research into making systems transparent is still in its infancy. Throughout the years, very few publications have focused on the need of transparent systems and even fewer have attempted to address this need. Each study provides its own definition of the keyword, without excluding others. To date, the transparency concept has been limited to explanations for abnormal behaviour, reliability of the system, and attempts to define the analytic foundations of an intelligent system.

2.1 The EPSRC Principle of Transparency

EPSRC's Principles of Robotics considers transparency as one of its key principles, by defining transparency in robotics as: "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent." .

The EPSRC definition of transparency emphasizes keeping the end-user aware of the manufactured, mechanical, and thus artificial

---

[1] University of Bath, UK, email: a.theodorou@bath.ac.uk
[2] University of Bath, UK, email: r.h.wortham@bath.ac.uk
[3] University of Bath, UK, email: j.j.bryson@bath.ac.uk

nature of the robot. However, the phrasing used allows to consider even indirect information, such as online technical documentation, as a sufficient methodology of enforcing transparency[4]. This places the burden of responsibility with the end-user. The user will have to find, read, and understand documentation or other information provided by the manufacturer. Some user groups, such as the elderly or non-specialist users, may have issues understanding the technical terms often found in technical manuals.

2.2 Transparency as a mechanism to report reliability

One of the early publications defined transparency in terms of communicating information to the end-user, regarding the system's tendency for errors in a given context [6]. While the Dzindolet's interpretation is only a part of our definition of a transparent system, the study presents interesting findings for the importance of transparent systems. The study showed that providing extra feedback to users regarding system failures, it helped participants place their trust in the system. The users knew that the system was not 100% reliable, but they were able to calibrate their trust to the autonomous system in the experiment, as they became aware of when they could rely on it and when not to. Military usage of robotic systems is increasingly becoming more popular, especially in the form of Unmanned Aerial Vehicles (UAVs), and transparency in combat systems is essential. Imagine if an agent identifies a civilian building as a terrorist and decides to take actions against it. Who is responsible? The robot for being unreliable? Or the human overseer, who placed his trust in the system's sensors and decision making mechanism? While the EPSRC Principle of Robotics considers the human operator responsible, the damage done is irreversible. Robots working autonomously to detect and neutralize targets need to have a transparent behaviour [17]. Humans should be able to calibrate their trust to the system and in cases of combat, medical, or other scenarios where if a robot acts unreliable may harm or kill humans, transparency as a mechanism to report the system's reliability is fundamental.

2.3 Transparency as a mechanism to expose unexpected behaviour

Later studies by Kim Hinds [11] and Stumpf et. al [14], concentrated on providing feedback mechanisms to users regarding unexpected behaviour of an intelligent agent. In their studies, the user was alerted only when the agent considered that its behaviour as abnormal. Kim and Hinds' study, interestingly, showed that by increasing autonomy the importance of transparency was also increased as responsibility shifted from the user to the robot. Their results are in line with [10] research, which together demonstrate that humans are more likely to blame a robot for failures than other manufactured artefacts and coworkers.

Being able to alert the user when the robot behaves in an unexpected way is essential to achieve transparency. In high-risk situations, it could help save human lives or valuable resources by alerting a human overseer of the system to take control or calibrate its trust respectively. However, in Kim and Hinds implementation, the robot was alerting the user only when it detected that it was behaving in an unexpected way. In our opinion, this implementation tries to fix a black-box by using another. There is no guarantee that the robot is behaving unexpectedly without it knowing about its atypical behaviour. Transparency should be enforced in real-time as a always-on mechanism, allowing the user to decide if the behaviour of the agent is considered expected or unexpected.

2.4 Transparency as a mechanism to expose decision making

It is to our belief that transparency mechanisms should be built-in to the system, providing information in real time of its operation, as well as providing additional documentation as dictated by the EPSRC current principle. The intelligent agent, i.e. a robot, should contain the necessary mechanisms to provide meaningful information to the end-user. To consider a robot transparent to inspection, the end-user should have the ability to request accurate interpretations of the robot's capabilities, goals, progress in relation to the said goals, sensory inputs - situation awareness, its reliability and unexpected behaviour, such as error messages. The information provided by the robot should be presented in a human understandable format.

