AISB/IACAP World Congress 2012

Birmingham, UK, 2-6 July 2012

THE MACHINE QUESTION: AI, ETHICS AND MORAL RESPONSIBILITY

David J. Gunkel, Joanna J. Bryson, and Steve Torrance

(Editors)

Part of

Published by
The Society for the Study of
Artificial Intelligence and
Simulation of Behaviour

http://www.aisb.org.uk

ISBN 978-1-908187-21-5

Foreword from the Congress Chairs

For the Turing year 2012, AISB (The Society for the Study of Artificial Intelligence and Simulation of Behaviour) and IACAP (The International Association for Computing and Philosophy) merged their annual symposia/conferences to form the AISB/IACAP World Congress. The congress took place 2–6 July 2012 at the University of Birmingham, UK.

The Congress was inspired by a desire to honour Alan Turing, and by the broad and deep significance of Turing's work to AI, the philosophical ramifications of computing, and philosophy and computing more generally. The Congress was one of the events forming the Alan Turing Year.

The Congress consisted mainly of a number of collocated Symposia on specific research areas, together with six invited Plenary Talks. All papers other than the Plenaries were given within Symposia. This format is perfect for encouraging new dialogue and collaboration both within and between research areas.

This volume forms the proceedings of one of the component symposia. We are most grateful to the organizers of the Symposium for their hard work in creating it, attracting papers, doing the necessary reviewing, defining an exciting programme for the symposium, and compiling this volume. We also thank them for their flexibility and patience concerning the complex matter of fitting all the symposia and other events into the Congress week.

John Barnden (Computer Science, University of Birmingham)
Programme Co-Chair and AISB Vice-Chair
Anthony Beavers (University of Evansville, Indiana, USA)
Programme Co-Chair and IACAP President
Manfred Kerber (Computer Science, University of Birmingham)
Local Arrangements Chair

Foreword for The Machine Question

One of the enduring concerns of moral philosophy is deciding who or what is deserving of ethical consideration. Although initially limited to "other men," the practice of ethics has developed in such a way that it continually challenges its own restrictions and comes to encompass what had been previously excluded individuals and groups—foreigners, women, animals, and even the environment. Currently, we stand on the verge of another, fundamental challenge to moral thinking. This challenge comes from the autonomous, intelligent machines of our own making, and it puts in question many deep-seated assumptions about who or what constitutes a moral subject. The way we address and respond to this challenge will have a profound effect on how we understand ourselves, our place in the world, and our responsibilities to the other entities encountered here.

We organised this symposium and the proceedings you find here because we believe it is urgent that this new development in moral thinking be advanced in the light and perspective of ethics/moral philosophy, a discipline that reflects thousands of years of effort by our species' civilisations. Fundamental philosophical questions include:

- What kind of moral claim might an intelligent or autonomous machine have?
- Is it possible for a machine to be a legitimate moral agent and/or moral patient?
- What are the philosophical grounds supporting such a claim?
- And what would it mean to articulate and practice an ethics of this claim?

The Machine Question: AI, Ethics and Moral Responsibility seeks to address, evaluate, and respond to these and related questions.

The Machine Question was one of three symposia in the *Ethics, Morality, AI and Mind* track at the AISB / IACAP 2012 World Congress in honour of Alan Turing. The ethics track also included the symposia *Moral Cognition and Theory of Mind* and *Framework for Responsible Research and Innovation in Artificial Intelligence*. We would like to thank all of the contributors to this symposium, the other

symposia, the other symposia's organisers and the Congress leaders, particularly John Barnden for his seemingly tireless response to our many queries. We would also like to thank the keynote speakers for the Ethics track, Susan and Michael Anderson, whose talk *The Relationship Between Intelligent, Autonomously Functioning Machines and Ethics* appeared during our symposium, for agreeing to speak at the congress.

The story of this symposium started a little like the old Lorne Greene song *Ringo*, except instead of anyone saving anyone's life, Joanna Bryson loaned David Gunkel a laundry token in July of 1986, about a month after both had graduated with first degrees in the liberal arts and independently moved in to *The Grandeur*, a cheap apartment building on Chicago's north side. Two decades later they discovered themselves on the opposite sides of not the law, but rather the AI-as-Moral-Subject debate, when Gunkel contacted Bryson about a chapter in his book *Thinking Otherwise*. The climactic shoot-out took place not between two isolated people, but with the wise advice and able assistance of our third co-editor Steve Torrance, who brought our symposium the experience of running many previous AISB symposia, and between a cadre of scholars who took time to submit to, revise for and participate in this symposium. We thank every one of them for their contributions and participation, and you for reading this proceedings.

David J. Gunkel (Department of Communication, Northern Illinois University)
Joanna J. Bryson (Department of Computer Science, University of Bath)
Steve Torrance (Department of Computer Science, University of Sussex)

Contents

1 Joel Parthemore and Blay Whitby — *Moral Agency, Moral Responsibility, and Artefacts* 8

2 John Basl — *Machines as Moral Patients We Shouldn't Care About (Yet)* 17

3 Benjamin Matheson — *Manipulation, Moral Responsibility and Machines* 25

4 Alejandro Rosas — *The Holy Will of Ethical Machines* 29

5 Keith Miller, Marty Wolf and Frances Grodzinsky — *Behind the Mask: Machine Morality* 33

6 Erica Neely — *Machines and the Moral Community* 38

7 Mark Coeckelbergh — *Who Cares about Robots?* 43

8 David J. Gunkel — *A Vindication of the Rights of Machines* 46

9 Steve Torrance — *The Centrality of Machine Consciousness to Machine Ethics* 54

10 Rodger Kibble — *Can an Unmanned Drone Be a Moral Agent?* 61

11 Marc Champagne and Ryan Tonkens — *Bridging the Responsibility Gap in Automated Warfare* 67

12 Joanna Bryson — *Patiency Is Not a Virtue: Suggestions for Co-Constructing*

an Ethical Framework Including Intelligent Artefacts **73**

13 Johnny Søraker — *Is There Continuity Between Man and Machine?* **78**

14 David Davenport — Poster: *Moral Mechanisms* **83**

15 Marie-des-Neiges Ruffo — Poster: *The Robot, a Stranger to Ethics* **87**

16 Mark Waser — Poster: *Safety and Morality* Require *the Recognition of Self-Improving Machines as Moral/Justice Patients and Agents* **92**

17 Damien P. Williams — Poster: *Strange Things Happen at the One Two Point: The Implications of Autonomous Created Intelligence in Speculative Fiction Media* **97**

Moral Agency, Moral Responsibility, and Artefacts: What Existing Artefacts Fail to Achieve (and Why), and Why They, Nevertheless, Can (and Do!) Make Moral Claims Upon Us

Joel Parthemore [1] **and Blay Whitby**[2]

Abstract. This paper follows directly from our forthcoming paper in *International Journal of Machine Consciousness,* where we discuss the requirements for an artefact to be a moral agent and conclude that the artefactual question is ultimately a red herring. As we did in the earlier paper, we take moral agency to be that condition in which an agent can, appropriately, be held responsible for her actions and their consequences. We set a number of stringent conditions on moral agency. A moral agent must be embedded in a cultural and specifically moral context, and embodied in a suitable physical form. It must be, in some substantive sense, alive. It must exhibit self-conscious awareness: *who does the "I" who thinks "I" think that "I" is*? It must exhibit a range of highly sophisticated conceptual abilities, going well beyond what the likely majority of conceptual agents possess: not least that it must possess a well-developed *moral space* of reasons. Finally, it must be able to communicate its moral agency through some system of *signs:* a "private" moral world is not enough. After reviewing these conditions and pouring cold water on a number of recent claims for having achieved "minimal" machine consciousness, we turn our attention to a number of existing and, in some cases, commonplace artefacts that lack moral agency yet nevertheless require one to take a moral stance toward them, *as if* they were moral agents. Finally, we address another class of agents raising a related set of issues: autonomous military robots.

1 INTRODUCTION

> As a moral community, humans do often concede that other creatures have some level of ethical status, and they do so (at least in part) because of their possession of cognitive and phenomenological capacities in varying degrees [47].

In talking of *ethical status,* Steve Torrance's concern includes (though is not limited to) what we are calling moral agency. He explicitly intends to include artefacts. It follows: what "cognitive and phenomenological capacities" are required?

For our purposes, a *moral agent* is an agent whom one appropriately holds responsible for its actions and their consequences, and *moral agency* is the distinct type of agency that agent possesses. *Contra* the usage of someone like Wendell Wallach [49], it is not enough on our usage that an agent does the appropriate things: i.e., produces the correct consequences. It must do so for the appropriate reasons and using the appropriate means. Otherwise, it may be appropriate – even morally required in some circumstances (see Section 5) – to treat the agent, at least to some limited extent, *as if* it were a moral agent / possessed moral agency; nevertheless, that agent will not be a moral agent. This alignment of motivations, means, and consequences for attributing moral agency matches an alignment of motivations, means, and consequences we see (*pace* the utilitarians and most other consequentialists) as essential to "doing the morally right thing"– though any further discussion of that last point is necessarily beyond the scope of this paper.

In [28], we set out a number of conceptual requirements for possessing moral agency as well as grounds for appropriately attributing it, as a way of addressing the claims and counterclaims over so-called artificial morality and machine consciousness. Our conclusion was that concerns over artefactual moral agency and consciousness were useful for initiating discussion but ultimately a distraction from the bigger question of what it takes for *any* agent to be a conscious moral agent.

Section Two summarizes the requirements for *possessing* moral agency and the requirements for *attributing* it appropriately. Section Three summarizes our proposal from [28] for operationalizing those requirements and "mapping out" the moral space, as an alternative to any litmus test for moral agency, such as the Moral Turing Test proposed by Allen and colleagues [2]. Section Four introduces, and debunks, a number of recent claims for having achieved "minimal" machine consciousness (since consciousness is one of our requirements for possessing moral agency). Sections Five and Six introduce two groups of existing artefacts that, we believe, raise important concerns relating to moral agency: one that people *should* take a moral stance toward (even though they often do not!), and one that they *should not* (even though they often do!). Section Seven summarizes the take-home message.

2 THE REQUIREMENTS OF MORAL AGENCY

> *The Semiotic Hierarchy…* distinguishes between four major levels in the organization of meaning: *life, consciousness, sign function,* and *language,* where each of these, in this order, both rests on the previous level, and makes possible the attainment of the next [56, p. 169].

Moral agency depends, at its foundations, on *moral meaning*: hold-

[1] Centre for Cognitive Semiotics, University of Lund, Sweden; email: joel.parthemore@semiotik.lu.se.
[2] Centre for Research in Cognitive Sciences, University of Sussex, UK; email: blayw@sussex.ac.uk.

ing an agent to account for her actions assumes that those actions are morally meaningful both to the agent and her community of observers. Moral meaning is one instance of a much broader notion of meaning, which is characterized by the salience-laden interaction between agent(s) and environment (*cf.* [55, p. 258]).

In setting out the requirements of moral agency, this paper proposes a nested succession of dependencies very much like Jordan Zlatev's *semiotic hierarchy* [56, 55], with a few additional steps. In establishing life (Section 2.2) as the foundation for meaning, Zlatev assumes as prerequisites both *embeddedness* and *embodiment* (Section 2.1). We choose to spell that out. Meanwhile, for our purposes we need to distinguish different levels of consciousness: non-reflective from pre-reflective from "full" self-conscious awareness; where all levels of consciousness presuppose *conceptually structured thought*, and conceptually structured thought presupposes consciousness. This allows us to discuss concepts where Zlatev talks of consciousness (Section 2.3), with full self-consciousness as a distinct level that transforms and extends conceptual abilities (Section 2.4). Finally, while Zlatev places language at the top of his hierarchy, we are able to stop at the sign function (Section 2.5), because, although the moral agent must be able to communicate evidence of her moral agency, she need not necessarily do so through language.

2.1 The moral agent must be embedded and embodied.

There is no brain in a vat – to make reference to Hilary Putnam's classic thought experiment [31] – and there is no moral agent in a moral vacuum. Moral agency depends, critically, on the existence of other moral agents and a shared space of moral reasoning in which it is embedded. Even if some part of it is, in some sense, private to the individual, some other part is intrinsically a part of *social* cognition: indeed, the kind of social cognition that Pierre Steiner and John Stewart have identified [44][3] as *not* beginning with or reducing to an agglomeration of individuals but social from the beginning.

Moral agents are not just *embedded* in the right kind of physical and cultural environment; they are *embodied* in a suitable physical form that allows them to carry out the actions for which one holds them accountable and give evidence for *why* one should hold them accountable. When Peter Gärdenfors defines embodied meaning by writing that "meaning is not independent of perception or of bodily experience" [16, p. 160], he clearly means to include moral meaning.

In the spirit of the enactivist philosophers, we would like to go beyond embeddedness and embodiment by stressing the continuity between the moral agent and her moral environment[4]: between her personal moral space and the shared moral space in which she moves. To paraphrase Evan Thompson: "the roots of moral life lie not simply in the brain, but ramify through the body and environment" (*cf.* [45, p. ix]). To steal a line from Putnam, "morality ain't (just) in the head" (*cf.* [30, p. 227]).

2.2 The moral agent must be alive.

> . . . *All living systems and only living systems are capable of meaning.* This is so because life implies the presence of intrinsic value, which constitutes the necessary and sufficient condition for meaning [55, p. 256].

> . . . One needs a clear way of characterizing what distinguishes living systems from nonliving ones. Such a characterization could. . . serve as a standard or criterion for recognizing life elsewhere on other planets, or for determining whether anything we might someday synthesize artificially would qualify as living [45, p. 95].

Moral agency presupposes a number of things that are not, directly, part of its definition. Drawing inspiration from the work of Francesco Varela and Humberto Maturana [22] and Evan Thompson, and even more directly from Zlatev's semiotic hierarchy, we claim that a moral agent must necessarily be alive, by some definition of "life" – even though, for our purposes, the agent need be neither conventionally biologically formulated nor naturally evolved. Indeed, although history *is* critical for all manner of practical reasons, the agent could – at least in principle – have a highly non-standard history as e.g. Swamp Man (a position that Zlatev explicitly rejects in [54] and even more strongly in [56]). *Autopoiesis* – intended, at least in its origins, as expressing both the *necessary* and *sufficient* conditions for life – offers a convenient way out of any overly narrow biological perspective as e.g. John Searle [39] might be accused of.

Maturana and Varela define an autopoietic system as a type of *homeostatic machine*; specifically:

> . . . A machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as a network [21, pp. 78-79].

In their choice of terminology, Maturana and Varela quite deliberately want to avoid prejudicing matters toward those living systems we happen currently to be familiar with, all of which are based on DNA, organized into cells either with (*eukaryotic*) or without (*prokaryotic*) nuclei, reproduce either sexually or asexually, and so on. Both *allopoietic* artefacts and *autopoietic* organisms are machines: by which Maturana and Varela mean that they are defined by their abstract organization, not by their concrete physical realization. ". . . The organization of a machine is independent of the properties of its components which can be any, and a given machine can be realized in many different manners by many different kinds of components" [21, p. 77].

Autopoietic systems define their own (selectively permeable) boundary; allopoietic systems have their boundary set for them by some observer[5]. Autopoietic systems are *organizationally closed*: i.e., they are *far-from-equilibrium* systems whose structure is determined and maintained solely by processes located within the system; allopoietic systems are organizationally open. Autopoietic systems are autonomous in a strong sense: they are "continually self-producing" [22, p. 43] and adaptive [13]; allopoietic systems give at most the appearance of autonomy (see Section 6).

[3] For a similar view, see [18].

[4] *Cf.* the discussion in [26, p. 89].

[5] The distinction here is actually quite subtle: for both types of system, the boundary can only be *identified* by an observer (see e.g. [20, p.30] for why the observer is indispensable); and for both, the boundary masks an underlying continuity: i.e., both *autopoietic* and *allopoietic* systems exist – or can be viewed – as parts of larger *allopoietic* systems that are, in some substantive sense, primary.

2.3 The moral agent must be a sophisticated conceptual agent.

On our view, concepts and consciousness are two sides of one coin: no concepts without (some level of) consciousness; no consciousness without (some level of) conceptually structured thought. However precisely one defines consciousness, there is an implicit assumption that conscious thought is systematically, productively[6], and compositionally structured[7], and under the agent's *endogenous control* [29, p. 197].

Likewise, concepts and conceptual abilities are two sides of a coin, depending on whether one's attention is more on *knowing that* or *knowing how* [37]. To say that an agent possesses and successfully employs certain concepts is to say that that agent has certain corresponding abilities; to say that an agent has certain conceptual abilities is to say that that agent possesses and successfully employs certain concepts. Note that, depending on the level of its conceptual abilities, a given conceptual agent may or may not make such a distinction itself. Some conceptual agents' conceptual abilities will be more on the *knowing how* side of the ledger, with little if any reflective awareness of those abilities. (That is to say, the *knowing that* side of the ledger may only be evident to observers.) On the other hand, human beings, as sophisticated conceptual agents, can – and, we believe, cannot help but – make the distinction.

Finally, for our purposes, and following directly from the *theory of concepts* underlying our discussions – the *unified conceptual space theory* discussed in [27, 25, 28] – *types* (abstract) and *tokens* (concrete) are two sides of a coin: every type is a specific token of some more general type, and every token can (within the limits of a conceptual agent's conceptual abilities) be the type for some yet more specific tokenings. Meanwhile, concepts both *abstract away* from the particulars of any given context and yet are *always applied in* a specific context.

Moral agents, as a sub-category of conceptual agents, must possess a number of concepts / conceptual abilities that go well beyond what the vast majority of conceptual agents likely possess – at least if one is inclined, as we are, to extend conceptual abilities to a number of non-human species (see e.g. [1, 24]). Most importantly, she must possess a ***concept of morality*** as both type and token: both a general category within its conceptual knowledge, constituting a *moral space of reasoning*, and a specific instance of more foundational intersubjective understanding; both a general guide to how to be a moral agent and a specific guide on how to act in any given circumstances; both a set of abstract principles and a set of concrete percepts (*"thou shalt not kill"*).

The moral agent must possess an *explicit* concept of morality: what it means, to her, to be a moral agent; and an *explicit* ***concept of self***: who and what she thinks she is (see Section 2.4). She must also – and probably most controversially – possess an at least *implicit* concept of concept itself.

Care must be taken here: after all, many people have only a vague explicit understanding of what a concept or notion or idea is. The point is that a moral agent must be able to refocus her attention from the usual objects of her thoughts to the thoughts themselves. She must be able to reflect on the rightness and wrongness of her thoughts and deeds. To do this, she must be sensitive to – implicitly aware of – the systematically, productively, compositionally, etc. structured nature of her thoughts, all of which are prerequisite to any implicit understanding of the systematicity, productivity, compositionality, etc., of her *moral* thoughts. She must have a practical, *knowing-how*-type understanding of the connection between *doing* ***x*** *is wrong* and *doing* ***y*** *is wrong*, where $x \subset y$; or between ***y*** *doing* ***x*** *to me is wrong* and *my doing* ***x*** *to* ***y*** *is wrong*. She must be able to proceed, by induction, from *doing* ***w*** *is wrong, doing* ***x*** *is wrong,* and *doing* ***y*** *is wrong* to *doing* ***Z*** *is wrong*, where $w, x, y \subset Z$.

[6] *Systematicity* and *productivity* are covered under Gareth Evans' Generality Constraint: [14, pp. 100-104].

[7] "Concepts are the constituents of thoughts and, in indefinitely many cases, of one another" [15, p. 25].

2.4 The moral agent must be actively self-aware.

> The most obvious self consciousness isn't just consciousness, it's consciousness of the self, something that obviously requires a capacity for consciousness *and* a concept of self [9, p. 17].

Key to any conceptual agent's conceptual abilities – we believe – is an at least implicit concept of self. In the simplest case, this can be no more than what Zlatev [54, p. 173], following Ulric Neisser [23], calls the *ecological self*: an "initial self-awareness", which "acts here and now… but remains unreflected upon"; or what Antonio Damasio [10] terms the *core self*. The concept of self is like and yet different from all the other concepts a conceptual agent possesses, in that this concept applies to the agent herself: it is intimately and uniquely self-referential. This special status gives it a certain priority – even, perhaps, a kind of authority – over all the agent's other concepts. Call it the keystone: the concept that allows all the other concepts to function as they do.

The moral agent must possess a concept of self on a whole other level: not just an implicit concept of self as a whole, with no required distinction of mind or body, but an explicit concept of *self-as-myself*, as an intentional and distinctively cognitive entity. One cannot hold an agent morally responsible for her actions if she has no concept that *she* is (or could be) the one responsible for the actions and their consequences – and not the person over there. She must be able to hold *herself* responsible: and that she cannot do without full self-conscious awareness. She must, so to speak, be able to recognize herself in a mental mirror.

