Project Proposal:
Re-thinking the Ethics of Technology, from a Conceptual Perspective

Abstract: This project addresses the following interrelated core questions in contemporary discussions within ethics of technology: (1) Are present-day artefacts moral agents: i.e., do they have the capacity to take responsibility for their actions? (2) If the answer to (1) is no, is it conceivable that future artefacts may appropriately come to be treated as moral agents? (3) If it is conceivable, then what can one say about the likelihood? (4) What standards should one hold such an agent to? (5) Is it possible to hardwire in such standards? If not, what are the alternatives? (6) What do the standards one seek to hold AIs to say about the moral standards one ought to apply to ourselves? (7) What do these standards reveal about the well-known "fact/value" divide and the seeming bifurcation on the factual side between truth and falsehood? Are there corresponding "correct" and "incorrect" values?

I. Introduction

Existing artefacts – from "autonomous" battlefield robots to self-driving cars to sexbots – are, appropriately or not, already being treated as moral agents capable of deciding moral questions and assuming at least some minimal responsibility for their actions. Whether or not such artefacts qualify as "true" moral agents – I have argued (Parthemore & Whitby 2014, 2013) that they are not – the field of artificial general intelligence is promising future artefacts that will be full-fledged exemplars; and there is no *a priori* reason to judge such entities impossible. It seems prudent to discuss now what such entities should look like.

Colin Allen and colleagues (Allen *et al.* 2010) have proposed a Moral Turing Test (MTT) as an analogue to Alan Turing's Imitation Game, to determine whether a given artefactual agent qualifies as a moral agent. Not only is such a test deeply ethically problematic – most people would not (in most instances) apply such a test to their fellow human beings, to determine whether they qualified as moral agents, so why should it be appropriate for an artefact? – but also impractical. What should one be meant to do with the results, other than to claim (on potentially dubious grounds) that the age of artefactual moral agents had finally arrived? For many if not most people, moral agency is – intuitively – not about a passing grade on a test. Too, a test like the MTT offers nothing to address people's often deeply held intuitions whether or not artefacts can, even in principle, qualify as moral agents.

In place of a litmus test for moral agency, I propose using the tools of Conceptual Spaces Theory (CST: a prototype-based theory about the nature of systematically and productively structured thought, in the tradition of Eleanor Rosch 1975,1999; see Gärdenfors 2000) along with its extension in the Unified Conceptual Space Theory (UCST; see Parthemore 2014) to map out the nature and extent of an artefact's (or any human agent's) alleged moral agency. What moral concepts does the agent show evidence of possessing, what contexts do they apply within, and what larger moral framework do they fit into?

Moral agency is a particularly useful application area for testing the more general usefulness of theories like CST and UCST, which aim (*pace* Edouard Machery 2000) to establish a common pattern for all systematically and productively structured thought: which is to say, all concepts. That's because of the way moral agency brings together very abstract concepts like rights and responsibilities with concrete decisions and consequences. Empirically testable

models can, in principle, be created using mind-mapping software based on UCST, a prototype for which already exists. Such software can produce a kind of abstract picture of an artefactual or human agent's moral territory – the better to understand the nature and minimum requirements for moral agency, be it artefactual or natural, with implications for the increasingly sophisticated AI agents human beings come to construct.

II. Deliverables

The deliverables will consist of:

a) A metric for mapping out an agent's moral agency along the lines described above, with the intention of applying it to future empirical studies of moral concepts.
b) A set of papers, described below.

2.1 Artefactual ethics as opportunity for rethinking "natural" ethics

This paper argues that, within the ethics community, the wider philosophical establishment and society in general, people have been far too lax about what to accept as morally "right" behaviour – far too quick to let themselves and, all too often, each other off the hook. By drawing comparisons to artefactual behaviour and the objections people raise to calling that behaviour the morally acceptable behaviour of authentic moral agents, this paper lays out a framework by which human ethics and metaethics can more fruitfully be approached. An earlier paper of ours argued that, for an action to be morally right, one must have a convergence of the right motivations, the right means, and the right consequences. The underlying insight is that deontological, virtue-ethics-based, and consequentialist accounts all have their necessary role to play, but each tends to get too focused on itself and its merits to the loss of the bigger picture; while utilitarian accounts, as perhaps the most prominent division within consequentialism, face the further problem of failing to allow for those occasions where the needs of the few, or the one, outweigh the needs of the many, as Ursula K. LeGuin (1973) so devastatingly addressed.

Although the requirement to align motivations, means, and consequences may seem impossibly onerous, it need not be, provided one is prepared to allow that moral behaviour is far more difficult to achieve, either for artefacts or human beings, than it might seem at first glance. Mistakes will be made, and perhaps it matters more to take responsibility for those mistakes than to assure oneself, despite reasonable argument to the contrary, that one has avoided them. It is time to hold artefactual and natural agent alike to a higher standard.