A transparent agent, with an inspectable decision making mechanism, could also be debugged in a similar manner to the way in which traditional, non-intelligent software is commonly debugged. The developer could see which actions the agent is making, why, and how it moves from one action to the other. This is similar to the way in which popular Integrated Development Environments (IDEs) provide options to follow different streams of code with debug points, and have abilities such as "Step-up" and "Step-in" over blocks of code.

3 DESIGNING TRANSPARENT SYSTEMSz

In this section of this paper, we will discuss the various decisions developers may face in designing a transparent system. Until now, prominent research in the field of designing transparent systems focused in presenting transparency only within the context of human-robot collaboration (HRC).Thus, they focused on designing transparent systems able to build trust between the human participants and the robot.[12]. We believe that transparency should be present even in non-collaborative environments, such as human-robot competitions [11] or even when robots are used by the military. In our view, developers should strive to develop intelligent agents, which can efficiently communicate information to the human end-user, and sequentially allow her to develop a mental model of the system and its behaviour.

3.1 Usability

In order to enforce transparency, additional displays or other methods of communication to the end-user must be carefully designed, as they will be integrating potentially complex information. Agent developers need to consider both the actual relevance and level of abstraction of the information they are exposing and how they will present this information.

3.1.1 Relevance of information

Different users may react differently to the information exposed by the robot. [16] demonstrates that end-users without a technical background neither understand nor retain information from technical inputs such as sensors. This is contrary to the agent's developer, who needs access to such information during both development and testing of the robot to effectively calibrate sensors and to fix any issues found. However, within the same study, Tullio demonstrates that users are able to understand at least basic machine learning concepts, regardless of their non-technical educational and work-history background.

Tullio's research establishes a good starting point at understanding which information maybe relevant to the user to help them understand intelligent systems. Nevertheless, further work is needed in other application areas to establish both domain-specific and user-specific trends regarding what information should be considered of importance.

3.1.2 Abstraction of information

Developers of transparent systems will need to question not only *which*, but also *how much* information they will expose to the user by establishing a level of complexity with which users may interact with the transparency-related information. This is particularly important in multi-robot systems.

Multi-robot systems allow the usage of multiple, usually small robots, where a goal is shared among various robots, each with its own sensory input, reliability, and progress towards performing its assigned task for the overall system to complete. Recent developments of biology inspired swarm intelligence allow the usage of large quantities of tiny robots working together in such a multi-robot system [15]. The military is already considering the development of swarms of autonomous little robotic soldiers. Implementing transparency in a such system is no trivial task. The developer must make rational choices about when low or high level information is required to be exposed. By exposing all information at all times, for all types of users, the system may become unusable as the user will be overloaded with information. We believe that different users will require different levels of information abstraction to avoid infobesity. Higher levels of abstractions could concentrate on presenting only an overview of the system. Instead of having the progress of a system towards a goal, by showing the current actions the system is taking in relation to achieve the said goal, it could simply present a completion bar. Moreover, in a multi-robot system, lower level information could also include the goal, sensor, goal-process, and overall behaviour of individual agents in a detailed manner. Conversely, a high-level overview could display all robots as one entity, stating averages from each machine. Intelligent agents with a design based on a cognitive architecture, such as Behaviour Oriented Design (BOD) [2], could present only high level plan elements if an overview of the system is needed. In the case of an agent designed with BOD, users may prefer to see and become informed about the states of Drives or Competencies but not individual Actions. Other users may want to see only parts of the plan in detail and other parts as a high level overview.

A good implementation of transparency should provide the user with such options, providing individuals or potential user-groups with both flexible and preset configurations in order to cater a wide range of potential users' needs. We hypothesize that the level of abstraction an individual needs is dependent on a number of factors including, but not limited to, the demographic background of the user.