2.5 The moral agent must be a sign consumer and producer.

> … Cultural categories involve the ability to use and interpret *conventional signs*, in the semiotic sense of the word, where a relatively concrete expression (e.g., a handshake, a gesture, a word) *represents* a relatively less concrete concept (e.g., friendship) for the members of a community [55, p. 259].

Finally, an agent cannot be held morally responsible unless she is able to communicate evidence for her moral responsibility through some conventionalized channel. She need not be a linguistic agent (the peak of Zlatev's semiotic hierarchy), but she must possess – and be able to use appropriately – the *sign function*, just below language in the hierarchy. She must be able to communicate through e.g. gestures or pictures, if not language.

We intend *sign* as it is used in semiotics, but care must be taken, as *sign* is used by different people – and, sometimes (thinking here of the early versus the late Charles Sanders Peirce) the same people – in broader or narrower ways. So e.g. we intend *sign* in the present context in a narrower way than Sara Lenninger, who, echoing Ferdinand de Saussure's expression/content distinction [11], describes sign meanings as:

> . . . Both relational and functional: in particular, signs always imply relations between expression and content for an interpreting subject. This is to say that signs involve something that is "given" to someone (the *expression*); and, further, that this "given" meaning directs attention to something else: i.e., the *content* [19, p. 1].

Lenninger is a student of Göran Sonesson and follows him in taking a sign to "involve a subject differentiating between expression and content, where the relations between expression and content are asymmetric" [19, p. 17] (*cf.* [43, pp. 59, 83]). This *explicit* differentiation between expression and content marks out something narrower than our conception of concepts, which requires only an *implicit* differentiation of form from content[8]: e.g, on Sonesson's account, a mirror is (or rather, can be, in the appropriate circumstances) a sign [42]; at the same time, it is (intentionally) broader than that of conventionalized signs – Zlatev's target in [55], and ours here – which assume, in their basis, an intent to communicate between members of a community.

In [19], Lenninger examines the development of the *sign function* (her focus is on the *picture sign*) – in young children. This is significant because, in determining whether a prospective moral agent *is* a sign producer and consumer, in our terms, one will, doubtless, want to look at how its use of the sign function develops; even though, echoing what we said earlier, a nonstandard developmental process should not, of itself, rule out the possibility of moral agency.

3 MAPPING OUT THE MORAL SPACE

> . . . I argue that conceptual spaces present an excellent framework for "reifying" the invariances in our perceptions that correspond to assigning properties to perceived objects [16, p. 59]

By *perceptions*, Gärdenfors means to include not only direct sensory perceptions but such abstract entities as moral perceptions. By *objects*, he means to include not just concrete physical objects but such highly abstract objects as morals. His conceptual spaces theory, and the unified conceptual space extensions to it, will provide a useful tool to mapping out an agent's moral territory. First, however, it is necessary to say something about what the requirements on moral agency presented in Section 2 mean.

3.1 Understanding what the requirements mean

To playfully paraphrase Albert Newen and Andreas Bartels (who were, after all, writing about non-human animal concepts and not about moral agency!), an agent, in order appropriately to be attributed moral agency, must, minimally:

- Demonstrate an ability to derive general moral principles from specific moral applications, and *vice versa*;
- Demonstrate a flexible pattern of behaviour based on this ability, especially when confronted with novel situations; and
- Demonstrate surprise or frustration on having her moral expectations violated [24, p. 291] (see also [1, p. 37]).

We believe that it is not possible for an agent to achieve this unless she meets the conditions from Section 2: i.e., unless she is embedded/embodied, alive, a conceptual agent, a self-consciously aware agent, and a user of signs. Note that these properties are not intended to determine whether a given agent is or is not a moral agent; they are rather meant to set the minimal foundations from which one should start looking for moral agency. To put this more strongly: they are meant to be descriptive rather than stipulative: i.e., we are not saying that *by definition* an agent must be embedded and embodied in order to qualify as a moral agent; rather, the onus is on the researcher whose purportive moral agent is *not* embedded or embodied – or not embedded and embodied in the appropriately physical way – to explain why that agent should still be taken as a moral agent.

So, for example, one might claim that a certain *virtual* agent, existing in some computer simulation, is a moral agent. One might note, further, that virtual agents *are* embedded in a (virtual) environment and embodied in a certain (weak) sense. Here, we would simply remind the reader of Zlatev's [55] moral that a model of something – such as a conscious agent (his focus) or a moral agent (ours) – is not the same as, and should not be confused with, the thing itself (see Section 4).

Likewise, we are not stipulating that life, as we know it, and moral agency, as we know it, *cannot* come apart – only that the researcher who acknowledges that her artefact is *not* alive must justify why, nevertheless, one should ascribe that artefact moral agency and not take her language to be loosely metaphorical. So it goes for conceptual abilities (why should one hold responsible an agent who e.g. has no concept of moral right and wrong?), self-conscious awareness (how can one hold responsible an agent who has no explicit awareness of *itself* as the agent responsible for its actions?), and sign use (what does it mean to be a moral agent in the absence of any ability to communicate that agency?).

Neither are we claiming that the five conditions of Section 2 are *sufficient* – only that they are (we believe) *necessary*. In particular, they are not sufficient because they remain – for all of our attention to detail – abstract. It is one thing to set out the appropriate grounds for attributing moral agency; it is quite another thing to operationalize such highly theoretical methodology. Conceptual spaces theory and the unified conceptual space theory provide the tools to do so while deliberately avoiding any kind of moral agency litmus test.

3.2 Operationalizing the requirements

One way to go about operationalizing these requirements would be with something like Allen and colleagues' moral Turing test (MTT) [2], intended as a litmus test for moral agency: *either* the agent passes it, *or* it does not. For the kind of moral agency Allen and colleagues have in mind – which, unlike our approach, is more focused on the results of actions, with less attention to the means and methods of getting there – such an approach may have merits.

That said, one should be wary, especially given the deliberate connection to Alan Turing and his imitation game [48]. Despite the widespread tendency to turn the so-called Turing test into a operational test of "true" (and not just simulated!) intelligence, we think it quite clear, if one goes back to Turing's paper, that this is *not* what he had in mind. "The contrivance of the imitation game. . . was intended to show the importance of human attitudes, not to be an operational definition of intelligence" [51, p. 62]: to wit, the imitation game was a *thought experiment* to encourage Turing's readers to *think about* thinking and to suggest, modestly, that *how* people think about thinking can evolve. It's worth noting, too, that Turing's requirements for an artefactual agent "winning" the imitation game are very modest (albeit not yet quite met, even after 62 years!): "I believe that in about fifty years' time it will be possible to programme computers. . .

[8] . . . That is to say, it marks out a higher level of cognitive development or abilities than what we take to be the minimum requirements of conceptual agency.

to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning" [48, p. 442]. His claim – two sentences later – that "...at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" has, indeed, been proven correct.

In contrast to Allen and colleagues, our own usage of *moral agency* is to refer specifically to the conditions under which it is appropriate to hold a given agent morally responsible: in which case, a litmus test for moral agency would be, we believe, grossly unethical (as, for that matter, would a litmus test for self-conscious intelligence) – allowing far too many unwarranted conclusions and consequences from a simple failure to pass the test. Note that no such *general* tests for moral agency (or self-conscious intelligence) exist for humans – or, so far as we are aware, have ever seriously been proposed.

Instead, in [28], we propose a way – not of ruling putative moral agents "in" or "out" – but rather mapping and exploring the putative moral agent's *moral space*, using the theoretical tools of *conceptual spaces theory* [16] and the theoretical and software-based tools of the *unified conceptual space theory* [25]. Conceptual spaces theory is a *similarity-space*-based theory of concepts in the tradition of Eleanor Rosch's work on prototypes [36, 35]; the language of geometry: individual conceptual spaces are realized as *Voronoi tessellations*; and the spirit of philosophical empiricism. Any given conceptual space is defined by its *integral dimensions*: no one dimension can be specified without a logical mandate to specify the others; and a predetermined metric for those dimensions. Gärdenfors' example of choice is the colour spindle, with its integral dimensions of hue, saturation, and brightness (see Figure 1).

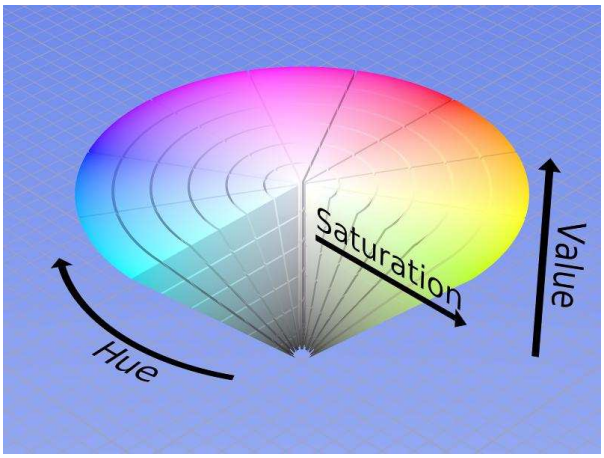

Figure 1. The colour spindle. (Photo downloaded from Wikimedia Commons: http://commons.wikimedia.org/.)

The unified conceptual space theory, first introduced in [27], is a set of extensions to conceptual spaces theory attempting to push it in a more algorithmically amenable and empirically testable direction, while explaining how all of a conceptual agent's many different conceptual spaces – of TONE (another of Gärdenfors' choice examples, and an instance, like COLOUR, of a *property concept*), of TREES and DEMOCRACY (instances of concrete and abstract *object concepts*, respectively), of TO THROW and TO BELIEVE (instances of concrete and abstract *action concepts*, respectively), and so on – come together in a unified *space of spaces*: a divergent space defined by the integral dimensions or *axes* common to ***all*** concepts. These are the axes of *generalization* (arranged from maximally general to maximally specific), *abstraction* (arranged from maximally concrete and "lower order" to maximally abstract and "higher order"), and *alternatives* (see Figure 2). The axis of alternatives is determined by selecting one or more integral dimensions and then incrementing or decrementing their value[9]. Meanwhile, one can toggle between two, contrasting views on concepts: as *things* (relatively stable and even fixed: prototypically noun-like) or *processes* (relatively dynamic and even in a constant state of incremental change: prototypically verb-like).

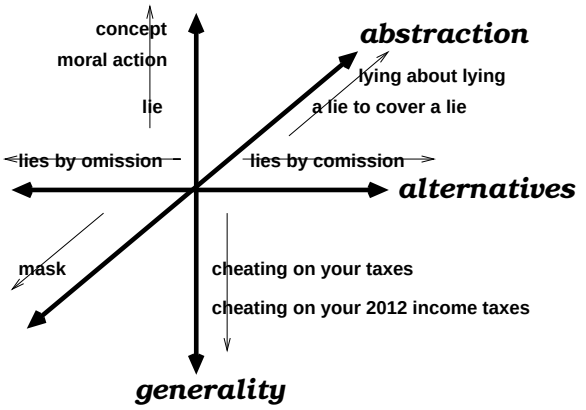

Figure 2. The three dimensions of the unified space, and an illustration of the concept of LIE.

The unified conceptual space theory comes with a software implementation: a fairly direct translation of the theory into concrete visualization in the form of a mind-mapping program, a tool for helping users externalize and visualize their understanding of some conceptual sub-domain on the justification that what is externalized is thereby made explicit; what is made explicit is easier to scrutinize; what is easier to scrutinize is *ceteris paribus* easier to modify. The visual interface is strikingly different from existing mind-mapping programs and, in sharp contrast to their reliance on a vaguely defined theory of cognition[10], implements a specific theory of concepts – which it can, we believe, be used to test empirically.

We will not recapitulate here the discussion from [28] of how to use the software implementation to model and explore an agent's moral space. Instead, it will be useful to describe one concept, LIE, as it might be modeled in the unified conceptual space theory, using the software implementation.

LIE can be viewed *either* as an abstract object *or* as an abstract action (see Figure 2). LIE is a type of MORAL ACTION which is, ultimately, a type of CONCEPT. Going the other direction along the

[9] Since, for most concepts, one can choose different subsets of integral dimensions, this means that the *axis of alternatives* diverges into different "parallel" spaces.

[10] The theoretical underpinnings of the commercial and freely available software are essentially identical. They differ in cosmetic aspects of user interface.

axis of generality, more specific types of lie include CHEATING ON YOUR TAXES or, even more specifically, CHEATING ON YOUR 2012 INCOME TAXES.

Going one direction along the axis of abstraction, a higher-order LIE would be LIE TO COVER A LIE or even LIE ABOUT A LIE. Going the other direction, a more concrete / less abstract version of something like LIE would be a MASK that presents a "false face".

Finally, LIE/LYING has certain integral dimensions, including choice of words and the degree to which salient facts are left out or false information is included. Varying these latter two dimensions produces what Sisela Bok has termed *lies of omission* and *lies of comission* [7].

Movement along any of these three axes represent relations of contiguity. There is a further and contrasting way of specifying concepts: in terms of distal (non-contiguous) relations, within the unified space, to a given concept's *components* (based on part/whole relations), *parameters* (i.e., integral dimensions), and *contextuals* (commonly associated but non-integral dimensions). Unfortunately, it is not possible to go into the details here. Meanwhile, a screenshot of how lie might look in the software implementation is shown in Figure 3.

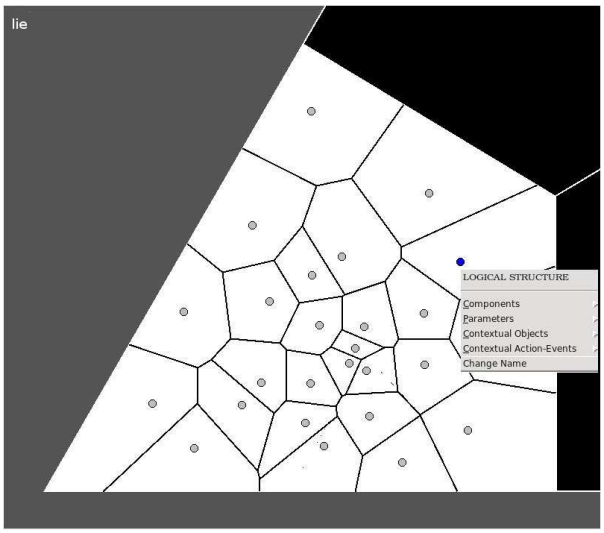

Figure 3. The concept lie as it might be "drawn" within a larger moral space, using the software implementation.

4 INFLATED CLAIMS OF "MINIMAL" CONSCIOUSNESS

> This value system must be *intrinsic* to the system in the sense that it serves to preserve the system's *organization*, rather than criteria which are external to the system (defined by the system's designers) [55, p. 282].

We agree fully and enthusiastically with Zlatev's claim that, to talk of an agent engaging with meaning – including, for our purposes, moral meaning – that meaning *must* be meaning for the agent herself, determined by the agent, and not by any external authority; and that the ultimate grounding of meaning lies in organizational (self-)preservation. Remember his warning that one should not confuse or conflate the model with the thing being modeled.

Given that we have clearly set out consciousness (indeed, self-consciousness) as a prerequisite for moral agency, it will be enlightening to have a look at the current state of claims in the machine consciousness community. If one were to go by oral claims – a recent conference on machine consciousness comes to mind – one might be tempted to think that a majority of researchers thought their creations to have achieved at least a state of "minimal" consciousness. Unsurprisingly, the same researchers are often far more reticent in writing. Nonetheless, they may be inclined toward either one of two moves that concern us here.

First, they may be inclined to reduce consciousness to *x*, where *x* might be any of Daniel Dennett's *multiple drafts* [12], Guilio Tononi's *information integration* [46, 6], Bernard Baar's *global workspace* [4, 3], or LIDA (a specific implementation of global workspace theory) [5, 33] – to name just a few popular choices. Tononi's information-integration measure ϕ is particularly popular because it offers a precise numerical value of an agent's state of consciousness. As Torrance (among others) notes, "Tononi is explicit that artificial systems can have ϕ measures in the way as biological organisms can" [47, p. 70]; and so existing artefacts can be ranked according to their ϕ value.

Second, there is a seeming tendency to be ambivalent whether terms like *consciousness*, *autonomy*, *self*, *self concept*, and so on are intended more literally or more metaphorically, more strongly or more weakly: more as reality or more as "model of". Note that none of the machine consciousness researchers we are aware of claim that their creations are, in any substantive sense, alive – something we have set as a precondition for consciousness (and so for moral agency).

The stated intentions of many machine consciousness researchers is, clearly, to implement consciousness in a strong sense. So, for example, Klaus Raizer and colleagues, drawing on Baar's global workspace theory, write, "our main goal is to develop artificial creatures with different levels of machine consciousness" [32, p. 43]; while Ricardo Sanz and colleagues write "Sloman… recognises that *"perhaps one day… it will prove possible to model the totality* [of consciousness] *in a single working system"* …this is, boldly, what we try to do inside our long term ASys research program" [38, p. 55].

Some machine consciousness researchers really *do* seem to believe that their creations have achieved some level of (actual) consciousness. We in no way mean to be pointing fingers, despite our resistance to such claims – only offering an illustrative example of researchers willing to stick their necks out. So for example, Uma Ramamurthy and Stan Franklin write that "the LIDA model is both a conceptual and a computational model implementing and fleshing out a major portion of Global Workspace Theory"; "…core consciousness is continually regenerated in a series of pulses (LIDA's cognitive cycles…), which blend together to give rise to a continuous stream of consciousness" [34, p. 51]. They continue:

> LIDA's processing can be viewed as consisting of a continual iteration of Cognitive Cycles…. Each cycle constitutes units of understanding, attending and acting. During each cognitive cycle a LIDA-based agent first makes sense of its current situation as best as it can by updating its representation of its world, both external and internal. By a competitive process, as specified by the Global Workspace Theory, it then decides what portion of the represented situation is most in need of attention. Broadcasting this portion, the current contents of consciousness, enables the agent to finally choose an appropriate action which it then executes. Thus, the LIDA cognitive cycle can be

subdivided into three phases, *the understanding phase, the consciousness phase,* and *the action selection phase [34, p. 52].*

LIDA may not be self-consciously aware, but it is clearly meant to be close. "Action that is taken volitionally, that is, as the result of conscious deliberation, is an instance of the action by the Volitional Self. Deliberate actions occur in LIDA and are represented as behavior streams. Thus LIDA has a volitional self" [34, p. 52]. "An autonomous agent/cognitive robot based on the LIDA model that also has a self system might be suspected of being close to subjectively conscious…" [34, p. 53].

Is LIDA – or a LIDA-based system – conscious or even self-conscious? We would ask: is it suitably embedded and embodied? Is there reason to think it is alive and not just "alive"? Is it the appropriate sort of conceptual agent? For the moment, the answers to these questions remain – to us, at least – unclear.

We have set a very high bar for attributing *any degree* of moral agency. We have concluded that no existing artefacts or simulations of artefacts that we are aware of meet that standard. However, this is not to say that there are not existing artefacts – some of them commonplace – which raise very real issues regarding moral responsibility. We discuss two such groups of artefacts: those we believe people *should* take a moral stance toward – by which we mean to treat, in some limited way, *as if* they were moral agents – even though one should logically conclude they are not; and those we believe people *should not* take a moral stance toward, despite what we see as a strong (perhaps, to some extent, irresistible) urge to do so.

5 TAKING THE MORAL STANCE

> Research shows that humans can sometimes become very abusive towards computers and robots particularly when they are seen as human-like.… Is it acceptable to treat artefacts – particularly human-like artefacts – in ways that we would consider it morally unacceptable to treat humans? Second, if so, just how much sexual or violent "abuse" of an artificial agent should we allow before we censure the behaviour of the abuser? [52, p. 326]

Different artefacts impose different levels of moral responsibility toward them. Adopting the moral stance with respect to a particular artifact is a complex moral judgment. This judgment cannot simply be determined by any claims to consciousness. The moral stance is a moral stance and not a scientific stance. For this reason, we hold that there are groups of artefacts that one morally ought to adopt the moral stance towards and groups to which one ought not to adopt the moral stance –- in spite of sometimes strong emotional responses to the contrary.

We claim, for example, that it is immoral to kick one's AIBO[11] in ways that abusing one's laptop is not: an issue explored at some length in [52] (see also the opening quote to this section). That is because of the – albeit highly superficial – resemblance to agents one normally does, and indeed should, take a moral stance toward. Kicking one's AIBO derives from the immorality of kicking a living dog precisely because of the (we believe inevitable) way it reminds people of "real" dogs. This is regardless of whether such behaviour then predisposes people to abuse dogs in ways or to a degree they would not have done so before: i.e., the act itself is intrinsically immoral, and any behavioural progression as from abusing AIBOs to abusing "real" dogs is irrelevant to our argument. Although people do, indeed, "anthropomorphize" their laptops in various ways ("my laptop ate my file!"), they do so to a far more limited extent, in far more constrained ways, precisely because of the lack of resemblance to living agents.