2.2 Revisiting the Trolley Problem, with lessons for AI: No one right answer; many wrong answers

Too many discussions of the Trolley Problem, in all its permutations, presuppose that there is one right solution, in line with a general attitude that, with values as with facts, there is one right answer, and it is knowable. Both, of course, are open to challenge: the truth, morally or otherwise, may neither be knowable nor singular; in particular, the indisputable existence of wrong answers does not establish that there is one fact of the matter, in the moral or the scientific domain, for matters of more than trivial complexity – however intuitively appealing or even self-evidently true that people might find the idea. The existence of one account that is taken to be right does not, logically, preclude the possibility of another account that appears, to human understanding, to be in conflict with it and yet also be right insofar as it is

supported by all the available evidence. Indeed, *pace* Bertrand Russell, this is what separates genuine paradox from contradiction, and why a paradox is meaningful (even if its ultimate truth escapes us) while a contradiction is not. This paper uses the Trolley Problem as a foil to argue for four things:

1. Truth may not be so clear in the moral as in the scientific domain, but neither is it purely subjective. Turning that around, truth may not be so obviously subjective in the scientific as in the moral domain, but neither is it purely objective.
2. Truth, in either the moral or scientific domain, is best understood when left open to the possibility (if not, indeed, necessity) of pluralism.
3. Both the fact/value divide and the mind/body problem, as typically presented, are based on a misunderstanding and both, from a certain critical perspective, can be seen to dissolve, to be replaced by a continuum.
4. The preceding points, if true, have profound implications for how one may go about designing future artefactual moral agents.

The bottom line is that no one set of "right" moral precepts can ever be arrived at for "hard wiring" into any agent: artefactual or natural. One's goal instead should be to train the potential moral agent in the general ways of moral agency and give that agency the capacity to take responsibility for its actions and their consequences.

2.3 Beyond objectification: From robots as sexual partners to a new theory of personhood

This paper is a fully updated version of a paper presented at the 2016 convention of the Society for the Study of Artificial Intelligence and Behaviour, in a symposium entitled Artificial Sexuality. That conference paper can be found here. Robots as sex toys and *faux* companions are available now. When, and how, will they stop being glorified dolls and become not our sexual playthings but our partners? This paper argues that sexual objectification is based on a fundamental misunderstanding of our physical nature that has filtered down from academic into popular discourse, reducing people to "nothing more" than physical objects. The proposed solution is an updated form of neutral monism (in particular, a form of dual-aspect monism) in contrast to the prevailing appeal of reductive physical monism. Although the state of the art in robotic sex is almost laughably primitive, nevertheless, in contemplating the future development of robots as sex workers, there are powerful opportunities for them to play a transformative role in our understanding of ourselves and what it means to be human.

2.4 Libertarianism and artefactual agency: Looking for free will in all the wrong places

This paper is a fully updated version of a paper presented at the 2014 Toward a Science of Consciousness conference. Using as a springboard the common misconception, traceable to Ada Lovelace, that computers and related artefacts "can only do what they are told", this paper argues that the difference between extant artefacts and human beings is that artefacts make mistakes and appear creative in relatively simple, uninteresting ways compared to human beings. The even bigger difference is that extant artefacts lack the capacity for free will, which libertarians on free will (*pace* the compatibilists and other determinists) take to be the capacity to generate some small but significant causal force in the moment of making a decision such that that causal force is heavily constrained *but not fully predetermined by* the

state of the world leading up to the moment of the decision. Existing artefacts lack that capacity because they are not self-consciously reflective artificial life.

2.5 Designing artefactual agents, in light of the limits of cognition and conceptual agency

This paper argues that there are real and knowable limits to human cognition and conceptual agency: in particular, to the capacity of the human mind to understand itself and its nature. Complete and consistent understanding of the human mind, or of consciousness, or of human nature would invite not paradox but outright contradiction: the equivalent of the dragon swallowing its own tail. This does not mean – as some might too hastily suppose – that human beings cannot create something that is their intellectual equivalent; indeed, they can and do, every time they have children. What it does mean is that, if human beings *do* succeed in creating their intellectual equals, then by the very virtue of that accomplishment, their ability to understand their creations will be limited. Human beings have a proud and long intellectual history of creating things that outstrip their capacity to explain or explain fully; AGIs, including artefactual moral agents, are logically guaranteed to fall into this category.

III. Bibliography

Allen, C., G. Varner & J. Zinser (2010). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence,* **12**(3): 251-261.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.

LeGuin, U.K. (1973). The ones who walk away from Omelas. In R. Silverberg (ed.), *New Dimensions, Volume 3*. Nelson Doubleday. Available online from https://sites.asiasociety.org/asia21summit/wp-content/uploads/2011/02/3.-Le-Guin-Ursula-The-Ones-Who-Walk-Away-From-Omelas.pdf (accessed 30 November 2020).

Machery, E. (2009). *Doing Without Concepts*. Oxford University Press.

Parthemore, J. (2014). Specification of the unified conceptual space, for purposes of empirical investigation. In P. Gärdenfors & F. Zenker (eds.), *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation* (223-244).

Parthemore, J. & B. Whitby (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why) and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness,* **6**(2): 141-161. http://doi.org/10.1142/S1793843014400162.

Parthemore, J. & B. Whitby (2013). What makes any agent a moral agent?: Reflections on machine consciousness and moral agency, *International Journal of Machine Consciousness*, **5**(2): 105-129.

Rosch, E. (1999). Principles of categorization. In E. Margolis & S. Laurence (eds.), *Concepts: Core Readings* (189-206).

Rosch, E. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology,* **7**(4): 573-605.