1. User: We have already discussed the way in which different users tend to react differently to information regarding the current state of a robot. Similarly, we can expect that various users will respond in a similar manner to the various levels of abstraction based on their usage of the system. End-users, especially non-specialists, will prefer a high-level overview of the information available, while we expect developers to expect access to lower level of information.
2. Type of robotic system: As discussed in our examples above, a multi-robot system is most likely to require a higher level of abstraction, to avoid infobesity of the end-user. A system with a singleagent would require much less abstraction, as less data are displayed to its user.
3. Purpose of the robotic system: The intended purpose of the system should be taken into account when designing a transparent agent. For example, a military robot is much more likely to be used with a professional user in or on the loop and due to its high-risk operation, there is much greater need to display and capture as much information about the agent's behaviour as possible. On the other hand, a robotic receptionist or personal assistant is more likely to be used by non-technical users, who may prefer a simplified overview of the robot's behaviour.

3.1.3 Presentation of information

Developers needs to consider how to present to the user any of the additional information regarding the behaviour of the agent they will expose. Previous studies used visual or audio representation of the information. To our knowledge, there are no prior studies comparing the different approaches.

Autonomous robotic systems may make tens of different decisions per second. If the agent is using a reactive plan, such as a POSH plan [5], the agent may make thousands of call per minute to the different plan elements. This amount of information is hard to handle with audio-oriented systems. Moreover, visualizing the information, i.e. by providing a graphical representation of the agent's plan where the different plan elements blink as they are called, should make the system self-explanatory and easy to follow by less-technical users. Finally, a graph visualization as a means to provide transparency-related information has the additional benefits in debugging the application. The developer should be able to follow a trace of the different plan elements called, viewing the sensory input that triggered them, until a specific elements was used.

3.2 Utility of the system

So far in this paper we have expanded on the importance and design choices regarding the implementation of transparency. However, we believe the developer also needs to consider whether implementing transparency may actually damage the utility of a system. [18] argues that the the utility of an agent is measured by the degree to which it is trusted. Increasing transparency may reduce its utility. This might, for example, have a negative effect for a companionship robot or a health-care robot, designed to assist children. In such cases, the system is designed against the EPSRC Principles of Robotics, as it exploits its users feelings to increase its utility and performance on its set task.

Another important design decision which effects the system is the physical transparency of the system. The physical appearance of an agent may increase its usability [7], but also it may contrast with transparency by hiding its mechanical nature. Back in our companionship robot example, a humanoid or animal-like robot may be preferred over an agent where its mechanisms and internals are exposed, revealing its manufactured nature [8].

Discussing the trade-offs between utility and transparency is far beyond the scope of this paper. However, developers should be aware of this as they design and develop robots.

4 CONCLUSION

We strongly believe that the implementation and usage of intelligent systems which are transparent in nature can help the public understanding of AI by removing the scary mystery around why is it behaving like that. Transparency will allow to understand an agents emergent behaviour. In this paper we re-defined transparency as an always-on mechanism able to report a system's behaviour, reliability, senses, and goals as such information could help us understand the autonomous system's behaviour.

Further work is needed to test and establish good practices regarding the implementation of transparency within the robotics community. Considering the benefits of transparent systems, we strongly suggest the promotion of this key principle by research councils, such as EPSRC, and other academic communities.

ACKNOWLEDGEMENTS

We would like to thank Swen Gaudl (University of Bath) for his most valuable insights.

REFERENCES

[1] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorell, Mick Wallis, Blay Whitby, and Alan Winfield. Principles of robotics. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC), April 2011. web publication.

[2] Joanna Bryson, 'The behavior-oriented design of modular agent intelligence', in *System*, volume 2592, 61–76, (2002).

[3] Joanna J. Bryson, 'Robots should be slaves', in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, ed., Yorick Wilks, 63–74, John Benjamins, Amsterdam, (March 2010).

[4] Joanna J Bryson, Darwin Caldwell, Kerstin Dautenhahn, Paula Duxbury, Lilian Edwards, Hazel Grian, Sarah Kember, Stephen Kemp, Paul Newman, Geo Peg, Andrew Rose, Tom Rodden, Tom Sorell, Mick Wallis, Shearer West, Alan Winfield, and Ian Baldwin, 'The making of the epsrc principles of robotics', **133**(133), 14–15, (2012).