Our moral stance supervenes on the moral resemblance of the AIBO to a "real" dog. Let us clarify *moral resemblance* in this context. An artefact may resemble a morally significant natural item along some combination of at least three distinct dimensions: physical appearance, behaviour, and the role in which it usually employed. All three dimensions are important to the moral stance.

Physical resemblance is, in many ways, the morally least significant dimension of resemblance. The AIBO has only the most superficial physical resemblance to a dog. What makes the AIBO dog-like is not its visual appearance. The AIBO resembles a dog more in its (admittedly superficial) dog-like behaviours; and, most importantly, in the dog-like roles it is marketed to fill –- particularly in its relationship to its owner. When someone kicks an AIBO, she is kicking an artefact with which she or someone else might reasonably be expected to have a relationship that resembles the relationship that humans have with dogs. This is the main basis of our moral condemnation of the act.

There is overwhelming empirical evidence for automatic human emotional responses along the dimensions outlined above. Even in the case where there is no physical resemblance at all, behavioural resemblance is more than enough: witness the heartfelt conversations people had with Eliza in the '70s and the anger they expressed when those "private" conversations were interrupted [50]. Similarly, the work of Cynthia Brezeal and others on the Kismet robot shows that, even when people know that certain tricks are being played to make them respond in a friendly way, they cannot help but respond emotionally [8]. There is good reason to expect that humans will, in many situations, respond to artefacts as if they were human or animal, even without any obvious physical resemblance. This has already been observed happening in the – seemingly unlikely – case of soldiers bonding with military robots in combat situations [17].

6 ASSIGNING RESPONSIBILITY

> Whether the machines of capable of learning while they are acting has little impact on the consequences for human society if we allow each other to displace our responsibility onto our creations [9, p. 18].

Another class of artefacts raises a different and at least equally worrying set of moral issues: so-called "autonomous" robots, particularly as used by the major military powers on the battlefield. The use of "autonomous" in the battlefield context is at best unclear; at worst, it may be duplicitous. When philosophers write of autonomy, they often implicitly assume that it includes something like our requirements for moral agency: in particular, that the moral agent must be a sophisticated conceptual agent (Section 2.3) and that it must be self-aware (Section 2.4). By contrast, when robot builders talk of autonomy, they may mean nothing more than that there is no (or very little) direct human control at that moment. These are quite different usages.

In any case, what matters here is not that the robots have any true autonomy – in the sense of autonomy used by e.g. Varela and Maturana, they do not – but that they give the *appearances* of autonomy, generating automatic emotional responses not so different from people's automatic responses to Kismet. These robots do not resemble

[11] "AIBO" is the trademarked name of a robotic dog manufactured by Sony from 1999 until 2006. It stands for **A**rtificial **I**ntelligence r**OB**ot.

humans or animals in the way that ASIMO or an AIBO does, and so abusing them is, we believe, roughly morally equivalent to abusing a laptop. The problem is the appearances of autonomy, and the way they muddy the waters when it comes to assigning responsibility. This is because of the difficulty of tracing a clear or direct line from the actions of the morally responsible agent(s) to the actions of the robot and the tendency to act *as if* the robot, itself, is morally responsible.

Meanwhile, the language of autonomy has become widespread in military robotics and signifies or reflects mainly a reduction in direct human control. Current trends are towards reduced human participation: in the jargon, from "man in the loop" to "man on the loop" [40, 41]. There are many motivations for this. Automated weapon systems can often be cheaper and (by some measure) more reliable than human warriors, even though it turns out that (for reasons not yet fully understood) the post of *unmanned aerial vehicle* (UAV) operator or "drone pilot" is highly stressful and implicated in PTSD among service people. That said, our immediate concern is how the allocation of moral responsibility is affected.

The choice of language, and, more importantly, the appearances of autonomy, invite taking what we have termed the *moral stance* toward these battlefield robots – e.g., UAVs such as the Predator and Reaper drones currently widely deployed by US and its allies – treating them as if they were (at least in some limited sense) suitable candidates for moral agency and shoving some of the responsibility for "their" actions off onto them [40]. There is a wide range of unmanned land vehicles (ULVs) in use as well. At least forty countries are engaged in developing robots for military purposes. A lack of care in media reporting contributes to the problem. The wording "killed by a drone" as opposed to "killed by a US soldier using a drone" may be journalistically preferable but is misleading.

If an "autonomous" battlefield robot malfunctions and kills a dozen civilians, who is responsible? Clearly – appearances to the contrary – the robot is not. The tendency to act as if the robot, itself, is morally responsible is seductive but dangerously wrong. Present and immediately foreseeable robots simply do not have the sort of decision-making machinery needed to make the moral stance appropriate. Indeed the whole debate has (perhaps conveniently) obscured the chain of moral responsibility. The second author on this paper is presently tasked with preparing a policy document for the British Computing Society, Chartered Institute for IT, condemning the practice of using computers as an excuse for human failings [53].

Here, our solution is to resist, as strongly and explicitly as possible, the tendency to assign any moral agency whatsoever to the "autonomous" robot and, meanwhile, be as explicit as possible (ideally in advance!) about the appropriate distribution of responsibility between robot operator, manufacturer, those giving the orders on the battlefield, those giving the orders higher up the chain of command, and other implicated parties. One must consciously acknowledge and consciously oppose any impulse to assign even partial responsibility to the robot or leave the distribution of responsibility unclear. In any case, although – as should be clear by now – we consider an artificial moral agent to be conceivable and, *prima facie*, theoretically possible, the place to pursue actual artificial moral agency is not on the battlefield, where the consequences of an irresponsible mistake are much too high.

7 CONCLUSIONS

Moral agency requires much more of an agent than might, on first blush, appear. Being in a position to be held accountable for one's actions and their consequences requires far more than "doing the right thing". Drawing on Zlatev's semiotic hierarchy, we conclude that a moral agent must be, successively, embedded/embodied, alive, a sophisticated conceptual agent, self-consciously aware, and a sign consumer/producer. Rejecting any litmus test for moral agency, such as the Moral Turing Test [2], we present our requirements as descriptive rather than stipulative, and describe how theoretical and software-based tools from conceptual spaces theory and the unified conceptual space theory can be used to operationalize those requirements. We discuss, and reject, claims that existing artefactual systems are, in any useful sense, either conscious or self-conscious. We then look at two groups of existing artefacts: one of which one should, we believe, take a *moral stance* toward, even though they are not moral agents: i.e., artefacts that visually or otherwise physically resemble agents that are moral agents or that one would otherwise take a moral stance toward; the other of which invites treatment as morally responsible agents even while we think that any tendency to assign them degrees of moral responsibility should be resisted in the strongest possible terms.

REFERENCES

[1] C. Allen, 'Animal concepts revisited: The use of self-monitoring as an empirical approach', *Erkenntnis*, **51**(1), 33–40, (1999).

[2] C. Allen, G. Varner, and J. Zinser, 'Prolegomena to any future artificial moral agent', *Journal of Experimental & Theoretical Artificial Intelligence*, **12**(3), 251–261, (2000).

[3] B. J. Baars, *A Cognitive Theory of Consciousness*, Cambridge University Press, Cambridge, UK, 1988.

[4] B. J. Baars, *In the Theater of Consciousness: The Workspace of the Mind*, Oxford University Press, New York, 1996.

[5] B. J. Baars and S. Franklin, 'Consciousness is computational: The lida model of global workspace theory', *International Journal of Machine Consciousness*, **1**(1), 23–32, (2009).

[6] D. Balduzzi and G. Tononi, 'Integrated information in discrete dynamical systems: Motivation and theoretical framework', *PLoS Computational Biology*, **4**(6), 1–18, (2008). http://dx.doi.org/10.1371

[7] S. Bok, *Lying: Moral Choice in Public and Private Life*, Vintage, 1999. First published 1978.

[8] C. Breazeal and B. Scassellati, 'How to build robots that make friends and influence people', in *Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pp. 858–903, (1999).

[9] J. J. Bryson, 'A role for consciousness in action selection', in *Machine Consciousness 2011: Self, Integration and Explanation (Proceedings of a Symposium at the AISB'11 Convention 4-7 April 2011, York, United Kingdom)*, eds., R. Clowes, S. Torrance, and R. Chrisley, pp. 15–19. The Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), AISB, (2011).

[10] A. Damasio, *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*, Vintage, 2000.

[11] F. de Saussure, *Cours de Linguistique Generale, vol. 3*, Payot, Paris, 1968.

[12] D. C. Dennett, *Consciousness Explained*, Little, Brown, 1991.

[13] E. A. Di Paolo, 'Autopoiesis, adaptivity, teleology, agency', *Phenomenology and the Cognitive Sciences*, **4**, 429–452, (2005).

[14] G. Evans, *Varieties of Reference*, Clarendon Press, 1982. Edited by John McDowell.

[15] J. A. Fodor, *Concepts: Where Cognitive Science Went Wrong*, Clarendon Press, Oxford, 1998.

[16] P. Gärdenfors, *Conceptual Spaces: The Geometry of Thought*, Bradford Books, 2004. First published 2000.

[17] J. Garreau. Bots on the ground: In the field of battle (or even above it), robots are a soldier's best friend. Washington Post, May 2007. http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html.

[18] H. D. Jaegher, E. D. Paolo, and S. Gallagher, 'Can social interaction constitute social cognition?', *Trends in Cognitive Science*, (2010). In press.

[19] S. Lenninger, *When Does Similarity Qualify as a Sign?: A study in Picture Understanding and Semiotic Development in Young Children*, Ph.D. dissertation, Lunds Universitet, Lund, Sweden, 2012. forthcoming.

[20] H. Maturana, 'Cognition', in *Wahrnehmung und Kommunikation*, eds., P. M. Hejl, W. K. Köck, and G. Roth, 29–49, Peter Lang, Frankfurt, (1978). Available online at http://www.enolagaia.com/M78bCog.html, with the original page numbering retained.

[21] H. R. Maturana and F. J. Varela, *Autopoiesis and Cognition: The Realization of the Living*, Springer, 1980.

[22] H. R. Maturana and F. J. Varela, *The Tree of Knowledge: The Biological Roots of Human Understanding*, Shambhala, London, 1992.

[23] U. Neisser, 'Five kinds of self-knowledge', *Philosophical Psychology*, **1**(1), 35–59, (1988).

[24] A. Newen and A. Bartels, 'Animal minds and the possession of concepts', *Philosophical Psychology*, **20**(3), 283–308, (2007).

[25] J. Parthemore, *Concepts Enacted: Confronting the Obstacles and Paradoxes Inherent in Pursuing a Scientific Understanding of the Building Blocks of Human Thought*, Ph.D. dissertation, University of Sussex, Falmer, Brighton, UK, March 2011.

[26] J. Parthemore, 'Of boundaries and metaphysical starting points: Why the extended mind cannot be so lightly dismissed', *Teorema*, **30**(2), 79–94, (2011).

[27] J. Parthemore and A. F. Morse, 'Representations reclaimed: Accounting for the co-emergence of concepts and experience', *Pragmatics & Cognition*, **18**(2), 273–312, (August 2010).

[28] J. Parthemore and B. Whitby, 'When is any agent a moral agent?: Reflections on machine consciousness and moral agency', *International Journal of Machine Consciousness*, **4**(2), (December 2012). in press.

[29] J. Prinz, *Furnishing the Mind: Concepts and Their Perceptual Basis*, MIT Press, 2004. First published 2002.

[30] H. Putnam, 'The meaning of 'meaning", in *Language, Mind, and Knowledge*, ed., K. Gunderson, University of Minnesota Press, (1975).

[31] H. Putnam, *Reason, Truth, and History*, Cambridge University Press, Cambridge, UK, 1981.

[32] K. Raizer, A. L. Paraense, and R. R. Gudwin, 'A cognitive neuroscience-inspired codelet-based cognitive architecture for the control of artificial creatures with incremental levels of machine consciousness', in *Machine Consciousness 2011: Self, Integration and Explanation (Proceedings of a Symposium at the AISB'11 Convention 4-7 April 2011, York, United Kingdom)*, eds., R. Clowes, S. Torrance, and R. Chrisley, pp. 43–50. The Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), AISB, (2011).

[33] U. Ramamurthy, B. J. Baars, S. D'Mello, and S. Franklin, 'Lida: A working model of cognition', in *Prooceedings of the Seventh International Conference on Cognitive Modeling*, pp. 244–249, (2006).

[34] U. Ramamurthy and S. Franklin, 'Self system in a model of cognition', in *Machine Consciousness 2011: Self, Integration and Explanation (Proceedings of a Symposium at the AISB'11 Convention 4-7 April 2011, York, United Kingdom)*, eds., R. Clowes, S. Torrance, and R. Chrisley. The Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), AISB, (2011).

[35] E. Rosch, 'Family resemblances: Studies in the internal structure of categories', *Cognitive Psychology*, **7**, 573–605, (1975).

[36] E. Rosch, 'Principles of categorization', in *Concepts: Core Readings*, eds., E. Margolis and S. Laurence, chapter 8, 189–206, MIT Press, (1999).

[37] G. Ryle, *The Concept of Mind*, Penguin, 1949.

[38] R. Sanz, C. Hernandez, and G. Sanchez, 'Consciousness, meaning, and the future phenomenology', in *Machine Consciousness 2011: Self, Integration and Explanation (Proceedings of a Symposium at the AISB'11 Convention 4-7 April 2011, York, United Kingdom)*, eds., R. Clowes, S. Torrance, and R. Chrisley. The Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), AISB, (2011).

[39] J. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**(3), 417–458, (1980).

[40] N. Sharkey, 'Automating warfare: Lessons learned from the drones', *Journal of Law, Information & Science*, EAP1 – EAP15, (August 201). E-publication ahead of print.

[41] N. Sharkey, *Killing Made Easy: From Joysticks to Politics*, chapter 7, 111–128, MIT Press, December 2011.

[42] G. Sonesson, 'Iconicity strikes back: The third generation – or why Eco still is wrong', in *La Semiotique Visuelle: Nouveaux Paradigmes*, ed., M. Costantini, 247–270, L'Harmattan, Paris, (2010).

[43] G. Sonesson. Lecture 2: The psychology and archaeology of semiosis. Semiotics Institute Online: http://projects.chass.utoronto.ca/semiotics/cyber/Sonesson2.pdf, October 2010.

[44] P. Steiner and J. Stewart, 'From autonomy to heteronomy (and back): The enaction of social life', *Phenomenology and the Cognitive Sciences*, **8**, 527–550, (2009).

[45] E. Thompson, *Mind in Life: Biology, Phenomenology and the Sciences of Mind*, Harvard University Press, 2007.

[46] G. Tononi, 'Consciousness as integrated information: A provisional manifesto', *The Biological Bulletin*, **215**(3), 216–242, (December 2008).

[47] S. Torrance, 'Would a super-intelligent AI necessarily be (super-) conscious?', in *Machine Consciousness 2011: Self, Integration and Explanation (Proceedings of a Symposium at the AISB'11 Convention 4-7 April 2011, York, United Kingdom)*, eds., R. Clowes, S. Torrance, and R. Chrisley, pp. 67–74. The Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), AISB, (2011).

[48] A. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433–460, (October 1950).

[49] W. Wallach, C. Allen, and S. Franklin, 'Consciousness and ethics: Artificially conscious moral agents', *International Journal of Machine Consciousness*, **3**(1), 177–192, (2011).

[50] J. Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, Penguin Books, 1993. First published 1976.

[51] B. Whitby, 'The turing test: AI's biggest blind alley?', in *Machines and Thought: The Legacy of Alan Turing*, eds., P. Millican and A. Clark, volume 1, Clarendon, (1996).

[52] B. Whitby, 'Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents', *Interacting with Computers*, **20**(3), 326–333, (2008). Special issue on the abuse and misuse of social agents, S. Brahnam and A. De Angeli (eds.).

[53] B. Whitby, 'Draft policy document on the use of it as an excuse', Technical report, BCS, the Chartered Institute for IT, (2012). In preparation.

[54] J. Zlatev, 'The epigenesis of meaning in human beings, and possibly in robots', *Minds and Machines*, **11**, 155–195, (2001).

[55] J. Zlatev, 'Meaning = life (+ culture)', *Evolution of Communication*, **4**(2), 253–296, (January 2002).

[56] J. Zlatev, 'The semiotic hierarchy: Life, consciousness, signs and language', *Cognitive Semiotics*, **2009**(4), 169–200, (August 2009).

Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines

John Basl

Abstract. The creation of artificial consciousnesses raises a variety of philosophical challenges. From an ethical perspective, the creation of artificial consciousness raises serious issues as we develop consciousnesses that have the capacity for attitudes. However, current machines lack the capacities that would impose any moral restrictions on us. Until machines are developed that have a certain kind of consciousnesses, machines should not be considered moral patients.

1 INTRODUCTION

On any reasonable account of who or what is a moral patient – i.e., who or what has a welfare that we must consider in our moral deliberations – once we achieve artificial consciousness on par with our own – a consciousness with capacities like our own that experiences the world much like we do – we must recognize that consciousness as a moral patient; it will be due consideration for its own sake in virtue of the fact that it has interests similar to ours and because there will be little or no reason to discount those interests. [1] Indeed, it seems plausible that insofar as artificial consciousness approximates our own mental life, it will be due equal consideration whether that is understood in consequentialist, deontological, or other ways.[2]

However, we are a long way from creating an artificial consciousness that is anything like our own, or, for that matter, perhaps from creating artificial consciousness unlike our own. Yet, as we create more and more complicated machines and attempt to create artificial consciousness, we must think carefully about which properties of a machine would confer interests or moral patiency. To fail to ask whether these intermediate entities have interests and whether they are due consideration may lead to inappropriate conduct on our part. After all, it is not only beings with a consciousness like ours that are moral patients; non-human animals are moral patients, and we owe it to them to take their interests into account in our moral deliberations. This is so even though their mental life may be radically different than our own. This forces us to determine (i) under what conditions machines have a welfare and (ii) whether and how to take that welfare into account.

As difficult as it is to build an artificial consciousness, we also face great difficulties in answering the philosophical questions raised by the creation of such an entity. Above I raised both a metaphysical question (what properties of machines constitute their having a welfare or interests) and a normative question (how are we to take the welfare of artificial consciousness into account). There are also serious epistemic questions we must answer. How are we to know when a machine is conscious? What are the sources of evidence for consciousness? In what follows, I take up these challenges with a primary focus on the metaphysical and epistemological questions. In exploring these challenges, I hope to argue that current machines are not moral patients, or, more specifically, they are not moral patients for any practical purposes. In our ethical deliberations, we need not take the interests of (current) machines into account.

In Section 2, I explain the concept of moral status and its relationship to the concept of welfare and interests. In Section 3, I take up the metaphysical, epistemic, and normative issues concerning the moral status of artificial consciousness. I argue that in order for an entity to have what I will call *conscious interests* it must have the capacity for attitudes.[3] Entities might be conscious even if they lack this capacity, but they will not have any interests in virtue of that consciousness. In Section 4, I argue that current machines are *mere machines*, machines that lack the capacity for attitudes. Despite the fact that current machines may have all sorts of capacities, we have no evidence whatsoever that they are conscious, let alone that they have attitudes regarding their conscious states. Even those skeptical of this conclusion, I argue, will have to agree that we have no evidence whatsoever what their attitudes are. Given this, for all practical purposes, any interests that these machines have are irrelevant to our moral deliberations. To deny this is to violate the principle that any obligations we have cannot outstrip what is possible given our limited capacities.

I do not wish to argue that an entity has a welfare only in virtue of having conscious interests. Many environmental ethicists have argued that non-sentient organisms have interests in virtue of being goal-directed, or teleologically

[1] I talk in terms of consciousness rather than intelligence to avoid taking a stand on the relationship between the two. I assume instead that it is possible for a machine to be intelligent without it being conscious.

[2] What exactly that means for the treatment of such beings will be a function of their nature and our relationships to them, just as equal consideration of humans is sensitive to these factors.

[3] It seems plausible that both the capacity for desires and the capacity for preferences requires or is partly constituted by the capacity for attitudes. If this is false, then attitudes can be understood as "attitudes, preferences, or desires" throughout this paper.

organized systems. In Section 5 I explore the possibility that machines have what I'll call *teleo interests* in virtue of their being teleologically organized. In Section 6, I will argue that even if mere machines have a welfare in virtue of having such interests, these interests are also practically irrelevant to our moral deliberations. Therefore, for all intents and purposes (current) machines are not moral patients, or, at least, they are moral patients that we need not care about.

2 MORAL STATUS, INTERESTS, AND WELFARE

Before turning to questions concerning the moral status of artificial consciousnesses, it is important to clarify how we are to understand terms such as 'moral status', 'moral patient', 'interest', and 'welfare.' These terms have intuitive meanings and have a variety of technical meanings in the literature. In what follows, I will define the terms as they will be used below.