[5] Joanna J. Bryson, Tristan J. Caulfield, and Jan Drugowitsch, 'Integrating life-like action selection into cycle-based agent simulation environments', in *Proceedings of Agent 2005: Generative Social Processes, Models, and Mechanisms*, eds., Michael North, David L. Sallach, and Charles Macal, pp. 67–81, Chicago, (October 2005). Argonne National Laboratory.

[6] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck, 'The role of trust in automation reliance', *International Journal of Human Computer Studies*, **58**(6), 697–718, (2003).

[7] Kerstin Fischer, 'How people talk with robots: Designing dialogue to reduce user uncertainty', *AI Magazine*, **32**(4), 31–38, (2011).

[8] Jennifer Goetz, Sara Kiesler, and Aaron Powers, 'Matching robot appearance and behavior to tasks to improve human-robot cooperation', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 55–60, (2003).

[9] Victoria Groom and Clifford Nass, 'Can robots be teammates?', *Interaction Studies*, **8**(3), 483–500, (2007).

[10] Peter H. Kahn, Rachel L. Severson, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, and Nathan G. Freier, 'Do people hold a humanoid robot morally accountable for the harm it causes?', *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12*, (February 2016), 33, (2012).

[11] Taemie Kim and Pamela Hinds, 'Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 80–85, (2006).

[12] Joseph B Lyons, 'Being transparent about transparency : A model for human-robot interaction', *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*, 48–53, (2013).

[13] R Parasuraman and V Riley, 'Humans and automation: Use, misuse, disuse, abuse', *Human Factors*, **39**(2), 230–253, (1997).

[14] Simone Stumpf, Weng-keen Wong, Margaret Burnett, and Todd Kulesza, 'Making intelligent systems understandable and controllable by end users', 10–11, (2010).

[15] Ying Tan and Zhong-yang Zheng, 'Research advance in swarm robotics', *Defence Technology*, **9**(1), 18–39, (3 2013).

[16] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty, 'How it works: a field study of non-technical users interacting with an intelligent system', *SIGCHI conference on Human factors in computing systems (CHI'07)*, 31–40, (2009).

[17] Lu Wang, Greg a Jamieson, and Justin G Hollands, 'Trust and reliance on an automated combat identification system', *Human factors*, **51**(3), 281–291, (2009).

[18] Robert Wortham, Andreas Theodorou, and Joanna J. Bryson, 'The iron triangle: Transparency-trust-utility'. submitted, 2016.

Robot Transparency, Trust and Utility

Robert H. Wortham[1], **Andreas Theodorou**[2] **and Joanna J. Bryson**[3]

Abstract. As robot reasoning becomes more complex, debugging becomes increasingly hard based solely on observable behaviour, even for robot designers and technical specialists. Similarly, non-specialist users find it hard to create useful mental models of robot reasoning solely from observed behaviour. The EPSRC Principles of Robotics mandate that our artefacts should be transparent, but what does this mean in practice, and how does transparency affect both trust and utility? We investigate this relationship in the literature and find it to be complex, particularly in non industrial environments where transparency may have a wider range of effects on trust and utility depending on the application and purpose of the robot. We outline our programme of research to support our assertion that it is nevertheless possible to create transparent agents that are emotionally engaging despite having a transparent machine nature.

1 INTRODUCTION

The EPSRC Principles of Robotics includes a specific reference to transparency: "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent." see [1]. This initially appears to be a straightforward normative assertion, drawing on the commonly held idea that agents should not be deceptive, since deception generally leads to exploitation. This paper considers whether in fact transparency is really such a simple idea, and also whether making certain types of agents transparent reduces their utility. In considering this question, we must also address the relationship between transparency and trust.

In this paper, we use the terms robot and agent interchangeably and by these terms we mean an embodied, autonomous intelligent artefact.

What does it mean to trust a robot? We might initially simply assert that if an AI is more transparent, then we are able to trust it more, and therefore its utility increases. We could also argue that trust is only required when an agent is not fully transparent, and therefore that increased transparency reduces the need for trust [4]. If the utility of an artefact is measured by the degree to which it is trusted, then increasing transparency may reduce that utility. This might, for example, be the case for a robot that's primary function is to provide companionship.

So, we start to see that there is a complex relationship between the ideas of utility, transparency and trust. This relationship will depend on the purpose of the AI. In this paper we review the literature relating to transparency and trust, and we also describe ongoing practical research to investigate the proposal that it is indeed possible build an emotionally engaging yet transparent robot.

2 THEORY OF MIND, TRUST AND TRANSPARENCY

Although we may presuppose that communication between animals, and particularly between humans must be complex, in fact natural communication systems tend to exploit relatively simple and minimal signals, the meaning of which derives from extensive models [16]. In other words, evolution, or a shared phylogenetic history, provides adequate priors such that minimal data is required to communicate context. Although some would argue otherwise [8], it is generally agreed that effective interaction, whether coercion or cooperation, relies on each party having some theory-of-mind (ToM) of the other [16, 14]. Individual actions and composite behaviours are thus interpreted within a pre-existing ToM framework. Whether that ToM is accurate is unimportant, provided that it is predictive in terms of behaviour. The robot's transparency model does not define the ToM employed by the human user, but it is the transparency model that we can directly adjust and this is therefore the focus of this paper. It is well known that observable behaviour can communicate the internal mental states of the individual. Breazeal [2] found that implicit non-verbal communication improves transparency over that of only deliberate non-verbal communication. Here implicit is defined as conveying information inherent in behaviour but which is not deliberately communicated by the robot designer. People have strong expectations for how implicit and explicit non-verbal cues map to mental states. Breazeal also found that transparency reduces conflict when errors occur, particularly when a joint task is being attempted. Reduced conflict implies that when an error occurs during task execution, recovery is still possible with less apportionment of blame. Breazeal terms this reduced conflict Robustness, and this robustness is one effective measure of utility.

2.1 Anthropomorphism and Mental Models of Robots

Humans have a strong predisposition to anthropomorphise not only nature, but anything around them [5] — the Social Brain Hypothesis [7] may explain this phenomenon, however humans do not treat robots identically to humans, for example with respect to moral standing [10]. Although there is significant debate about the ontology of robot minds versus human minds, what is of more practical importance is how robot minds are understood psychologically by humans, i.e. what is the perceived, rather than actual, ontology. Stubbs [15] considers it essential to form a mental model of robots in order to build common ground — which we might also interpret as the basis for human trust. Stubbs [15] also found that this common ground can be effectively established via an interactive dialogue with the robot. Although this study primarily considered remote robots working in an industrial or exploratory setting, rather than robots operating in

---

[1] University of Bath, UK, email: r.h.wortham@bath.ac.uk
[2] University of Bath, UK, email: a.theodorou@bath.ac.uk
[3] University of Bath, UK, email: j.j.bryson@bath.ac.uk

domestic environments, we should take note of the importance of dialogue in establishing trust. Indeed Mueller [13] sees dialogue as one of the three main characteristics of transparent computers, the others being explanation and learning.

Meerbeek [12] investigates the relationship between a robot's perceived personality and the level to which the user feels in control during the interaction. In order to be believable, Meerbeek found that the personality expression should be linked to an internal model that deals with the behaviour (e.g. decision making) based on personality and emotion. More expressive, informal behaviour is associated with a higher perception of user control.

Non-specialist humans either have little ToM for robots, or have a model based on contemporary science fiction, and therefore interpret behaviours using a default other agent theory, which assumes the agent to share human-like motivations. This can be understood in evolutionary terms through our ancestors' need to rapidly categorise proximal activity as either neutral (the rustling of leaves in the wind), friendly (the approach of a tribe member) or hostile (the approach of a predator or foe). When sensory information is uncertain, evolving a bias towards an assumption of both agency and hostility is selective for individual longevity in an environment where one is frequently the prey, not the predator. Even in our technological environments we often experience fake agency, such as robotic dialling sales calls, automated twitter postings and auto-generated personalised spam emails.