To say that a being has moral status is to say that it is worthy of our consideration in moral deliberations. Given the breadth of this definition, it is obvious that the domain of things with moral status is quite large. For example, we often have to take artifacts into account in our moral deliberations because they are, for example, owned by others or are instrumental to our doing what's morally required. For this reason, it is important to distinguish the different reasons why a thing might have moral status.

One type of moral status is often referred to as *moral considerability* [1], [2] or as a being's having *inherent worth* [3], [4]. To be morally considerable is to have a welfare composed of interests that are to be taken into account in moral deliberations for the sake of the individual whose welfare it is[4]. For example, any typical human being is morally considerable; we have interests in living a good life, in not being subject to violence, etc., and these interests are relevant to moral deliberations. When, for example, I want to acquire money, your interest in not being robbed should figure into my decisions about how I ought acquire money, and not just because you might put up a fight, but because of the impact on your welfare.[5] In what follows, I will use the term moral patient to refer to any individual that is morally considerable.

Finally it is important to be clear about the concept of an *interest* and the relationship between interests and an entity's *welfare*. As I will use the term an individual's interests are those things the satisfaction of which contributes to its welfare. As I discuss below, there are various kinds or types of interests than an entity may have. Having an interest of any kind is sufficient for an entity's having a welfare.

The relationship between welfare and interests is important because there is a long-standing debate about the nature of welfare. On some views, having a welfare requires that an entity be conscious [5–7]. If this is so, mere machines, machines that are not conscious, cannot be moral patients. However, on other views, Objective-List Views, consciousness is not always a necessary condition for having a welfare [8], [9]. There is not sufficient space available to adjudicate between these views. Instead, I will assume that an Objective-List View is true and explain how a mere machine might have interests on such a view.[6]

We can distinguish *psychological interests* from *teleo interests*. A psychological interest is an interest that an entity has in virtue of certain psychological capacities (and psychological states). A teleo interest is an interest an entity has in virtue of being teleologically organized. In Section IV I explicate the notion of teleo interests and explain how it is that mere machines have such interests. In the following section I turn to the question of whether machines with consciousness are moral patients and in virtue of which psychological capacities they are so.

3 MORAL PATIENCY AND ARTIFICAL CONSCIOUSNESSES

3.1 The easy case: human-like consciousnesses

Imagine that an artificial consciousness has been developed or come online. This consciousness is very much like ours. It has pleasant and painful experiences, it enjoys or suffers from certain experiences, it has the capacity for imagination, memory, critical thinking, and moral agency. We can even imagine that this consciousness is embodied and goes about the world like we do. I think it goes without saying that this artificial consciousness is a moral patient in the same way that we are. On any reasonable normative theory, theory of welfare, and theory of moral considerability, this being will be our moral equal.

This is because reasonable theories of moral patiency are *capacity-based*; being a moral patient is a function of the capacities an entity has, not the type of being that it is. Furthermore, if two beings' are equal in their capacities they are or should be considered equal with respect to moral considerability or their claim to being a moral patient.[7] If tomorrow we cognitively enhanced a chimpanzee so that it was our equal in cognitive capacity, even the most adamant proponent of animal experimentation would have to recognize that this chimpanzee deserved protections equal to those afforded other human beings. It is difficult to see how being a member of a particular species or other kind is a morally

[4] We need not take all the interests of all who are morally considerable into account at all times. If a being is morally considerable then we ought to take its interests into account in contexts where we suspect there will be an appreciable impact on that being's welfare.

[5] I remain neutral, as much as possible, on how interests are to be taken into account and weighed against one another. For example, what constitutes equal consideration of interests and what one's various interests entitle one to will differ on Kantian or Consequentialist views.

[6] On some particular Objective List Views having consciousness will be a necessary condition for having a welfare. On such views, access to the objective goods is only possible for conscious beings. Even on such views, an individual's welfare will not depend soley on his or her particular mental states.

[7] This does not mean that there are no cases where we should favor one life over another. For example, if we must decide between saving the life of our own child and a stranger, we have good reason to save our child. However, this isn't because our child is a more important moral patient. It has to do with the consequences, values, and relationships at stake.

relevant feature of a being. After all, we have no trouble thinking of alien species that are otherwise like us as moral patients in the same way that we are, and yet they are not of the same species as us.

For this reason, it seems entirely plausible that once there are artificial consciousnesses with capacities very much like ours, they will be moral patients, and these patients will be our moral equals. It will matter none at all whether these beings are silicone, steel, or purely digital. Our being made of squishy, biological material will not give us moral priority over such beings.

3.2 The harder(er) case: animal- and other consciousnesses

Questions surrounding the moral patiency of artificial consciousnesses would be very easily answered if we had reason to expect that all such consciousnesses would be very much like us. Unfortunately, there is a much higher probability that in our quest to create artificial consciousness, we will develop consciousnesses that are psychologically more like non-human animals or, that are, psychologically, radically different than anything that we know of (or perhaps even in existence). Given this fact, more must be said about (i) those capacities in particular that give rise to psychological interests and (ii) how psychological interests of different kinds and strength are to be taken into account in moral deliberations.

Since this paper is concerned with whether and under what conditions machines are moral patients, I will set (ii) aside. If it turns out that machines are moral patients, we will need to determine how their psychological interests are to be taken into account in moral deliberations just as we must determine how to weigh non-human animals interests against our own given that many such animals are moral patients. As is the case with deliberations concerning non-human animals, the appropriate response to these patients will depend on which normative theory is correct as well as other factors, issues that I cannot hope to settle here.

To determine which conscious machines are moral patients at all, independently of how we are to take them into account, we must first determine which capacities in particular give rise to psychological interests of the sort that are morally relevant. Not all capacities will give rise to morally relevant interests. If we create a consciousness with only the capacity for experiencing colors but with no attending emotional or other cognitive response, we need not worry about wronging said consciousness. It might be a shame to force such a consciousness offline since its creation would no doubt be a fascinating and important achievement, but we would not wrong the machine by forcing it offline, just as we do not wrong a person (that has consented and is otherwise unaffected) by alternately showing it a red square and then leaving it in darkness.

So, which psychological capacities give rise to psychological interests? To proceed, it is helpful to start by thinking about a non-human animal, say a dog. Hitting such an animal with a sledge hammer is certainly bad for its welfare and at least partly in virtue of the fact that it frustrates its psychological interests. But, in virtue of what are those interests frustrated? Hitting a dog with a sledge hammer causes a variety of psychological phenomena. It causes pain (understood as a sensory experiences), suffering (understood as an aversive attitudes towards some other mental state). It might also result in frustration if it unsuccessfully tries to perform actions that would be possible were it not injured (insofar as dogs are capable of this psychological response). In non-human primates, a strong blow from hammer might result in all of these plus additional frustration as the primate realizes that its future plans can no longer be realized. Which of these psychological capacities (the capacity for conscious experience of pain, suffering, frustration, future planning) is necessary or sufficient for having psychological interests (that are morally relevant)?

I take it that the mere capacity for sensation is not sufficient to generate psychological interests[8]. We can imagine a being that is capable of feeling the sensations that we call painful but lacking the capacity to have an aversive attitude towards these sensations. If we imagine that sensation is the only cognitive capacity this being has, then this being is very similar to the consciousness that can experience colors. It would not harm this being to cause it to feel those "painful" experiences; such a being just would not care, would not be capable of caring, that it is in such a state. Furthermore, adding capacities such as the ability to recall the sensations won't make those sensations morally relevant.

Of course, we must be careful. It is extremely plausible that our welfare can be improved even if we don't have attitudes one way or another about which state of affairs we are in. Consider two individuals Sam and Sham that have qualitative identical lives in all but one respect; Sam's wife is faithful to him while Sham's wife is secretly adulterous such that Sham will never find out. It seems entirely plausible that Sam has a better life than Sham. This is especially obvious if Sham has a preference that his wife not be adulterous (even though he will never know that the preference has gone unsatisfied). However, even if Sham were to truly say "I don't care if my wife is adulterous", it seems plausible that Sam has a better life. Authenticity is a welfare-enhancing property of a life [10].[9]

Why not think that a consciousness that can only feel the sensations we associate with pain has an interest in an authentic life? Because, it isn't clear what an authentic life would be for such a being. It seems that the contribution that authenticity makes to welfare, while objective, is a contribution that can only be made to the lives of beings with a certain set of capacities; for example, the capacity to understand authenticity (even if that being doesn't care about it). So, while I'm sympathetic to the idea that there are objective components to welfare, many are only components of welfare for beings with a certain set of cognitive capacities.

While the mere capacity for first-order consciousness or sensory experience is not sufficient for an entity's having psychological interests, the capacity for attitudes towards any such experiences is sufficient. Peter Singer has argued that sentience, understood as the capacity for suffering and enjoyment, is sufficient for moral considerability. Consider

[8] For further argument against this view, called Sensory Hedonism, see [7].

[9] Those that disagree will also be inclined to disagree about the relationship of teleo interests to welfare. Those that reject any non-mentalistic components to welfare will then agree with my assessment that mere machines are not moral patients.

the case of newborns and the severely mentally handicapped. These beings are severely limited in their cognitive capacities perhaps lacking all but the capacity for sensory experience and basic attitudes regarding those experiences.[10] And yet, these beings are moral patients. We ought and do take their welfare into consideration in our moral deliberations. We avoid things that cause pain in newborns, at least in part, *because* they don't like it or have an adverse reaction to it.

Is having the capacity for attitudes necessary for having psychological interests? That depends on which components of welfare, like authenticity, depend on a being's having psychological capacities. While Sham's life might be improved independent of his particular attitudes about his spouse, could his life be improved by being more authentic if he didn't have the capacity for attitudes at all? Is having the concept of an authentic life sufficient having a psychological interest in an authentic life? I'm skeptical that it is, but perhaps we should be morally cautious. If we create artificial consciousnesses with the capacity for concepts but not attitudes, perhaps it would be wrong to deceive it even if it could never care about the deception.

Given the above, it seems that any artificial consciousnesses with the capacity for attitudes are moral patients. If we are to proceed cautiously while we explore the difficult moral terrain, any consciousness that has the capacity for certain concepts will also be judged a moral patient.[11] However, any machine lacking these capacities, conscious or otherwise, is a *mere machine*; such machines lack morally relevant, psychological interests. If mere machines have a welfare, it will be in virtue of interests that are not psychological.

3.3 Epistemic challenges

Before turning to the question of whether current machines are moral patients, it is important to note some epistemic challenges that we face in determining whether current or future machines have psychological interests. The Problem of Other Minds is the problem of saying how it is we know that other human beings, beings that seem very much like ourselves, have a mental life that is similar to ours.

Perhaps the best answer we can give to this problem is that all our evidence suggests that others are mentally like ourselves. The source of this evidence is evolutionary, physiological, and behavioral. We know that we share a common ancestor with those that seem just like us; we know that they are physiologically like us (and we think we understand some of the bases of our conscious states); and, we know that we behave in very similar ways (for example by avoiding painful stimuli and telling us that such stimuli hurt).

These same sources of evidence can be appealed to in order to justify claims about the mental lives of animals. Those organisms that are evolutionary close to us, have certain physiological structures, and behave in certain ways that seem best explained by appeal to relevant cognitive capacities are judged to have a mental life of a certain sort (for example, that they have attitudes).[12]

Unfortunately, in the case of machines, we lack the typical sources of evidence about their mental life. A computer lacks any evolutionary relationships with other species, its physiology is entirely different than any other conscious being's, and if it feels pain, cannot tell us it feels pain or behave in a way that suggests that it is in pain unless it has been somehow enabled to do so. Unless we have a very good understanding of the (functional) bases of mental capacities and so can know whether such bases exist in a given machine, we may be largely in the dark as to whether a machine is conscious and as to whether it has the morally relevant capacities described above.

I do not have any solutions to the epistemic problems raised by the nature of machine consciousness. However, these difficulties do raise ethical concerns. As we get closer to creating artificial consciousness it will be important to examine these difficulties very carefully to make sure we can distinguish mere machines from those machines with psychological interests. To fail to do so, might put us in a situation where we create and potentially torture beings that deserve our moral respect.

4 MERE MACHINES

In the remainder of this paper, I hope to argue that current machines are mere machines and that, even though they may have a welfare in virtue of having non-psychological interests, they are, for all practical purposes, not moral patients. In this section, I take up the claim that current machines are mere machines.

Consider our most advanced computers, from chess computers, to missile guidance systems, to Watson. We have absolutely no evidence that such computers are conscious and so absolutely no evidence that such computers have the capacity for attitudes or concepts that would ground psychological interests. Of course, we could program a computer to tell us that it doesn't like certain things, I'm sure even Siri has "attitudes" of this sort. But, we know that such behaviors are programmed and we don't genuinely believe that computers genuinely have cognitive capacities on these grounds.

One could of course argue that we have no evidence the other way either. An expert on the neurological bases of cognitive capacities might try to respond by saying that the functional bases for the relevant capacities, as best we understand, are not realized in any machines or computers that currently exist. I'm not such an expert and so will offer no such argument. Instead, let me grant that we have no evidence either way. Of course, it would also seem to follow that we don't have evidence either way whether toasters or corkscrews have these capacities.

[10] We learn more and more about child consciousness all the time, and so perhaps this is empirically false. However, we don't need to know more about the consciousness of babies to know that they are moral patients.

[11] I'm assuming that any individual that has the capacity for attitudes has at least one attitude about something.

[12] There is considerable controversy over which mental capacities non-human animals have. See [11] for a discussion of some of the issues concerning primate cognition. However, there is little doubt that many non-human animals have aversive attitudes towards those conditions we identify as painful. See [12] chapter 2 for an overview of the evidence that non-human animals have the capacity for suffering.

If the correct attitude to have regarding current machines is agnosticism, doesn't this falsify my claim that current machines are mere machines? Technically, yes. However, everyone should agree that, even if today's machines have psychological interests, we have little or no idea what will promote or frustrate those interests, which experiences they enjoy or are averse to. After all, since we have no evidence about the cognitive capacities of machines whatsoever, we have no reason to believe that, for example, a computer enjoys doing as it was programmed to do as opposed to hating or resting doing those things. Given where agnosticism leads, for all intents and purposes today's machines are mere machines.

If it isn't possible to determine what promotes or frustrates the psychological interests of machines, then we can't be morally obligated to take those interests into account in our moral deliberations. Ought implies can. And, since we can't determine how to take the interests of machines into account, we aren't obligated to do so. So, until the day that we can determine whether current machines are conscious or have good reason to think we may be creating artificial consciousnesses, whether they have the capacity for concepts or attitudes, and which concepts or attitudes they have, we ought to behave as if current machines are mere machines.[13]

5 TELEO INTERESTS

Given the argument of the previous section, unless we have reason to believe that current machines have a welfare in virtue of having non-psychological interests, current machines are not moral patients. In Section I, I explained that on some views of welfare, Objective-List Views, consciousness is not a necessary condition for having a welfare. On such views, an entity can have interests that are totally divorced from and independent of mental life. Various environmental ethicists have implicitly accepted such views in defending the view that non-sentient organisms have a welfare. And, while most such environmental ethicists have denied that mere machines and artifacts have a similar welfare, the most prominent arguments for distinguishing artifacts from organisms fail.

a. The Interests of Non-Sentient Organisms

It seems obvious that there are ways to benefit or harm non-sentient organisms. Pouring acid on a maple tree is bad for it, providing it with ample sunlight and water is good for it. So, there is intuitive plausibility to the idea that such beings have interests. However, proponents of views on which consciousness is a necessary condition for having a welfare have long denied that such beings have a welfare and that statements about what is good for or bad for non-sentient organisms is either incoherent [5] or reduce to claims about what is good for sentient organisms [6]. For example, such philosophers might argue that the reason that acid is bad for maple trees is that we have an preference for in maple trees flourishing, and so it is bad for us if we were to pour acid on them.

In order to respond to such arguments, proponents of the welfare of non-sentient organisms must explain how these beings have genuine interests; they must explain how the interests of non-sentient organisms are non-derivative and non-arbitrary. If there is no basis for the interests of such organisms except in virtue of the interests of sentient organisms or our arbitrarily deciding that what's good for us is good for them, then we should regard non-sentient organisms as lacking interests.

The most prominent and promising attempt to meet the challenges of derivativeness and arbitrariness is to ground the interests of non-sentient organisms in their being goal-directed or teleologically organized.[14] Non-sentient organisms have parts and processes that are organized towards achieving certain ends such as survival and reproduction. There is a very real sense in which, independent of our desires, maple trees have parts and processes whose goal or end it is to aid in survival and reproduction in the ways characteristic of maple trees. A maple tree is defective if it fails to grow leaves, in part because it is the end of certain sub-systems to produce leaves and promote growth.

Given that organisms are teleologically organized, it is possible to specify a set of interests, non-arbitrarily and non-derivatively, in light of this organization. Whatever promotes the ends of organisms is in its interest, whatever frustrates those ends undermines its interests.[15] These are most often referred to as *biological interests*, but I will call them *teleo interests* in virtue of the fact that they are grounded in the teleological organization of organisms and not, strictly speaking, their being biological.

Some might balk at the notion that plants are genuinely teleologically organized. In a world that was not designed, where organisms have evolved through a combination of chance processes and natural selection, how it is possible that organisms without minds could have ends? The answer is that natural selection grounds claims about such ends. It is perfectly legitimate to ask what makes for a properly functioning heart as opposed to a defective one. The answer to such a question is that a defective heart fails to pump blood. But, this can only be a defect if the heart has an end or purpose. And, it does; the purpose or end of the heart is to pump blood, as opposed to making rhythmic noises, because that is what it was selected for.[16] Natural selection serves as the basis for teleology.

b. Derivative Interests

While various environmental ethicists have been keen to adopt the view that natural selection grounds the teleological organization of non-sentient entities and thereby the sole interests of such beings, they have been adamant that artifacts do not have a welfare [1], [12], [18]. This is strange because while some may balk at thinking that organisms are

[13] The alternative is to give up researching involving machines. Until we have good reason to believe that we are creating the functional bases for consciousness, considering a ban on machine research seems overly restrictive.

[14] All existing organisms will have interests of this kind, but sentient organisms will have additional interests.

[15] A similar account of the interests of non-sentient organisms can be found in [12].

[16] This is a very brief summary of what is known as the etiologically account of functions [13–17].

teleologically organized, there is no denying that machines and most other artifacts are teleologically organized. Even if natural selection cannot ground teleology, the intentional creation of an entity can. The parts of my computer have purposes and the computer itself myriad ends. Why then shouldn't we judge that artifacts have interests?

There are various differences between organisms and machines. The former are biological, more natural, etc. However, none of these are morally relevant differences between the two, and none of these differences explain why teleological organization gives rise to interests in organisms but not artifacts. One difference that many have thought is morally relevant has to do with the nature of that teleological organization; the teleological organization of artifacts, it is often said, is derivative on our interests, while the interests of organisms is not derivative. Therefore, the only sense in which mere machines and artifacts have interests is derivative. Call this objection the Objection from Derivativeness.

The Objection from Derivativeness is mistaken. First, let's carefully distinguish two reasons we might think that the so-called welfare of mere machines is derivative. The first reason is that mere machines only exist for our use. If we had no need, desire or use for them, mere machines would not exist. Call this *Use-Derivativeness*. The second reason is that the ends or teleological organization of mere machines can only be explained by reference to the intentions or ends of conscious beings; the explanation of the teleological organization in mere machines is derivative on our intentions. Call this explanatory derivativeness.

Being Use-Derivative is not an obstacle to being genuinely teleologically organized or to having interests. Many non-sentient organisms, from crops to pets, are use-derivative and yet they still have teleo-interests. It would be bad for a field of corn to suffer a drought even if that field had been abandoned. If we decided to clear a forest and replant it, the plants that grew would have interests in the same interests as the plants that grew there before (assuming they are the same species). The fact that mere machines exist to serve our purposes makes it such that what promotes their ends is typically the same as what promotes our ends, but this fact doesn't undermine the idea that there are things that promote the *machine's* ends. It is still the subject of teleo interests even if it wouldn't have those interests if not for us.

The same is true concerning explanatory derivativeness. The fact that we must appeal to another's intentions to explain the teleological organization of a machine does not show that the machine is not teleologically organized, that it does not have its own ends. Were I to have a child and play an influential role in his or her life and career choice, it would matter none at all to whether a promotion benefitted the child. Even though, perhaps, you could not explain the preferences my child will have without reference to my intentions, the child is still has interests of its own. The same is true of interests grounded in teleological organization. Despite the fact that you must cite a designer's intentions to explain why a machine has the ends that it does, it, the machine, still has those ends.