In a study conducted in 2006 in a community hospital in the USA, the nursing staff were constantly searching for reasons why the robots acted as they did. They would ask themselves and others, "What is going on here? Is the robot supposed to do this or did I do something wrong?". This research asserts that low levels of transparency led people to question even the normal behaviours of the robot, sometimes even leading people to think of correct behaviours as errors [11].

3 RESEARCH PROGRAMME

We are beginning a programme of practical research to investigate the transparency, trust, utility triangle. Initially using non-humanoid robots, we are conducting experiments to determine the effect of various expressions of transparency on the emotional response of humans. At the heart of our experiments we are using reactive planning techniques to build autonomous agents. We have developed the Instinct reactive planner based on Bryson's Behaviour Oriented Design (BOD) approach [3]. The Instinct planner reports the execution and status of every plan element in real time, allowing us to implicitly capture the reasoning process within the robot that gives rise to its behaviour. Our experiments will investigate and demonstrate how this transparency data from the planner can be used to make the behaviour of the robot more understandable. Initially we are primarily interested in making the behaviour transparent for the robot designer, since robots with complex plans are typically very hard to design and debug. However, these initial experiments may also improve transparency for non-specialist observers.

We will subsequently investigate how we can harness the transparency mechanism embedded with the Instinct Planner to produce a more effective domestic robot. The research will investigate whether transparency makes people feel more or less bonded to their robot, and whether they are more or less able to accurately assess the needs of the robot, as it works to achieve its goals.

It is anticipated that these trials should take place within a domestic or near-domestic environment, such as a retirement home. We must gain feedback from non-specialist observers/users about the qualitative level of intelligence of the robot, and also about how comfortable they would be to have such a device in their home environment. The research will attempt to assess initial levels of fear, anxiety, mistrust of AI and robots in general, and of domestic robots in particular. Having established a reference position, transparency of the robot must be enabled by providing feedback to the user based on the real time execution within the reactive planner. The methods we currently envisage are:

- Real-time presentation of textual statements relating to plan execution.
- Graphical real-time visualisation of plan execution.
- Audio (i.e. verbal) statements relating to robot plan execution.

For each of these methods the transparency information could either be presented on/from a remote device, or on/from the robot itself. There are thus six possible combinations. Of course additional transparency fusion, such as audio combined with graphical, could also be tested based on the success or failure of initial experimental results.

As the literature indicates that dialogue is important in establishing trust, this research should give some consideration to the possibility of accepting speech input, albeit restricted to simple commands, as a means for users to inquire of the robot what it is doing, and to have the robot respond appropriately.

4 DISCUSSION

EPSRC Principle 1 asserts that robots are tools. Within industrial and engineering environments this is fairly clear, in the sense that a human uses the robot to complete a technical task. The designer and user of the robot share the goal of the robot: to complete the task. However, within domestic and healthcare environments, robots may have rather a different relationship with those they interact with. They may be intended to provide companionship and simultaneous covert monitoring of patient well-being. They may be tools for the healthcare professional, but for the patient they are companions. In such an environment the utility may be negatively affected by increased transparency. Our sense of companionship is related to the measure of agency we project onto the robot. If we are able to understand the workings of the intelligence does it inherently appear to become less intelligent in the folk sense, such that we then project less agency, and as a result experience less benefit from the robot? We might compare this with television. We know it has no agency, but its presence in the corner of our sitting room does provide companion like benefits. Maybe this has to do with the conscious suspension of disbelief, or maybe we have an unconscious agency detector which is more easily fooled by technology.

Common-sense notions of intelligence are conflated with folk psychology ideas of agency and also of living. Things that are intelligent are alive, in the sense that they have their own beliefs, desires and intentions that we understand are fundamentally self serving, or selfish. We implicitly recognise selfishness as a fundamental characteristic of all life [6]. If such an agent engages with us then it considers us to be important in the pursuit of these selfish objectives. Such agents are worthy of becoming our companions because they ascribe true value in their relationship with us, and this increases our value in society. Conversely, agents who have no self-serving agency are not worthy of our attention because they convey no social value. Perhaps therefore, artificial agents whose sole purpose is companionship and are truly transparent in this respect are thus disqualified from being worthy companions. In some situations robot transparency may therefore be at odds with utility, and more generally it may be orthogonal rather than beneficial to the successful use of the robot. Whilst we may invent scenarios and continue to discuss the theoretical and philosophical interplay between transparency, trust and utility, as scientists we await the outcome of our experiments.