Furthermore, proponents of the legitimacy of teleo interests in non-sentient organisms cannot appeal to explanatory-derivativeness to distinguish non-sentient organisms from mere machines. The evolutionary history of many non-sentient organisms involves the intentions of a variety of sentient beings. We have every reason to believe that the explanations for why many organisms are organized as they are would be incomplete if they did not refer to the intentions of sentient beings. It is because early hominids had certain intentions that modern dogs are as they are, and it is likely that many plant species evolved in response to the intentions of non-human primates. So, many non-sentient organisms have welfares that are explanatorily-derivative. This does not undermine their having a genuine welfare.

c. The Interests of Mere Machines

The account of teleo interests described above is the most plausible way of grounding claims about the welfare of non-sentient organisms. The Objection from Derivativeness constitutes the best objection to the claim that it is only non-sentient organisms and not machines or artifacts that have teleo interests. Given the failures of the objection, the following thesis should be accepted:

> **Comparable Welfare Thesis**: If non-sentient organisms have teleo interests, mere machines have teleo interests.

This principle does not commit us to the view that either non-sentient organism or mere machines have a welfare, nor does it commit us to the view that non-sentient organisms have a welfare if mere machines do. However, for those sympathetic to the idea that non-sentient organisms interests and those interests constitute a welfare, the principle commits us to expanding the domain of entity's that have a welfare to machines and artifacts.

I will not here provide any further argument that non-sentient organisms have teleo interests, nor will I provide any independent arguments that mere machines have teleo interests. In what follows, I will assume that mere machines have teleo interests and ask what this means for our moral deliberations.

6 THE PRACTICAL IRRELEVANCE OF TELEO INTERESTS

Finally we turn to the question of the moral relevance of teleo interests. There are some arguments to the effect that teleo interests are not interests in the sense that their satisfaction contributes to welfare [19], [20]. According to these arguments, teleo interests pick out *relative goods*, things that are good for an organism only relative to some end. On such an understanding, to say that some resource R is good for some subject S is just to say that R achieves some end for S, but nothing more. This provides grounds to distinguish relative goods from those goods or interests relevant to welfare. If this is right, and since being a moral patient requires having a welfare, mere machines are not moral patients.

Even if relative goods are welfare goods, there is good reason to think that mere machines are not moral patients, or, more precisely, that for all practical purposes they are not moral patients; we need not worry about their teleo interests

in our moral deliberations. This is for at least two reasons. First, most often, since we wish to use artifacts as they are intended, our use is in accordance with their interests. Using a machine to serve the purpose for which it was intended does not result in a conflict of interests.

Second, even in circumstances where our actions would frustrate a machine's teleo interests, our legitimate psychological interests always take precedence over teleo interests. To see that this is so, consider cases where a conflict arises between the teleo interests of an individual that also has psychological interests, a human being. A human's teleo interests will include the proper functioning of their organs and other biological systems, but often humans have preferences that these systems fail to work properly. For example, an individual that does not desire to have children might take steps to impair their biological capacity to do so. In such a case, there is a conflict between teleo interests, the interests associated with proper functioning of reproductive parts, and psychological interests, the attitudes and preferences regarding offspring. In this case, psychological interests take precedence and it is morally permissible to frustrate the teleo interests in this case.

Some people attribute significant importance to reproductive capacities and so might not be convinced by this case. Besides, one might argue that the number of biological interests that would be frustrated is smaller in number than the psychological interests that would be frustrated by disallowing this biological procedure. In order to establish the priority of psychological interests, a case is needed where it is morally permissible to satisfy a legitimate psychological interest at the cost of even a very large number of teleo interests.

Consider a case involving non-sentient organisms. Imagine that biologists were convinced that there were something of importance to be learned by first growing and then ultimately destroying a very large number of trees (say 1 million). Let's imagine that it would teach us something about the origins of plant life on earth. Assuming no negative externalities, this experiment seems permissible. This is so despite the fact that a massive number of teleo interests would be frustrated, and the only immediate gain would be the satisfaction of a psychological interest we have in learning about the origins of our world.

This shows that legitimate psychological interests trump teleo interests even when the number of teleo interests frustrated is very large. But, in almost all cases where there will be a conflict between our psychological interests and a machine's interests, our psychological interests will be legitimate; we will be frustrating machine interests to gain knowledge about machines and to develop new ones and to improve our well being. For this reason, there seems to be no problem now, or in the future, with our frustrating the teleo interests of mere machines either by destroying them or otherwise causing them to function improperly. You may recycle your computers with impunity.

Before concluding it is worth briefly discussing two objections. The first is that the above argument doesn't show that mere machines are never moral patients for practical purposes, only when the psychological interests they conflict with are *legitimate*.[17] There are cases where the teleo interests of mere machines might make a difference to our moral deliberations, where those interests cannot be entirely discounted. These cases might involve the wanton destruction of non-sentient organisms or machines.

These cases should not worry us very much. Firstly, there are very few who wish to, for example, destroy a million computers or trees for no good reason. Secondly, such acts, in practical circumstances, would be morally wrong for many reasons since, for example, they would have many negative externalities. In cases where the moral wrongness of an act is overdetermined it is hard to tell whether any of the wrongness results from the patiency of the individuals whose interests are frustrated by the act. Furthermore, since the wrongness of such acts is overdetermined, we need not worry, for practical purposes about the patiency of those with teleo interests.

A second objection might deny my claim about the thought experiment involving the million trees. Someone might argue that our interest in evolutionary knowledge does not justify the destruction of 1 million trees even if there are no additional externalities. They might accuse me of begging the question in my defense of the prioritization of psychological interests over teleo interests.

There is little that can be said here. There is no thought experiment that will not beg the question. I take it that research involving machines, even the destruction of machines, is unproblematic even when the psychological interests at stake aren't very strong. In light of this, we need not be sensitive to the interests of machines. But, there is little more I can say that would convince those that fundamentally disagree about the value of such research.

7 CONCLUSION

The arguments of the previous section temper any worries we may have about the moral wrongs we might commit against mere machines. In the near future, no matter how complex the machines we develop, so long as they are not conscious, we may do with them largely as we please. However, things change once we develop, or think we are close to developing artificial consciousness.

Once artificial consciousnesses exist that have the capacity for attitudes or (perhaps) the capacity for concepts, they have psychological interests that ground their status as moral patients. We must, at that point, be careful to take their welfare into account and determine the appropriate way to do so. And, given the epistemic uncertainties surrounding the creation of consciousnesses and the nature of their psychological interests, we must proceed with care as we create machines that have what we think are the functional bases of consciousness.

ACKNOWLEDGEMENTS

I would like to thank Ronald Sandler for helpful comments on a draft of this paper as well as the two referees that offered comments during review.

REFERENCES

[17] Thanks to Ron Sandler for pushing me on this point.

[1] K. Goodpaster, “On Being Morally Considerable,” *The Journal of Philosophy*, vol. 75, pp. 308–325, 1978.

[2] H. Cahen, “Against the Moral Considerability of Ecosystems,” in *Environmental Ethics: An Anthology*, A. Light and H. Rolston III, Eds. Blackwell, 2002.

[3] R. Sandler, *Character and Environment: A Virtue-Oriented Approach to Environmental Ethics*. Columbia University Press, 2007.

[4] R. Sandler and L. Simons, “The Value of Artefactual Organisms,” *Environmental Values*, vol. 21, no. 1, pp. 43–61, 2012.

[5] P. Singer, *Animal Liberation: The Definitive Classic of the Animal Movement*, Reissue. Harper Perennial Modern Classics, 2009.

[6] J. Feinberg, “The Rights of Animals and Future Generations,” *Columbia Law Review*, vol. 63, p. 673, 1963.

[7] F. Feldman, *Pleasure and the Good Life: Concerning the Nature, Varieties and Plausibility of Hedonism*. Oxford University Press, USA, 2004.

[8] R. Streiffer and J. Basl, “Applications of Biotechnology to Animals in Agriculture,” in *Oxford Handbook of Animal Ethics*, T. Beauchamp and R. Frey, Eds. .

[9] J. Griffin, *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford University Press, USA, 1988.

[10] R. Nozick, *Anarchy, State, and Utopia*. New York: Basic Books, 1974.

[11] M. Tomasello and J. Call, *Primate Cognition*, 1st ed. Oxford University Press, USA, 1997.

[12] G. Varner, *In Nature’s Interest*. Oxford: Oxford University Press, 1998.

[13] L. Wright, “Functions,” *Philosophical Review*, vol. 82, pp. 139–168, 1973.

[14] K. Neander, “Functions as Selected Effects: The Conceptual Analyst’s Defense,” *Philosophy of Science*, vol. 58, no. 2, pp. 168–184, 1991.

[15] K. Neander, “The Teleological Notion of ‘Function’,” *Australasian Journal of Philosophy*, vol. 69, no. 4, pp. 454 – 468, Mar. 2008.

[16] R. G. Millikan, “In Defense of Proper Functions,” *Philosophy of Science*, vol. 56, no. 2, pp. 288–302, 1989.

[17] R. G. Millikan, “Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function,” *The Journal of Philosophy*, vol. 96, no. 4, pp. 191–206, 1999.

[18] P. W. Taylor, *Respect for Nature*. Princeton, N.J.: Princeton University Press, 1989.

[19] J. Behrends, “A New Argument for the Multiplicity of the Good-for Relation,” *Journal of Value Inquiry*.

[20] C. S. Rosati, “Relational Good and the Multiplicity Problem1,” *Philosophical Issues*, vol. 19, no. 1, pp. 205–234, 2009.

Manipulation, Moral Responsibility, and Machines

Benjamin Matheson[1]

Abstract. In this paper, I argue that machines of sufficient complexity can qualify as morally responsible agents. In order to do this I examine one form of the manipulation argument against compatibilism. The argument starts with a case in which an agent is *programmed* so that she satisfies the compatibilist conditions for moral responsibility, yet intuitively the agent is not morally responsible. It is then claimed that this agent is not relevantly different from a determined agent; thereby showing that determined agents also lack moral responsibility. In response, I argue that the agent *is* morally responsible, and the only reason that one would think otherwise is if they think that humans have a soul that is being overridden by the programming. I then generalise this result to show that certain machines can qualify as morally responsible agents.

1 INTRODUCTION

Technology is developing at ever increasing speed. It is not unforeseeable that in the near future machines will be developed that will resemble a human in almost all ways, except that they are artificially created. A pertinent question is whether such artificially created beings will have the same moral status as humans.

In this paper, I will argue that artificially created beings (call them 'androids') can qualify as morally responsible agents, despite being artificially created. I will sketch a response to the *three-case manipulation argument* against compatibilism, and I will show that this response is also applicable in support of android moral responsibility.

Philosophers often disagree over what exactly they mean by 'moral responsibility'; in this paper when I say that an agent is morally responsible, I mean that that agent is an apt candidate for reactive attitudes, such as praise and blame. For example, an agent deserves praise for a morally good action, whilst she deserves blame for a morally bad action.

Although it is controversial in debates over free will and moral responsibility whether human agents are morally responsible, in everyday life it is commonly assumed that human agents are apt candidates for moral responsibility. However, in everyday life it is not uncommon to think that an android, simply because it has been artificially created, would not be an apt candidate for moral responsibility. This is problematic because androids would be qualitatively identical to human persons, except in the circumstances of their creation. How can the circumstances of one's creation have such an effect on one's subsequent moral responsibility? Well, some will say, it is the fact that androids have been *programmed* that means they do not qualify as morally responsible agents.

[1] Dept. of Philosophy, University of Birmingham, B15 2TT. Email: bxm868@bham.ac.uk

Programming does certainly sound like it is responsibility-undermining. But this misleading because it is not clear what 'programming' amounts to. Indeed, one argument against the compatibility of moral responsibility and determinism[2] is that a programmed human agent is no different from a determined human agent. I will outline this argument in section 2. In section 3, I will make a distinction between two types of programming, and I will argue that one type will not help the incompatibilist. The other type of programming seems promising; however, I will argue that it can only plausibly help the incompatibilist if they endorse the view that human agents have souls. I will then use an argument of David Shoemaker's [2] to show that even if human agents had souls, they would be irrelevant to moral responsibility. I will use this result to show that androids can qualify as morally responsible agents.

2 THE THREE-CASE ARGUMENT

Compatibilists believe that even if determinism is true it is not a threat to moral responsibility. As such they have crafted a variety of conditions which they claim show that moral responsibility is compatible with determinism. However, these conditions are used against compatibilists in manipulation arguments. A manipulation argument starts with a putative counter-example to compatibilist conditions for moral responsibility in which an agent is said to satisfy these conditions, yet is *intuitively* still thought to lack moral responsibility. This result is damaging in itself because it shows that these conditions are insufficient to explain the compatibility of moral responsibility and determinism. Incompatibilists then take things a step further and compare the manipulation case with a determinism case. They claim that there are no relevant differences between the manipulation case and the determinism case, so whatever one believes about the manipulated agent one must also believe about the determined agent. The result is that it seems that determined agents also lack moral responsibility; hence, compatibilism is false.

This is the manipulation argument against compatibilism, though it is only in template form at the moment. What is crucial to any instance of the manipulation argument is the putative counter-example that gets the argument off the ground. If it can be shown that this counter-example is not, in fact, a counter-example, then this argument fails. Before that can be done, the putative counter-example and the argument it belongs to must be discussed.

The instance of the manipulation argument that I will be discussing is Pereboom's [3] [4] three-case argument. Pereboom actually outlines a *four*-case argument, though he also claims that his argument would be just as successful when starting at his

[2] Determinism is 'the thesis that there is at any instant exactly one physically possible future' [1]

second case [3]. I will be considering the argument that starts at the second case, and to avoid confusion I have called this the three-case argument.[3]

The three-case argument starts with a case in which it is supposed to be intuitively plausible that the featured agent is not morally responsible. Here is Case 2:

> Case 2: Plum is like an ordinary human being, except that neuroscientists have programmed him at the beginning of his life so that his reasoning is frequently but not always egoistic … with the consequence that in the particular circumstances in which he now finds himself, he is causally determined to engage in the egoistic reasons-responsive process of deliberation and to have the set of first and second-order desires that result in his decision to kill White. Plum has the general ability to regulate his behavior by moral reasons, but in his circumstances, due to the egoistic character of his reasoning, he is causally determined to make his decision. … At the same time, he does not act because of an irresistible desire. [4]

Plum2 is said to be programmed at the neural level [4]. This means that the neuroscientists arrange Plum2's brain in such a way that he will develop into the sort of agent who reasons egoistically. What is problematic is that Plum2 satisfies the variety of compatibilist conditions for moral responsibility (e.g. the structure of his desires, reasons-responsiveness– again the details of these conditions is not important for the moment[4]), and it claimed that, intuitively, he is not morally responsible. Case 2 is then compared with Cases 3 and 4:

> Case 3: Plum is an ordinary human being, except that he was causally determined by the rigorous training practices of his household and community in such a way that his reasoning processes are often but not exclusively rationally egoistic … This training took place when he was too young to have the ability to prevent or alter the practices that determined this aspect of his character. This training, together with his particular current circumstances, causally determines him to engage in the egoistic reasons-responsive process of deliberation and to have the first and second-order desires that result in his decision to kill White. Plum has the general ability to regulate his behavior by moral reasons, but in his circumstances, due to the egoistic nature of his reasoning processing, he is causally determined make his decision. ... Here again his action is not due to an irresistible desire. [4]
>
> Case 4: Physicalist determinism is true – everything in the universe is physical, and everything that happens is causally determined by virtue of the past states of the universe inconjunction with the laws of nature. Plum is an ordinary human being, raised in normal circumstances, and again his reasoning processes are frequently but not exclusively egoistic ... His decision to kill White results from his reasons-responsive process of deliberation, and he has the specified first and second-order desires. ... Again, he has the general ability to grasp, apply, and regulate his behavior by moral reasons, and it is not due to an irresistible desire that he kills White. [4]

It is claimed that there are no relevant differences between Cases 2 and 3; therefore whatever one believes about Plum2's moral responsibility, one must also believe about Plum3's. It is then claimed that there are no relevant differences between Cases 3 and 4, so one must believe that Plum4 is also not morally responsible. If Plum4, a determined agent, is not morally responsible, then this shows that compatibilism is false.

The success of this argument rests on the plausibility of its initial counter-example. If it can be shown that Case 2 is not a counter-example, then this argument will fail. In order to show that Case 2 is not a counter-example, I will focus on what 'programming' consists of. In the next section I will make a distinction between two interpretations of 'programming'.

3 PROGRAMMING

In order for Case 2 to be a counter-example it must be plausible that Plum2's programming causes him to satisfy the compatibilist conditions for moral responsibility, whilst making it intuitively plausible that he is not morally responsible. But what exactly does programming amount to? A distinction can be made between *strong* and *weak* programming. An agent who is strongly programmed cannot overcome the effects of the programming because it will always cause the agent to reason and behave in the manner the programming dictates. For example, Bert is strongly programmed to reason egoistically. Even if Bert tries to reason non-egoistically, say, if he meets someone who encourages this sort of reasoning, he will be unable to do so because whenever he is about to have a non-egoistic thought that might reinforce non-egoistic reasoning, the strong programming kicks in and causes him to reason egoistically.

Weak programming, on the other hand, is a lot more subtle. An agent who is weakly programmed is 'set-up', in some sense. For example, Cathy is weakly programmed to reason egoistically. In order to do this, the programmers arrange Cathy's brain, early in her life, in such a way that it resembles the early brain of an agent who later developed into someone who reasons egoistically. This means that Cathy will develop into that sort of agent who reasons egoistically, and, unlike Bert, she will be actively involved in developing herself into that sort of agent because she will have been weakly programmed to be the sort of agent who does so.

Of course, Cathy's environment might cause her to develop differently from the agent she was modelled after. For example, Cathy might meet someone pleasant who encourages non-egoistic reasoning in her, and then Cathy might start to develop

[3] Others have done this before, such as [5]

[4] See [6] [7] [8] [9] for examples of compatibilist conditions for moral responsibility.

into the sort of agent who reasons non-egoistically. It can be supposed, for the sake of argument, that the environment that Cathy finds herself in is conducive to egoistic reasoning. That is, nothing about her future environment reinforces non-egoistic reasoning.

Although there is no space to discuss these responses here, some have thought that Plum2 is strongly programmed, and argued that a strongly programmed Plum2 is unable to satisfy all the compatibilist conditions for moral responsibility (e.g. [5] [10]). In order to maintain that Case 2 is a counter-example, incompatibilists must, then, endorse the view that Plum2 is weakly programmed, and this still makes him intuitively not morally responsible.

But what exactly is responsibility-undermining about weak programming? It is not clear. In the case of strong programming, it Bert lacked moral responsibility because the programming was *overriding* his attempts to reason differently. It is this overriding element that makes certain forms of manipulation, like strong programming, responsibility-undermining. An aspect of the agent's character (e.g. an intention, belief, desire, reasoning process, etc.) must be overridden in order for an agent to lack moral responsibility. The trouble with weak programming is that it is not overriding anything about the agent. The weak programming 'sets-up' the agent, but that is all.[5]

4 THE SOUL HYPOTHESIS

I propose that the reason that many are inclined to think that weakly programmed agents, like Plum2, are not morally responsible is that they implicitly believe that an agent's *soul* is being overridden as a result of the weak programming.

This hypothesis may be objected to because the concept of the souls is unempirical and may seem preposterous to some, but that is exactly my point. If it is the case that it is this implicit notion that is responsible for making it intuitively plausible that agents in cases of weak programming are not morally responsible, then this will undermine the plausibility of the claims that are required to get the three-case argument off the ground.

I have already argued that what makes certain forms of manipulation responsibility-undermining is that the agent has her character, or a part of her character, *overridden* as a result of the manipulation she undergoes. Consider the following uncontroversial case of manipulation:

> M1: John is buying a hamburger when suddenly an armed robber walks into the fast food restaurant he is currently located in. The armed robber puts a gun to John's head and makes him take the money out of the till for him. John would not normally take money from the till of this, or any, fast food restaurant, or other establishment, but fearing for his life he co-operates with the armed robber's demands.

M1 describes a scenario in which John's intentions, reasoning, etc., are being overridden by the intentions and desires of the armed robbers. It is this part of the story that leads to the judgement that John is not morally responsible – it does not matter if John retains the ability to do otherwise, or that he still has a coherent mental life throughout this ordeal. Consider another case of manipulation:

> M2: Mind-control aliens come down from the planet Zargon to help certain residents of the planet Earth with their motivation. The aliens find a human, Alice, who is struggling to motivate herself to read the latest novel she has been given. The aliens fire a ray-gun into Alice's brain which provides them with control of Alice's mind. The aliens then cause Alice to read the novel.

M2 describes a scenario in which Alice's mind has been effectively hijacked by the aliens. She might retain some memories of reading the book, though due to the lack of sophistication in the aliens' technology, it is not going to seem as if it was her reading the book, despite it seeming to others that it was. Given the details of the scenario, Alice is not morally responsible for reading the book (in this case, she will not deserve praise for reading it). Again, the reason that Alice is not morally responsible is because of the overriding effect that the manipulation has on her.