5 CONCLUSION

We have seen that unpacking transparency and trust is complex, but can be partly understood by looking at how humans come to understand and subsequently trust one another, and how they overcome evolutionary fears in order to trust other agents, through implicit non-verbal communication. Unacceptable levels of anxiety, fear and mistrust will result in an emotional and cognitive response to reject robots. Hancock [9] asserts that if we cannot trust our robots, we will not be able to benefit from them effectively. However, given that we happily interact in society with others whom we do not completely trust, and increasingly we interact with computers knowing that their recommendations maybe faulty, we must conclude that Hancock is over simplifying. Finally, there may be applications where transparency is at odds with utility. Our ongoing programme of research is intended to validate our hypothesis that we can indeed create transparent robots that are nevertheless emotionally engaging and useful tools across a wide range of domestic and near-domestic environments. Meanwhile, there remains a great of work to be done to unpack the relationship between transparency, utility and trust.

REFERENCES

[1] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorell, Mick Wallis, Blay Whitby, and Alan Winfield. Principles of robotics. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC), April 2011. web publication.

[2] C. Breazeal, C.D. Kidd, A.L. Thomaz, G. Hoffman, and M. Berlin, 'Effects of nonverbal communication on efficiency and robustness in human-robot teamwork', in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 708–713, Alberta, Canada, (2005). Ieee.

[3] Joanna J. Bryson, 'Intelligence by design: principles of modularity and coordination for engineering complex adaptive agents', (2001).

[4] Joanna J Bryson and Paul Rauwolf, 'Trust, Communication, and Inequality'. 2016.

[5] Kerstin Dautenhahn, 'Methodology & themes of human-robot interaction: A growing research field', *International Journal of Advanced Robotic Systems*, **4**(1 SPEC. ISS.), 103–108, (2007).

[6] Richard Dawkins, 'Hierarchical organisation: A candidate principle for ethology', in *Growing Points in Ethology*, eds., P. P. G. Bateson and R. A. Hinde, 7–54, Cambridge University Press, Cambridge, (1976).

[7] R I M Dunbar, 'The Social Brain Hypothesis', *Evolutionary Anthropology*, 178–190, (1998).

[8] Shaun Gallagher, 'The narrative alternative to theory of mind', in *Radical Enactivism: Intentionality, Phenomenology, and Narrative*, ed., R Menary, number Gallagher 2001, 223–229, John Benjamins, Amsterdam, (2006).

[9] P. a. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, 'A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **53**(5), 517–527, (2011).

[10] Peter H. Kahn, Hiroshi Ishiguro, Batya Friedman, and Takayuki Kanda, 'What is a human? - Toward psychological benchmarks in the field of human-robot interaction', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, **3**, 364–371, (2006).

[11] Taemie Kim and Pamela Hinds, 'Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 80–85, (2006).

[12] Bernt Meerbeek, Jettie Hoonhout, Peter Bingley, and Jacques Terken, 'Investigating the relationship between the personality of a robotic TV assistant and the level of user control', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 404–410, (2006).

[13] Erik T. Mueller, *Transparent Computers: Designing Understandable Intelligent Systems*, Erik T. Mueller, San Bernardino, CA, 2016.

[14] Rebecca Saxe, Laura E Schulz, and Yuhong V Jiang, 'Reading minds versus following rules: dissociating theory of mind and executive control in the brain.', *Social neuroscience*, **1**(3-4), 284–98, (jan 2006).

[15] Kristen Stubbs, Pamela J Hinds, and David Wettergreen, 'Autonomy and Common Ground in Human-Robot Interaction: A Field Study', *IEEE Intelligent Systems*, **22**(2), 42–50, (2007).

[16] Robert H Wortham and Joanna J Bryson, 'Communication', in *Handbook of Living Machines* {*in press.*}, Oxford University Press, Oxford, (2016).