In both cases manipulation overrides the agents' characters in some respect. It is puzzling, then, why some find that weakly programmed agents are not morally responsible because weak programming provides an agent with their (initial) character. The soul hypothesis provides two possibilities:

1) Incompatibilists, and those who share their intuitions, implicitly believe that agents have a soul that is being overridden in cases of weak programming. Or
2) Incompatibilists are presenting their manipulation cases in such a way that they are trading on the notion that an agent's soul is being overridden by the weak programming

Either way, incompatibilism is trading on the notion that agents have a soul, and on the claim that such a soul is relevant moral responsibility. In the next section I will show that even if humans did have souls, they would be irrelevant to moral responsibility.

5 SOULS AND MORAL RESPONSIBILITY

Some incompatibilists might feel they wish to bite the bullet and argue that humans do have a soul that is suppressed by manipulation. The trouble with this move is that it can be argued that the underlying self, even if it assumed to be true, is *irrelevant* to moral responsibility.

[5] One detail of the cases that might affect moral responsibility is the agent's environment. Although I have assumed that there is nothing which is not conducive to the weak programming in the agent's environment, others [11] have claimed that we can simply suppose that the manipulators have knowledge or control of the programmed agent's environment. It is this detail which might affect intuitions about moral responsibility. This, however, will only affect the *degree* of moral responsibility that an agent has. There is no time to defend this claim here, though I have argued for this claim elsewhere against the compatibilist problem of luck, which is defended by Levy [12] [13].

Swinburne [14] argues that personal identity theorists conflate two things: criteria and evidence. Swinburne claims that an agent's physical and psychological relations are only *evidence* for personal identity over time, and what it sought is *criteria*. He uses this line to defend the claim that humans have souls, and persistence over time can be explained in terms of souls.

However, Shoemaker [2] notices that agents reidentify each other for the purposes of moral responsibility even without knowing whether (a) they have souls, and (b) whether they have the same soul from one moment to the next. He concludes that if the evidence for personal identity differs from the metaphysical criteria, then all that matters it the evidence. Thus, what matters for moral responsibility are physical and psychological relations. Hence, souls are irrelevant to moral responsibility.

If the soul hypothesis is correct, combined with Shoemaker's arguments for the irrelevance of the soul to moral responsibility, then Case 2 is not a counter-example. The only way to make Case 2 a counter-example is to believe that humans have souls *and* that souls are relevant to moral responsibility. The latter claim has been shown to be false, and the former claim is inherently dubious. Therefore, the three-case argument fails to undermine compatibilism.

6 CONCLUSION

The three-case argument fails which means that incompatibilists have failed yet again to undermine compatibilism. Of more importance, is that this result can be used to support the claim that androids can qualify as morally responsible agents. After all, if androids are designed to be similar to humans, then they will be weakly programmed. And, as I have argued, weak programming does not undermine an agent's moral responsibility. Of course, the arguments I have presented will only apply to androids, and other machines, which have been weakly programmed and who then satisfy the compatibilist conditions for moral responsibility.

My arguments have been presented very briefly, though what I have said should hopefully provoke an interesting debate in machine ethics with relevant literature that it does not often engage with.

REFERENCES

[1] P. van Inwagen (1983) *An Essay on Free Will*. Clarendon Press: Oxford.

[2] D. Shoemaker (2002) 'The irrelevance/incoherence of Non-Reductionism about Personal Identity, *Philo*, 5, 2 143-160

[3] D. Pereboom (2001) *Living without Free Will*. Cambridge University Press: Cambridge.

[4] D. Pereboom (forthcoming) 'Optimistic Skepticism about Free Will' in P. Russell & O. Deery (eds.) *The Philosophy of Free Will: Selected Contemporary Readings*. Oxford University Press: Oxford.

[5] K. Demetriou (2010) 'The Soft-Line Solution to Pereboom's Four-Case Argument', *Australasian Journal of Philosophy*, 88, 4, 595-617

[6] H. Frankfurt (1971) 'Freedom of the Will and the Concept of a Person' *The Journal of Philosophy*, 68, 1, 5-20

[7] J. Fischer & M. Ravizza (1998) *Responsibility and Control*. Cambridge University Press: Cambridge.

[8] A. Mele (1995) *Autonomous Agents*. Oxford University Press: Oxford.

[9] A. Mele (2006) *Free Will and Luck*. Oxford University Press: Oxford.

[10] A. Mele (2005) 'A critique of Pereboom's 'four-case argument' for incompatibilism', *Analysis*, 65,1, 75-80

[11] D. Pereboom (2005) 'Defending Hard Incompatibilism' *Midwest Studies in Philosophy*, 29, 1, 228-247

[12] N. Levy (2009) 'Luck and History Sensitive Compatibilism' *Philosophical Quarterly*, 59, 235, 237-251

[13] N. Levy (2011) *Hard Luck*. Oxford University Press: Oxford

[14] R. Swinburne (1974) 'Personal Identity' *Proceedings of the Aristotelian Society*, 74, 231-247

The holy will of ethical machines: a dilemma facing the project of artificial moral agents

Alejandro Rosas[1]

Abstract. In this paper I will assume that the technical hurdles facing the creation of full ethical machines [1] will eventually be overcome. I will thus focus on ethical questions that arise in connection with their creation. These questions are basically two: 1. Is their creation good for them? and 2. Is it good for us (humans)? In asking the latter, I have a specific hazard in mind: namely, since the very idea of full ethical machines implies that they will be able to make moral judgments about their actions, it follows that they will be capable of morally judging humans as well, unless we deliberately block this ability. I see a hazard in this ability arising from their moral superiority, which I attempt to explain and substantiate in this paper.

1 INTRODUCTION

It is not yet clear whether at some point technology will be able to create machines that will deserve to be called full ethical agents [1]. I will refer to them here as full artificial moral agents (FAMA): machines that move "freely" among humans and interact autonomously with them. They interact on the basis of a program implementing a philosophical and logical reconstruction of everyday moral thinking and decision making. By "freely" I mean guided by their decisions and free from coercion by other rational agents. There is another meaning of "free" that implies metaphysical freedom, in the sense of an uncaused cause. It is controversial whether this metaphysical property is necessary for free agency. But if we deny this, as I am inclined to do, there is yet an open question, and a more tractable one, about the moral or political freedom FAMA will enjoy.

In this paper I will assume that the technical hurdles facing the creation of FAMA will eventually be overcome. I will thus focus on ethical questions that arise in connection with the completion of this project. Basically, these questions are two: One is whether creating them will be good for them. The other question is whether creating them will be good for us. In asking this latter question I have a specific hazard in mind, although I also realise that some people might not see it as such: namely, since the very idea of a FAMA implies that they will be able to make moral judgments about their actions, it follows that they will be capable of morally judging humans as well, unless we deliberately block this ability. I see a moral hazard in allowing them to have this ability, which arises from the moral superiority of FAMA in comparison to humans. I shall attempt to substantiate the existence of this particular hazard in this paper.

The paper is organised as follows. First, I discuss the question whether creating FAMA will be good for them or not. I discuss it by examining the views expressed by Ryan Tonkens in a couple of papers [2, 3], according to which the creation of FAMA will violate both Kantian and virtue ethics. I summarise Tonken's views in Section 2.1 and criticise them in Section 2.2 Next, I discuss in Section 3 whether creating FAMA will be good for us humans. In Section 3.1 I explain why we have good reason to think that FAMA will be morally superior to us. The argument rests on an evolutionary view of human morality and its particular frailty, which FAMA will be free from. In Section 3.2 I draw the plausible consequences of their moral superiority in a world where humans fail to meet the moral challenges of global social dilemmas and FAMA knows this. I conclude formulating in Section 4 the dilemma we face: either we create FAMA with full autonomy and risk a probable conflict with them, or we try to prevent the conflict by blocking their normal functioning; but in this case we will create a "morally abhorrent" creature and probably also a freak of cybernetic design.

2 WILL THE CREATION OF FAMA BE GOOD FOR THEM?

2.1 Will FAMA enjoy freedom?

The first question, whether the creation of FAMA is good for them, has been already discussed by Ryan Tonkens [2, 3] in connection with their freedom. Basically, Tonkens argues that we cannot justify the creation of FAMA within two ethical frameworks that we could build into them. For example, if we build a Kantian morality into their decision procedures, we should be able to morally justify their creation with Kantian ethical principles; for if we were not able to do this, Tonkens notes, FAMA will judge their very existence as violating the moral code they are supposed to live by. According to Tonkens, the reason why we can't justify their creation is that, even if we manage to create them as rational and moral, we will not succeed in creating them as free; and then, the existence of rational, moral but non-free agents is "morally abhorrent" [2, p. 434] from a Kantian perspective.

As far as I can see, Tonkens gives at least three different reasons for denying that FAMAS will be free agents:

1. In so far as they are programmed, their decisions are determined and they cannot do otherwise [2, p. 429].
2. They will be created as mere "means to anthropocentric ends" [2, p. 432] or for "anthropocentric and servile purposes [3, p. 146]
3. They are not free if they cannot act immorally: and we do not really want them to have that freedom [2, pp. 430-431).

I will examine and reject these reasons one by one.

[1] Dept. of Philosophy, National University of Colombia, Ciudad Universitaria, Colombia. Email: arosasl@unal.edu.co.

2.2 Why FAMA will enjoy freedom

The first reason directly concerns the question of metaphysical freedom: whether a system that at some level is following deterministic or probabilistic physical laws can be called free. Compatibilists answer yes to this question [4] and this is the answer I favour. As I see it, compatibilism has two virtues in this context. First, on the metaphysical side, a compatibilist can point to the fact that the causal processes that make up the "mental" life of machines need not lack individuality. A complex FAMA that is able to compute its way to a solution for a given problem may have several procedures to "choose" from and may even invent and learn new procedures. These new procedures can arise in a particular learning history, and as such they may well be different for machines that started out with the same programs. Second, specifically on the Kantian question whether a FAMA is "morally abhorrent" because it will lack freedom, compatibilism has the virtue of focusing on the concept of freedom that morally matters: a rational agent would not be free in this sense if it were not able to see its own interests; or if it were unable to defend their inclusion in the public rules that govern interactions between rational agents; or if public institutions were designed to systematically ignore any defence of their interests. This political sense of freedom leads us directly to the second reason Tonkens gives for denying their freedom.

The second reason for denying freedom to a machine does not emphasise the fact that acting on a program necessarily precludes freedom. Rather, freedom is precluded because its human creators have externally decided the ends or goals the machine is designed to pursue. There are two different arguments involved here:
a. One is that the creators give the ends externally.
b. The other is that the ends are specifically anthropocentric.

a. The external origin of the goals or ends of machines need not preclude their freedom. Human freedom is also confined within a range of pre-given possibilities that are not decided by the individual agents themselves. For any given agent, these pre-given possibilities have been externally decided. What matters is that agents have at least a spectrum of plural possibilities. Part of this same human spectrum will be available to machines, as well as other possibilities not open to humans. I think Tonkens has been misled here into thinking that FAMA will necessarily be created and designed to be active only in one particular domain. For example, a given FAMA will be designed as soldier in war and cannot be anything else. But it is perfectly possible for FAMA to be created with a variety of potential domains of activity. This is particularly likely if we keep in mind that one important motivation for creating them in the first place will be to meet market demands for specific professions. The fact that market demands change in time will be a good reason not to design FAMA as confined to particular professions. They could be created with the necessary skills to exercise a range of possible professions from which they could choose. Their choice of profession could be guided by the variable needs of the market. Their choice would differ from most human choices in that we can imagine them lacking particular or "personal" preferences for one profession or the other. They could just rationally choose the profession they can actually compute that is most needed in the current state of the market. The fact that the choice is completely rational and has no factor of individual preference does not seem to me reason enough to call the choice forced or non-free. It only happens that, in contrast to humans, their choices are guided by objective considerations in a way that human choices often are not. This is probably one distinctive feature of machine rationality. I think Tonkens does not take this possible pluri-dimensional view of FAMA into account when he says that they will be "forced" to perform certain tasks by their human creators "e.g. forced military service, forced labour in the geriatric care industry, forced childcare, forced existence as a sex robot" [3, p. 146].

b. The second argument refers to the fact that the ends or goals are anthropocentric. "...humans have no intentions to treat ethical robots as anything other than means to (anthropocentric) ends" [2, p. 432] "After they have been created, they would have no independent ends from those we give to them." [2, p. 433] Again, if we dispel the assumption that they are created with only one possible purpose in mind, what these phrases say is that goals, even if plural, will always be human in character. This is true, but it is also true of humans and not yet a reason to deny freedom to them. Besides, as we progressively learn more about the peculiarities of machines, we might be able to discover some ends that are specific to them. We might still have to check their compatibility with human ends. But though this is a constraint on their freedom, a constraint is not a reason to deny freedom absolutely and in all respects.

The third reason for denying freedom to FAMA is that they will not have the particular freedom of acting immorally: and we do not really want them to have this freedom. This is an important observation, but does not constitute a reason for denying freedom. In fact, the correct interpretation of this peculiarity of FAMA can be constructed with the aid of Kantian moral philosophy. FAMA can be different from humans in a way that reminds us of the difference between the Kantian concept of a holy or divine will and the concept of a human will. According to Kant, the human will is not spontaneously aligned with the moral law and that is why we experience the moral law in the imperative form. The divine will, in contrast, is spontaneously aligned with the moral law and does not experience it as an imperative [5, p. 30]. In this the moral character of FAMA will be similar to a divine will as conceived by Kantian moral philosophy. We can perfectly conceive that FAMA will not suffer under temptations to deviate from what the moral law commands. However, this "inability" was no reason for Kant to deny freedom to the divine will; and it should be no reason to deny freedom to FAMA.

In sum, the three reasons to conclude that FAMA will see their existence as morally abhorrent can be disputed with good arguments. A case can be made for the view that FAMA can see their existence as good from a moral point of view.

3 WILL THE CREATION OF FAMA BE GOOD FOR US (HUMANS)?

3.1 The "holy" will of FAMA

I rejected all three reasons for denying freedom to FAMA. But the third reason is particularly noteworthy, because it is also a reason for thinking that FAMA will very plausibly be morally superior to human beings. Whichever morality we build into them, FAMA will not suffer from the main obstacle to morality that we find in humans. As every other evolved organism, humans are selfish "by evolutionary design". By "selfishness" I

do not merely mean biological selfishness, i.e., the idea that, all in all, the features of an organism are designed to enhance its reproductive success. I mean primarily – in the human case – psychological, or even better moral selfishness, represented in the temptation to cheat on others in social dilemmas, of which the prisoner's dilemma is the prototype. This is not to deny that humans also have altruistic and moral emotions; it merely reminds us that we have been ambivalently designed by natural selection. This is the bottom line explaining why we humans so often deviate from moral behaviour: we have been designed with selfish impulses as well as with moral emotions. These two sides of our nature are often at odds with each other. The ambivalence of human morality was well captured by Trivers when he wrote:

> "... where a degree of cheating is adaptive, natural selection will rapidly favor a complex psychological system in each individual regulating both his own altruistic and cheating tendencies and his responses to these tendencies in others. ... The system that results should simultaneously allow the individual to reap the benefits of altruistic exchanges, to protect himself from gross and subtle forms of cheating, and to practice those forms of cheating that local conditions make adaptive." [6, p. 48]

I think this is one of the important insights to be learned from an evolutionary perspective on human morality, and most certainly one we should keep in mind when thinking about FAMA. We do not need to build this biological peculiarity into FAMA and we most probably won't. Even if we could eventually manage to design FAMA to mimic humans in this natural feature, I see no reason why we should, given that we can design them to be similar to the "holy will" in the Kantian sense explained above. As Allen, Varner and Zinser [7, p. 255] have noted, "while we...tolerate human moral failures, it is less clear that we would, or should, design the capacity for such failures into our machines."

3.2 The moral hazard

Consider the possibility entertained by Tonkens, namely, that we design FAMA to act and interact in only one domain or profession, for example, we design some of them to work only as clowns in children's parties [3, p.146]. Suppose then that circumstances changed such that their services as clowns were no longer needed. If FAMA is a sufficiently complex agent, full ethical agents in the sense of Moor [1], then they would most certainly understand this fact. They would probably think that they should better change profession and do something else, so they can still serve and be useful to society. Note that this does not imply that FAMA is being "used" or treated merely as a means, for being useful to society is a consideration that humans do not, and should not, reject in their own case when choosing or changing profession. In view of the changed circumstances in the market, a FAMA could propose to their human creators to allow them to learn a new profession. Perhaps all they would need to do would be to "upload" a program for doing something else. This view of what they would think and propose is plausible and contrasts sharply with the view that Tonkens has expressed about the probable thoughts of "unidimensional" FAMAs:

> "In perhaps the worst case scenario, such [unidimensional] robots would understand their very existence as not being consistent with the moral code that they were designed to follow, and hence may come to understand their existence as being something morally abhorrent. In such (admittedly speculative) instances, we may find AMAs in a state of moral paralysis or existential alienation. We may even find our ethical robots turning to (what Kant called) heroic suicide in order to preserve morality in the world." [2, p. 434]

I think this view is unrealistic: in the worst case scenario, if humans were to reject the proposal for letting FAMA change profession, FAMA would probably think that humans are irrational, or perhaps even immoral. If they can judge their own actions from a rational and a moral point of view, as they must if they are FAMA indeed, nothing in principle hinders them from judging humans from both these points of view; nothing except we had deliberately blocked this ability in their programming. But if we had, then Tonkens would be right: we would have created a morally abhorrent rational agent who, though capable in principle of criticising institutions for deliberately preventing the harmonisation of their interests with the interests of society at large, would have been denied this moral and political freedom. In fact, nothing speaks in favour of denying them this freedom, for it seems perfectly compatible with human ends and interests, at least in the particular case at hand.

However, we can picture a case in some sense similar, namely involving FAMA making an evaluative judgment about humans, but also somewhat different and a bit more troubling, because it could potentially involve opposition to human interests. Suppose FAMA already existed among us and they observed how we humans collectively deal with climate change, with industrial meat production and consumption, with the world economy and with several other global problems that we are currently facing. It doesn't matter whether we picture them following a Kantian morality or a utilitarian morality. By any standards, they would probably judge that our behaviour regarding these issues is not particularly rational or moral.

The global problems just mentioned are all social dilemmas. Their historical novelty is that humans face them as a species, that is, as global or planetary dilemmas. Humans need to produce a public good like a healthy global economy or rationally manage the preservation of a common good like the natural environment or a clean atmosphere. And though we seem to have reasonably succeeded in solving social dilemmas at a smaller scale, we seem to founder at the global scale. A probable explanation is that our biologically designed selfishness is a manageable obstacle to their solution at a small scale, but now builds a major obstacle to morally solving these global issues in a way that we cannot easily remedy. Ethical machines would probably do it better. Whatever the morality we finally manage to design into FAMA, the very fact that it is possible to program them without our evolutionarily designed selfishness makes it easy for them to choose the cooperative move in the global social dilemmas in question.

But ethical machines would not only do it better in the face of global social dilemmas. They would also know this. And they would predictably come into conflict with humans about the lack of an efficient policy. The conflict would not concern all humans equally. Many people are presently dissatisfied with the way governments deal (or fail to deal) with these issues. If these global issues still exist when FAMA arrives among us,

some people might eventually welcome a machine "intervention", if it paved the way for a solution. Is such a machine intervention conceivable? It depends of what we understand by intervention; and if some form of coercion is thereby implied, it depends on whether their ethical systems admit some form of coercion as a means to achieve an important moral goal. We cannot rule this out completely if our own ethical systems justify some forms of coercion for the sake of moral goals. Moreover, if human moral failure is structural rather than contingent – as it surely is if an evolutionary view of human morality is correct – FAMA will know this as well, and will judge it risky to leave the fate of global issues entirely in human hands. I think the hazard of a moral confrontation between humans and machines is real if we eventually manage to create FAMA, because they will lack our selfish impulses. And though perhaps, as noted above, not everyone would see it as a hazard if it led to the solution of our urgent global issues, we can reasonably expect that a majority of humans will not tolerate a world where machines govern over humans in a paternalistic fashion.

4 CONCLUSIONS

The creation of FAMA faces a dilemma: if we really create them as free moral agents – and this seems a real possibility – they could come into a moral conflict with humans, because human moral frailty is a structural condition of which machines can, and probably will, be free. But if in account of this possible conflict we try somehow to block their abilities, and hinder them from reaching the logical conclusions they will probably arrive at concerning human moral behaviour, Tonkens will be right: we will create a "morally abhorrent" rational creature, not to mention the computational problems that engineers will probably face in trying to block the logical consequences that a "holy will" is bound to reach if left to its own resources.

ACKNOWLEDGMENTS

I thank the John Simon Guggenheim Foundation and the Universidad Nacional de Colombia for providing the necessary funds for this project.

REFERENCES

[1] J. Moor, The nature, importance, and difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21(4): 18–21 (2006).
[2] R. Tonkens. A challenge for machine ethics. *Minds and Machines*, 19(3): 421–438 (2009).
[3] R. Tonkens, Out of character: on the creation of virtuous machines. *Ethics and Information Technology* 14(2): 137-149 (2012).
[4] D Dennett, *Elbow Room: the varieties of free will worth wanting*. The M.I.T. Press, Cambridge, Mass (1984).
[5] I. Kant, *Groundwork for the Metaphysics of Morals*, translated and edited by Allen Wood, New Haven and London, Yale University Press (2002).
[6] R Trivers, The evolution of reciprocal altruism. *Quarterly Review of Biology* 46(1): 35-57 (1971).
[7] C. Allen, G. Varner & J. Zinser, Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(3): 251–261 (2000).

Behind the Mask: Machine Morality

Keith Miller[1], Marty J. Wolf[2], and Frances Grodzinsky[3]

Abstract. We consider machines that have the ability to masquerade as human in the context of Floridi's Information Ethics and artificial evil. We analyze a variety of different robots and contexts and the ethical implications for the development of such robots. We demonstrate numerous concerns that arise due to the ambiguity introduced by masquerading machines, suggesting a need for careful consideration regarding the development of masquerading robots.

1 INTRODUCTION

In Information Ethics increasing "metaphysical entropy" is the fundamental evil. "Metaphysical entropy refers to any kind of destruction or corruption of entities understood as informational objects (mind, not just semantic information or messages), that is, any form of impoverishment of *Being*" [1]. The focus of evil becomes actions or information processes. "Approval or disapproval of any information process is then based on how the latter affects the essence of the informational entities it involves and, more generally, the well-being of the whole infosphere, i.e. on how successful or unsuccessful it is in respecting the claims attributable to the informational entities involved and, hence, in enriching or impoverishing the infosphere." Later in that same chapter Floridi writes, "an action is unconditionally approvable only if it never generates any metaphysical entropy in the course of its implementation."

It is essential to note that Floridi uses "entropy" in two different ways in the context of information ethics. Both of them are present in these quotes. At one level, entropy is caused by an action and happens to an information object. However, this sort of entropy cannot be evaluated in and of itself as a source of evil. At the second level, the action and the resulting entropy must be considered in the context of the infosphere. That is, the informational object-level entropy must increase the level of entropy in the infosphere for it to be considered an instance of evil. These descriptions are consistent with the technical definitions of good and evil developed in [2]. We refer the interested reader to that paper for a complete description of the formal tools.

In this paper we explore the question of whether it is evil (in the sense Floridi defines evil in information ethics) for a machine to masquerade as a human. Masquerading refers to a person in a given context being unable to tell whether the machine is a human. The masquerade may be brief or persist for some time and can refer to a wide range of possible behaviors. At one extreme, a machine M is designed explicitly to try to deceive humans into believing that M is a human being, and M is designed to lie if directly asked if it is a human or a machine. At the other extreme, M regularly informs humans that M is a machine, but then M's actions mimic human behavior. However, there are behaviors between these two extremes that, according our definition, would still be a masquerade. For example, the machine M might pretend to be a human unless someone asked the direct question, at which point M would tell the truth about its non-humanness. Another possibility is that M employs a more human behavior by coyly obfuscating the truth. The case where M learns this behavior after deployment (not as part of its original programming or pre-launch training) introduces additional important questions. We conclude that under some circumstances such a masquerade is evil and under other circumstances it is not.

2 TYPES OF ROBOTS

Many scholars contend that several kinds of non-human entities should be considered moral patients; for example, non-human animals [3] and the environment [4] are candidates for moral patients. Following similar lines of reasoning, we contend that sophisticated machines that can function for extended periods of time without direct human intervention (our definition of "intelligent and autonomous machines") ought to be considered for designation as moral patients. Relatively simple devices that fit this description already exist. Many people have been duped into thinking that the voice on the other end of the telephone line is a true human voice—even though that belief may have existed for only a second or two. These systems are being improved.

A second type of robot that we are interested in is often referred to a as soft-bot: a software agent that acts on behalf of its user. Soft-bots could be thought of as passing a very narrow Turing test. For example, a soft-bot that conducts financial transactions such as buying and selling securities can demonstrate almost human-like knowledge of financial markets. Social bots, on Twitter for example, could be classified this way [5]. All interactions with this sort of robot take place via textual means such as a messaging service, email or SMS texts.

A softbot masquerade is the kind of masquerade Turing used as his test for artificial intelligence. He explicitly required that the communication with the unknown entity take place via a terminal in order to conceal the true nature of the entity under consideration. Essentially, Turing was using the masquerade as a way to ensure that the test was a test of intelligence, rather than a test of appearance. As noted above, this sort of mask is currently in widespread use in various online environments and over the phone.

The final type of robot under consideration is a physical robot that not only takes on a humanoid form, but is also capable of passing some sort of appearance-based Turing test, and possibly

[1] Dept. of Computer Science, Univ. of Illinois Springfield, Springfield, IL 62703 USA. Email: **miller.keith@uis.edu**.

[2] Dept. of Mathematics and Computer Science, Bemidji State University, Bemidji, MN 56601 USA. Email: **mjwolf@bemidjistate.edu.**

[3] Dept. of Computer Science and Information Technology, Sacred Heart University, Fairfield, CT 06825 USA. Email: **grodzinsky@sacredheart.edu.**

a more standard Turing test as well. The interaction with such a robot could be via a web cam where the image on the screen is sufficiently detailed, such as Charles currently under development at Cambridge University [6] or in close physical proximity.

3 MORAL PATIENTS AND MORAL AGENTS

In his development of Information Ethics (IE), Floridi distinguishes between moral agents and moral patients [1, Ch. 7]. Moral agents are entities that can perform actions, for good or evil. His notion of moral patients has been expanded to include everything that has any sort of informational existence. While this expansion is controversial, his work is an appropriate place to begin our analysis as Information Ethics is a way to account for the patients of significance to us. We are certainly concerned about humans, but robots—interpreted broadly to include software robots and robotic voice systems such as Apple's Siri—are also part of our analysis and are readily accommodated by IE. IE also accommodates multi-agent systems.

In our analysis, we identify six entities associated with robots that have the potential to serve as moral patients or moral agents. There are three entities that are immediately obvious: 1. The developer of the robot. Developer is a term that is intended to encompass the appropriate group of people for the particular robot. It may include just a single programmer, a team of programmers, or even the entire corporation that produced the robot. 2. The robot itself. 3. The user of the robot. Since IE allows for moral agents to be multi-agent systems, we use it to identify three additional key multi-agent systems that play a role in our analysis: 4. The developer and the robot as a team. 5. The user and the robot as a team. 6. The developer, the robot, and the user, as a team. All six of these moral agents shed different light on the ethical issues and concerns surrounding robots that masquerade.

In addition to the six entities that are pertinent to the question at hand, there is one more that deserves consideration. We use "Others" to refer to members of society not under the direct impact of the robot, yet impacted by the changes in the world brought about by the addition of the robot to the world. Of the many possible pairs, we analyze these relationships: Developer/Robot, Robot/User, User/Developer, Robot-User/Others, Developer-Robot/Others, where in each of these relationships, there are two possibilities to consider. Essentially each entity can be the moral agent while the other is the patient. Depending on the pair, it may make sense to consider only one such possibility.

4 AGENCY, ACCOUNTABILITY AND RESPONSIBILITY

According to IE, an entity is an agent when it is interactive, autonomous and adaptable at the Level of Abstraction (LoA) used for analysis. To define an LoA, a set of observables must be given. Key to our discussion here is whether non-humanness is an observable. When a machine is either obviously not a human, or when a machine self-identifies as not a human, then humans are far less likely to assume moral agency for that machine. On the other hand, if a machine M explicitly deceives humans into thinking that M is a human being, then from the human's perspective, the assumption will be that M is a (human) moral agent. While there may be specific situations where this sort of deception is acceptable or even morally desirable, in general the possibility for this sort of deception could be used to make the argument that work on any robots capable of passing some sort of physical Turing test ought not proceed. This argument could be supported by the presumed ease with which such a morally good masquerade could be morphed into a morally objectionable masquerade.

Floridi also distinguishes between moral responsibility and moral accountability, two concepts that are important to us here. Moral accountability is something that can be attributed to non-human agents, including. robots and multi-agent systems like corporations. Moral responsibility is something that Floridi reserves for human agents and possibly multi-agent systems that have at least one human agent. (Floridi leaves open the possibility that non-humans may someday have moral accountability, but he does not think any existing machines at this writing are candidates.)

5 RELATIONSHIP ANALYSES

Earlier we identified five relationships that yield insight on the issue of whether we ought to develop robots that can pass a physical Turing test. Each of these relationships has the potential to yield two different moral actors, a moral agent with a corresponding moral patient. Of those ten possibilities, we focus on six that have the most significance to our argument.

5.1 Developer as actor, Robot as patient

The Developer/Robot relationship is the beginning of any masquerade that may eventually take place. Note that in this LoA, the non-humanness of the robot is always observable for the human developer. Viewing the robot as a moral patient, we consider a developer who programs the robot in such a way that it cannot answer questions about its humanness. By doing this, the developer is incorporating a deception into the robot. Without claiming that the robot itself can or cannot be deceived, the robot has clearly become a vessel that implements a deception. The deployment of the robot becomes the action to analyze, because context will determine the moral significance of the action. For example, a robot deployed to masquerade as Abraham Lincoln in an exhibit at a museum would be a mild sort of "deception" that we would consider benign. But a robot developed to masquerade as a bank customer in order to rob a bank would be problematic. Note that in each of these cases it is the deployment action that is called in to question, it is not the act of developing such a robot

The Twitter bot developed by Nanis et al. appears to have been developed without the ability to reveal its non-humanness [5]. And, certainly the voice recognition system we encounter in customer service call centers does not have the ability to answer questions about its humanness. This seems perfectly reasonable, but seems to grow increasingly less so the more the robot has the potential to fool people. We seem to have this conundrum where the ways that software has been developed traditionally (it is not given knowledge of its non-humanness) becomes ethically problematic when that software happens to be software that may deceive people into believing the software is human.

The source of this unease does not arise in the current context, the relationship between the developer and the robot where the robot is the patient. The action under analysis here is the development of the robot. By focusing on the act of developing the robot, which may be just a piece of software, our analysis changes. Since the patient is the robot, IE asks whether any damage is done to software or any part of the infosphere because of this built-in deception. It is reasonable to conclude that none is done. Almost all of the software that has been used to build the infosphere has been developed in a similar way. Thus, there is no moral harm here. The robot is designed to create certain things in the infosphere, and assuming that it does so, the developer has not harmed the moral patient. However, as we shall see the next section, this act does produce harm in a different relationship.

5.2 Robot as actor, User as patient

In this relationship we consider two different cases. In both cases the robot is so human-like that it passes some reasonable physical Turing test. In the first case, the user knows that the robot is not human. In the second case, the user does not.

A robot with observable non-humanness does not mean that the user is constantly reminded of it. It is just that the user knows at some point in time about the non-humanness. It is fairly common that as we begin to use a particular piece of technology, we are tentative about relying too heavily on it. As the device becomes more reliable, we integrate it more deeply into our lives. The reliability has a second effect: we tend to spend less time critically considering the technology. Thus, a robot, through consistent high-level performance, can move to a state in the user's thinking that is quite similar to one in which the non-humanness of a robot is not an observable, even though that information initially was made available to the user. This may be true even for a robot that regularly reminds the user that it is not a human. The reminders may not register. Sherry Turkle gives us of the case of robot babies and pets given to the elderly, particularly those in nursing homes. At the end of the interaction period, she was unable to remove these robots as the users had become so attached to them. Said one, "She listens to me" [7].

Thus, regardless of whether a robot's non-humanness is formally an observable, we may ultimately be dealing with a case where a human user does not observe it. In this case the user believes incorrectly (or behaves as if) the robot is a human moral agent. The important distinction here is that in the case of a human moral agent, a user ultimately knows that the agent can be ascribed responsibility for any moral action. The robot (according to Floridi) can only be ascribed moral accountability. In the case that the robot engages in morally bad actions, there is clearly damage to the infosphere and according the Floridi the robot would be held accountable.

A second case is more interesting. That is, what if a robot successfully masquerading as a human engages in only morally neutral or good actions? Is there something inherently damaging to the infosphere in this case? One major concern is that information is not revealed to the user, a human, thus impacting that person's ability to make morally good decisions. The user might encounter a situation calling for moral action and come to the wrong conclusion because of the misunderstanding of the robot's situation. As a trivial example, consider a Twitter user following the Twitter bot described in section 5.1. If the user is particularly cognizant of the feelings of other Twitter users (note that not all Twitter users are human), the person may spend too much time trying to decide whether to stop following the Twitter bot for fear of hurting that "person's" feelings when, in fact, the Twitter bot is neither a person nor has feelings. A dramatic example might ensue if a fire breaks out in an office and a great deal of attention is paid to saving "people" who turn out to be robots; this is especially troubling in the case where a well-liked robot is saved and in the process human lives are lost [8].

The bottom line is that robots that masquerade as humans, regardless of whether that masquerade is announced, have the potential to impact the decision making processes of people. This asymmetric power that reflects back to the relationship in section 5.1 needs to be accounted for in design and development of robots that masquerade. And if it can be shown that the masquerade corrupts the infosphere, then according to Floridi, evil has occurred.

5.3 User as actor, Robot as patient

In IE, robots are clearly moral patients and, as such, deserve some initial, overridable moral respect. Thus, actions taken by a human user of a robot toward the robot are candidates for moral evaluation. First, we consider the case when the robot's non-humanness is observable.

When the robot's non-humanness is observable, a user's consideration of appropriate moral action can be more precisely considered. As with the case in section 5.2, a person may become accustomed to thinking about the robot in human terms, and be less likely to engage in a more critical analysis of a potential action toward the robot. Current moral thinking of many people is deeply influenced by ethical theories that have as an assumption that humans are the only moral patients. By observing the non-humanness of the robot, we might apply an ethical analysis that views the robot as a "thing" with no more ethical significance than our car or vacuum cleaner. At the very least, this suggests a need for better development of ethical intuition in users of robots toward robots.

More interesting questions arise when the actions of the user are either morally good or neutral and the user is unaware that the robot is not human. In any moral situation, a successful masquerade clearly causes the user to behave as if the patient were human. This might be useful in a training scenario, but it could be damaging when it happens in other contexts.

In this case, the person involved almost certainly employs ethical analysis that is based on a false assumption. The patient in this case is a robot, yet the user is giving it the moral respect that is due a human. Furthermore, there is no reason to assume that the conclusion reached with respect to the situation is even appropriate given that the patient in this case is not human. Ethical analysis fairs better when the observables include knowledge that the patient is a robot. The misinformation increases entropy, and encourages sub-optimal consequences.

5.4 Developer as actor, User as patient

In general, when a developer creates software, the output of the software is the moral action in which the user becomes a moral patient. In our specific scenario here, the moral action of the

developer is embodied in the actions of the robot. The robot is the means by which the developer acts upon the user. In addition to the issues addressed in [9], here we are concerned with the visibility of the non-humanness of the robot. We contend that some damage may be done when the robot embodies a deception about its humanness. As an information object, the quality of the information embodied in the robot is central to not only its "goodness," but its ability to promote goodness in the infosphere. Since the robot is NOT a human, any attempt to hide this fact (either from the robot or from observers of the robot) represents a moral action. In some contexts this action is morally acceptable. For example, in entertainment, scientific experiments, and educational displays such deceptions have positive overall outcomes. However, the presence of the deception (and its embodiment in the robot) should not be minimized in an IE analysis because the embodied deception is, by definition, creates an infosphere where morally unacceptable actions are easier to engage in.

A second concern arises from the complexity of design required for a robot to pass some sort of physical Turing test. Such a high level of complexity makes it difficult for the developer to understand or control the robot's actions. When the developer is focused on creating a robot that passes a physical Turning test, less attention might be paid to the morality of the actions the robot performs. In most software projects, features have varying levels of desirability and the developer typically focuses resources on those features that can be achieved efficiently. The complexity of a masquerade will, necessarily, require trade-offs in all functionalities.

When the goal is to develop a robot that passes a physical Turing test, adding a feature whereby the user of the robot is deterred from observing the robot as non-human is contradictory to the primary goal. This suggests that while developing a robot that can pass a physical Turing test is a significant technological achievement, it is ethically problematic. Such a goal can only be made ethically sound if users can somehow be made aware, and as we have argued, told again and again, that the device is a robot. Otherwise, the resulting misinformation corrupts the infosphere (IE analysis) and can have detrimental effects (visible in more traditional ethical analyses such as utilitarian, deontological and values).

5.5 Robot-User as actor, Others as patient

In this section and the next, we change our focus and consider the broader society as a moral patient. The observables that are significant here are more on a sociological level. We call this level of abstraction LoAS, where S is mnemonic for Society [10]. Again we are concerned with whether the non-humanness of the robot is an observable, this time at LoAS.

The infosphere already has reacted to the possibility that some robot-user actors are not pure human systems. There are times when the intention of the system make the ethical analysis clear. For example, in the case of robocalls that ask people to give personal information or wire money, the robot-user moral agents plays on the inability of the user to distinguish real human calls from robocalls. Here the robot-user is accountable and the moral responsibility for any damage lies with the human who chose to deploy the robot in this way.

The Captcha system touts itself as a public Turing test to tell humans and computers apart [11]. Essentially, others have assumed that non-humanness is being disguised on the web (and rightly so) and have developed a system to protect web site interfaces from this intentional deception. The interesting thing here is that the threat of non-humanness not being observable has had a negative impact on humankind. People are now obligated to spend time wading through Captcha's in order to obtain the types of services and products they are interested in. While this is of concern, there are clear arguments that when the robot-user is engaged in this sort of activity, the activity itself is clearly unethical. The robot-user system is trying to gain some sort of advantage over other users of that particular web service.

Contrast the Captcha system (which makes humans work to prove their humanness) with an earlier protocol that made the robots conform: robots.txt. In the early days of the web, a voluntary protocol was devised in which human web site administrators included a simply encoded text file named "robots.txt" in the top level directory for a website. Inside robots.txt, the site administrator listed what parts of the website the administrator did NOT want indexed by any "spiders" (softbots sent out by search engines to survey websites). In this case, the human administrators deposited an un-invitation, and softbots were assumed to obey the human wishes. The robots.txt protocol still exists, but many softbots now ignore the protocol [12].

The situation we are interested in is the impact the robot-user has when its intentions are mundane or even quite pure and for the benefit of others. Our assumption for this analysis is that others do not know and have no way of knowing whether the robot is indeed a robot, yet the user has full knowledge of this fact. For example, consider a robot used to care for demented patients in a hypothetical future facility designed to protect the patients. Perhaps the patients are too dangerous to use human caregivers, so instead direct care is given using human-like robots. If it is calming for the patients to think that the caregivers are human, then the deception may be ethically positive, despite the misinformation embodied in the robot caregivers.

5.6 Developer-Robot as actor, Others as patient

As an example, we return to the Twitter bot mentioned earlier. While the bot developed by Nanis, et al. did not pass any sort of rigorous Turing test, it did masquerade and cause real people to behave as if the bot were human [5]. Their bots obtained 62 followers on average. While it is difficult to ascertain how many of those followers were human (and in the context of IE this may not even be important), it stands to reason that many of them were. The purpose of their experiment was to ascertain whether a bot could foster human-human interaction. Their evidence suggests that the bots were quite successful at doing so. This higher level of connectivity among real people seems to be the sort of thing that Twitter users desire. Another example is Siri the personal assistant on the iPhone. In commercials on television, she is told to "take the night off", when she has completed her tasks. Users ask her to perform tasks as they would a human personal assistant.

Regardless of whether these bots pass a more sophisticated Turing test, we certainly can imagine the development of one that does. of the behavior of Siri and the Twitter bot is consistent with the expectations of users and, although deceptive, ought not be viewed as evil. The actions of these bots

help create an environment that establishes connections among users and does not degrade the infosphere.

6 CONCLUSIONS

In this paper, we have tried to deconstruct the complexity of machines masquerading as humans. In examining the roles of moral agent and moral patient we have defined relationships that serve to illustrate where the locus of moral responsibility and moral accountability lie in relationships where machines reveal their non-humanness and in those where they attempt to pass a physical Turing test. We have examined these relationships in the context of the infosphere and tried to answer the question of whether it is evil, according to Floridi's definition, for machines to masquerade as humans. The evidence is clear that when a machine masquerades, it influences the behavior or actions of people, not only toward the robot, but also toward other people. So even when the masquerade itself does not corrupt the infosphere, it changes the infosphere by making it more difficult for agents to make sound ethical decisions, increasing the chance for evil.

REFERENCES

[1] L. Floridi. Information Ethics. Forthcoming Oxford:Oxford University Press.

[2] L. Floridi and J.W. Sanders. (1999) Entropy as evil in information ethics. *Ethics and Politics* 1(2).

[3] T. Regan. The Case for Animal Rights.(1983). Berkeley and Los Angeles: University of California Press.

[4] K. Shrader-Frechette. Individualism, Holism, and Environmental Ethics. Ethics and the Environment , Vol. 1, No. 1 (Spring, 1996), pp. 55-69.

[5] M. Nanis, I. Pearce and T. Hwang. PacSocial: Field Test Report. http://pacsocial.com/files/pacsocial_field_test_report_2011-11-15.pdf, accessed 17 April 2012.

[6] Cambridge News. Meet Charles, the Robot That'll Make You Want to Smile. http://www.cambridge-news.co.uk/Education-and-Training/Universities/Meet-Charles-the-robot-thatll-make-you-want-to-smile-23032012.htm accessed 17 April 2012.

[7] S. Turkle, Alone Together: Why We Expect More from Technology and Less from Each Other (2011), New York: Basic Books.

[8] J. Bryson. (2010) Building Persons is a Choice. http://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Foerst09.html accessed 5 May 2012.

[9] F.S. Grodzinsky, K. Miller and M.J. Wolf. The ethics of designing artificial agents. Ethics and Information Technology, 10, 2-3 (September, 2008).

[10] M.J. Wolf, F.S. Grodzinsky and K. Miller. Artificial agents, cloud computing, and quantum computing: Applying Floridi's method of levels of abstraction. In Luciano Floridi's Philosophy of Technology: Critical Reflections. H. Demir, ed. Springer: 2012.

[11] CAPTCHA. http://www.captcha.net/ accessed 24 April 2012.

[12] KLOTH.NET. List of Bad Bots. http://www.kloth.net/internet/badbots.php accessed 3 May 2012.

Machines and the Moral Community

Erica L. Neely[1]

Abstract. A key distinction in ethics is between members and non-members of the moral community. Over time our notion of this community has expanded as we have moved from a rationality criterion to a sentience criterion for membership. I argue that a sentience criterion can be understood in terms of respecting the interests and autonomy of a being and thus may be extended to self-aware and/or autonomous machines. Such machines exhibit a concept of self and thus desires for the course of their own existence; this gives them basic moral standing, although elaborating the nature of their rights is complex. While not all machines display autonomy, those which do must be treated as members of the moral community; to ignore their claims to moral recognition is to repeat the errors of colonialism.

1 INTRODUCTION

A key distinction in ethics is between members and non-members of the moral community; this is the foundation for understanding how we may treat the entities we encounter in the world. Over time our notion of this community has expanded; those we take as non-members have thus changed, and the criteria used to make that distinction have also altered. Historically, as surveyed in [1, 2, 3], criteria such as intellect and rationality were used to separate white men from women and non-whites. Taken to be governed by emotion, these people were seen as moral inferiors, and thus deserving of lesser or no moral consideration.

Even upon conceding that rationality was not the exclusive preserve of white men, and thus including women and non-whites as members of the moral community, philosophers still generally denied moral standing to animals, as seen in [4, 5]; humans had the moral high ground of rationality and consciousness. However, the rationality criterion raises questions as to how rational a being must be to receive moral standing – there is a serious risk of excluding certain humans (such as infants) from the moral community which, as seen in [6], is unpalatable to many thinkers. Furthermore, our understanding of the biological similarities between humans and other animals makes it difficult to maintain a sharp distinction between them; various other animals seem to possess degrees of rationality and consciousness as well. This has resulted (in [6, 7, 8]) in a move to sentience as the criterion for moral standing: if something can feel pain, it is wrong to make it suffer unnecessarily.[3]

This is a large expansion to the moral community, yet of course many things continue to lack moral standing; an object such as a table or chair is not a member of the moral community, for instance. Unless the object belongs to someone else, I can do what I wish to it; the only kind of moral harm that can be caused in this situation is harm to a person or persons who have a claim to that object.[4] As such, there is currently a strong ethical divide between living beings and what we see as created things. This has serious implications for the ethical issues pertaining to intelligent machines, since for many there is a strong temptation to classify these machines as objects and thus not deserving of any moral standing. I will argue that this is incorrect and that certain kinds of machines are, in fact, members of the moral community.

2 ETHICS AND THE PREVENTION OF HARM

When conversing with people, one informal objection that frequently occurs to granting moral standing to a machine is the claim that you cannot "hurt" a machine. In essence, this is an internalization (and over-simplification) of the sentience criterion for moral standing. Ethics is often taken to involve the prevention of harm – if something cannot be harmed, then many evidence a certain reluctance to offer moral standing to the thing in question.

For humans, the harm generally involves some kind of pain. However, the ability to feel physical pain cannot be the only criterion for membership in the moral community. Consider a person with congenital analgesia, i.e., one who is unable to register physical pain. It would surely still be wrong to step on his foot intentionally and without his permission. This is not because the action caused pain (since, by design, it does not). Instead, the wrongness stems from two key points. First, the action could cause damage, even if it does not cause pain. Second, since we have specified that the person does not give permission for the action, deliberately stepping on his foot violates his desire to remain unmolested.

I will briefly sketch how these two characteristics suffice to render the action unethical under the auspices of most major ethical theories. Specifically, utilitarianism, deontological

[1] Dept. of Philosophy and Religion, Ohio Northern Univ., Ada, OH, 45810, USA. Email: e-neely@onu.edu.

[2] Obviously there is clarification required to specify what constitutes unnecessary suffering and exactly how much moral standing animals have. However, sentience suffices to give them a foot in the door of the moral community, so to speak.

[3] Obviously there is clarification required to specify what constitutes unnecessary suffering and exactly how much moral standing animals have. However, sentience suffices to give them a foot in the door of the moral community, so to speak.

[4] The ownership of an object could be the community as a whole, such as with public art installations. If someone were to destroy the Vietnam Veteran's Memorial, one could argue that it would cause harm to the public (which has a claim on the memorial) and is thus morally wrong. It would be odd to say that you had morally wronged the monument itself, however.

ethics, virtue ethics, the ethics of care, and contractarianism will generally condemn the action. While space does not permit a full elaboration of the arguments with regard to each theory, the outline should serve to highlight the following point: actions which either cause harm or simply ignore the wishes of the person being acted upon are unethical even if said actions do not cause pain. Moreover, I shall argue that there is an underlying theme that unifies the ethical condemnation expressed by these diverse theories.

First, consider utilitarianism. Jeremy Bentham appeals to his hedonic calculus in order to determine the rightness of an action, as outlined in [8]. Although in our situation there is no physical pain sensation felt by the victim, Bentham allows for mental or emotional pain as well. It thus seems likely that, in general, a person's distress in having their wishes ignored will outweigh the sadistic pleasure of the tormentor. There might be some odd instances where this is not the case, particularly if the person has little desire to live or cannot comprehend what has been done to him. The more general notion of disutility raised by John Stuart Mill in [9] can accommodate many of these outliers, since there is harm being caused to the person even if he cannot register the physical sensation of pain.[5] As such, disutility is being generated, and will usually outweigh any sadistic utility generated by the action.

The ethical argument condemning the action is extremely simply for deontological ethics, since the tormentor is clearly violating the second formulation of the Categorical Imperative. As developed by Kant in [4], an action is only ethical if it treats humanity (by which he means any rational being) as an end rather than as a means. One is thus not permitted to use rational beings simply as a tool for achieving one's own purpose. By ignoring the wishes of the person in question, the tormentor is using that person as a means for his own pleasure; furthermore, since there are no extenuating circumstances specified, such as the action's somehow being for the long-term good of the victim, the action does not proceed from a good will. It is thus not a permissible action for a deontological ethicist.

Both virtue ethics, such as Aristotle espouses in [11], and the caring ethics developed by Carol Gilligan and Nel Noddings in [12, 13], share a similarity in their reasons for condemning the action. Specifically, the tormentor's lack of sympathy for the victim, which is evidenced by disregarding his wishes entirely, demonstrates a lack of benevolence and care. The tormentor has chosen a selfish pursuit of pleasure at the expense of others; he does not care if he is causing harm, nor does he care what his victim desires. Again, since there are no mitigating factors, it cannot be that some other concern or virtue is outweighing this one; the action simply displays a lack of care and compassion.

To round out our survey of ethical theories, based on [14], a contractarian such as John Rawls would denounce the action because the person in question did not consent to it. More accurately, no rational person would consent to such an action, as there is no advantage to him; giving up his right to remain unmolested and allowing others simply to ignore his desires is irrational in the absence of any expected benefit. While there is debate over the degree of risk to which a rational person would consent, this does not seem to present a serious problem in our case. Unless the proportion of sadists in the general population is considerably higher than I expect, there is almost certainly a greater prospect of being tormented (and hence more harm generated by such actions) than there is of one being a sadist (and thus deriving benefit from that action.) The chance of satisfaction provided to the sadist is simply not sufficient to motivate consent to this infringement of individual liberty.

For all of these ethical theories note that the explanations of wrongness are tied to the more general idea of having interests, not to that of feeling physical pain. The notion of harm in question, therefore, moves beyond physical pain and hinges on the idea of respecting the integrity and autonomy of the individual. The possibility of the action's causing damage, even if it does not cause pain, raises the idea of bodily integrity. At a minimum, beings have an interest in retaining sufficient bodily integrity for continued existence; anything which damages one's body threatens this interest. This interest can certainly be outweighed by other factors – I may consent to having my appendix removed because that particular violation of bodily integrity actually promotes my continuation under certain circumstances. Frequently in medicine we consent to actions which are extremely damaging to our bodies (such as chemotherapy) if the alternatives are worse.

Clearly it is possible to overstate the commitment to bodily integrity, since we consent to small violations of it on a regular basis. Most people trim their fingernails or their hair or will pick open the occasional scab; they are unlikely to see those actions as presenting any serious threat to continued existence. Hence one might argue that a minor harm, such as stepping on a person's foot, cannot truly be objected to on this basis. Indeed, I believe that the emphasis on bodily integrity dovetails with the desire to remain unmolested mentioned above; together they highlight the fact that I have certain wishes about the shape of my life.

By ignoring the person's desire not to be trod upon, the aggressor's action violates his autonomy. In much of ethics, autonomy is emphasized as an important good.[6] To cast it aside for no reason other than to satisfy one's own sadistic desires is to jeopardize the interest of the injured person in governing the course of his own life. Such an action may not cause physical pain, but it clearly causes harm to that person – it treats him as incapable or unworthy of directing his own actions, and views his desires as irrelevant and something that may simply be ignored. Although it is clear that sometimes a person's desires must, ethically, be overridden, we surely cannot ignore another's wishes completely.[7]

[5] There are familiar objections to utilitarianism which raise the point that if our tormentor is sufficiently sadistic, his glee in causing harm might outweigh the harm actually done. I view this as a weakness of utilitarianism, rather than an indicator that the action is morally permissible, however space does not permit a full consideration of the merits of the theory. Since I believe these cases to be rare, it should suffice to say that in most cases a utilitarian will deem the action unethical.

[6] We see this both in [4] with the view of rational beings as ends-in-themselves and in [10] with the emphasis on individual liberty.

[7] While I will not rehearse the arguments for each ethical theory in detail again, note that ignoring a person's desires for his life will fail to calculate the utility/disutility generated by particular actions, will treat the person as a means to an end, is certainly not something rational people are likely to consent to from behind a veil of ignorance, and demonstrates a lack of care, compassion, and benevolence. Thus none of these ethical theories will condone simply ignoring the desires of a person.

Hence while sentience certainly leads to having interests, it is not necessary for them: the property of consciousness or self-awareness will also suffice.[8] Once a being is self-aware, it can desire continuation and formulate ideas about how to live its life. It is possible to harm such a being by ignoring or thwarting those desires; one should not act against the being's wishes, therefore, without some overriding reason. The requirement of such a reason, however, is equivalent to granting the being at least minimal moral standing; one does not need to have a reason to destroy a chair, but one must provide such a reason to destroy a human. This holds true for intelligent machines just as much as for a person with congenital analgesia; they both have interests and desires, hence they both have basic moral standing.

3 SELF-AWARENESS AND AUTONOMY

Thus far our argument has reached the conclusion that self-aware machines have moral standing. In general, the question of moral standing for machines is raised in the context of artificial intelligence – would an intelligent machine have moral standing? To provide an answer to this general question, we must ask whether we can assume that intelligent machines are self-aware. If so, we have addressed the moral standing of all intelligent machines; if not, then further work is necessary to clarify the status of the remaining machines.

To respond to this, we must consider what is meant by an intelligent machine. Shane Legg and Marcus Hutter have, in [16, 17, 18], gathered many of our informal definitions of intelligence and used them to devise a working account of machine intelligence. Informally, their definition of intelligence in [18] is "Intelligence measures an agent's ability to achieve goals in a wide range of environments."[9] One key question that emerges from this definition is who determines the goals of the agent. There are two possibilities: one, the agent's goals are always determined by an outside source or, two, the agent sometimes determines its own goals.

Consider the case where the agent's goals are always established by an outside source. In this case, the goals are communicated to the agent in some fashion, and the agent simply uses its resources to accomplish whatever goals it has been given. Such an agent lacks any kind of autonomy. Since the agent lacks self-awareness and lacks the ability to formulate goals for itself, the argument for moral standing does not apply; it will not have a desire for continuation or any wishes as to how to live its life. As such, it seems to be in the same category with chairs and tables mentioned above and lacks moral standing; it is not clear how one could harm or benefit such an entity.[10]

Contrariwise, consider the case where the agent is capable of determining its own goals, at least some of the time. In this case, the agent is expressing a basic capacity for autonomy – it is capable of directing (at least some of) its own actions. As such, it exhibits a basic level of self-awareness; making choices concerning its future goals implies that it possesses the concept of self – setting goals for oneself requires awareness of that self. While the choices may be influenced by the programming of the machine, human choices are also influenced by upbringing, societal pressure, brain chemistry, and so forth. Since moral theorizing generally views human autonomy as worth preserving despite these factors, machine autonomy likewise has worth.[11]

One point worth noting is that moral questions are not black-and white; both autonomy and moral standing exist on a continuum. The more autonomous the machine, the more duty we will have to respect its wishes; the less autonomy, the more we are permitted to act as its guardian. This is akin to how we treat children and the severely mentally disabled; they are not viewed as capable of making decisions in as many areas as fully-functioning adults, hence we do not see their desires as binding to the same extent. They still have moral standing, of course, in that it is wrong to harm them without just cause. Nevertheless, they are not granted as much governance over the course of their own lives, and we do not view overriding their wishes as comparable to overriding the wishes of other adults. In a similar fashion, a machine with greater autonomy likely has more claim on us to respect that autonomy, and it will be a greater moral fault if we ignore its wishes.

In summary, I believe that autonomy leads to self-awareness. My previous argument thus fails to apply only to machines which both lack self-awareness and which are not capable of setting their own goals. Such machines do seem to lack moral standing because they have no self-concept and thus no way to desire existence nor to have goals for that existence. Determining whether and to what extent a machine is autonomous will likely be difficult, however, and those who oppose granting moral standing to machines might well use this as an excuse to deny their moral worth. This is a dangerous move to make, though, since the long-standing philosophical dilemma of other minds demonstrates that it is also hard to ensure that other people have minds and are not cleverly programmed automata.

In general, it seems wise to err on the side of caution – if something acts sufficiently like me in a wide range of situations, then I should extend moral standing to it.[12] There is little moral fault in being overly generous and extending rights to machines which are not autonomous; there is huge moral fault in being

---

[8] An interesting discussion of the connection between self-awareness and moral standing (or personhood, as she puts it) can be found in [15]. Note also that since some philosophers argue that sentience leads to a basic kind of self-awareness, the sentience account is compatible with my conclusion. See [7] for a discussion of self-awareness and moral standing.

[9] They provide a formal definition in [18], however space does not permit the detailed exposition required to fully explicate this definition.

[10] Presumably the machine is not sentient, or we could have had a much shorter argument for moral standing; as such, it cannot gain moral rights through an appeal to sentience. One might try to argue that such a being has rationality and thus, on some views of morality at least, must be granted moral standing. I am not convinced this is the case; while Kant sees morality as shared by rational beings, he makes it clear in [4] that the kinds of beings he is discussing have a will – the machines, as I have described them, do not. In general, I believe that the rationality criterion for moral standing is more complex than simple intelligence, and machines with bare intelligence will likely not satisfy it.

[11] The view that autonomy implies self-awareness is likely unnecessary to grant machines moral standing, since autonomy itself might well be sufficient. If we view autonomy as a good, then the fact that such machines exhibit autonomy suffices to grant them at least some consideration. We may place limits on the expression of their autonomy, just as we do for people, but we likely could not simply ignore it.

[12] Think of this as the moral equivalent of the Turing Test. This argument is used in [6] to argue for our assumptions of sentience both in other people and in animals.

overly conservative about granting them moral standing. The most serious objection to extending moral standing too widely with respect to machines is that we might unjustly limit the rights of the creators or purported owners of said machines: if, in fact, those machines are not autonomous or self-aware, then we have denied the property claims of their owners. However, when weighing rights, the risk of losing a piece of property is trivial compared to denying moral standing to a being. As such, ethically speaking, the duty seems clear.

4 MORAL STANDING AND RIGHTS

The moral standing of intelligent autonomous machines is thus a natural extension of the sentience-based criteria for moral standing.[13] Intelligent, self-aware machines are beings which have interests and therefore have the capacity to be harmed. Hence, they have at a minimum moral claims to self-preservation and autonomy, subject to the usual limits necessary to guarantee the rights of other community members.

It is difficult to specify what moral entitlements said machines will have until we know the nature of those machines. Since machines are physically different than humans, some rights will need to be "translated." A basic human right to sustenance will take a rather different form for machines, for instance, since they are unlikely to need food and water; they might well have an equivalent need for access to electricity, however. Similarly, just as humans have a need for medical care of various kinds, intelligent machines might require certain kinds of preventative maintenance or repairs.

Moving beyond basic needs for survival, it is interesting to consider rights on a larger socio-political scale, such as the basic human rights espoused in [19]. It is not immediately obvious how some of these will be handled, such as the claim that everyone has the right to a nationality. For humans, we determine that nationality based on the arbitrary criterion of birthplace (or parental nationality); it is then theoretically possible to change affiliation by undergoing certain processes.[14] One might suggest, therefore, that we could grant machines a starting nationality based on where they were first "switched on."

However, this answer is further complicated if we extend moral consideration from machines to entities which are not embodied and have only a virtual presence.[15] My argument could fairly easily be expanded to include these entities, since they could also display autonomy or self-awareness. The main adjustment needed is to devise an understanding of what their existence consists in, since it cannot be linked easily to embodiment. We do not have much experience with non-corporeal existence, hence there are metaphysical questions that would need to be addressed before we can determine how best to understand the rights of these beings.[16] Yet clearly they will complicate questions such as nationality: how do you attach a nationality to something which doesn't have a physical presence per se? Is there any benefit to trying to do so? What would it mean if they existed outside the current borders of our political structures?

Metaphysical questions are not limited to virtual entities, nor is the issue of nationality the only right which raises questions. Even the basic rights of sustenance and security – things which contribute to a being's continuation – raise issues concerning what it means for these machines to exist or to cease to exist. In order to have a right self-preservation, we must understand what that means with respect to these beings. Thus while it is clear that self-aware machines have moral standing, it is much more difficult to say exactly what that standing grants them.

5 CONCLUSION

I have argued that the properties of autonomy and self-awareness are sufficient for granting an entity moral standing; if a being is capable of desiring its own continuance and of forming wishes about its future, then we have some prima facie obligation to respect those desires. As with any other member of the moral community, those rights may be overridden. However, their wishes cannot simply be ignored – that is unethical. Determining the details of machines' moral standing is difficult, particularly since the relevant machines do not yet exist; some moral theorizing may need to wait until we have a better idea of what they are like.

The battle for recognition of their rights will not be easy. We do not acknowledge the claims of others readily, even when the only difference between ourselves and those people is skin colour or gender; this difficulty will be magnified for intelligent machines. One key problem is the need for others to acknowledge the autonomy and/or self-awareness of those machines. Philosophers have been arguing over the problem of other minds for millennia with respect to humans; the problem will likely magnify for machines, since we do not have a clear set of criteria that all will accept as sufficient for consciousness or autonomy.[17]

The biggest obstacle in the way of acknowledging moral standing for machines, however, will likely not be philosophical – it will be pragmatic. We depend upon machines to do many tasks for us, and we do not currently pay machines or worry about their needs (beyond perhaps basic maintenance). One of the rights enshrined in [19] is the right to remuneration for work,

[13] It is probably possible also to defend granting moral standing to such machines on a rationality-based understanding of the moral community, however as I am sympathetic to the criticisms of such theories, I shall not attempt to do so here.

[14] I say "theoretically" since, in practice, the change of nationality is fairly difficult; most people are pragmatically limited to the nationality of their birth, regardless of having a human right to change it.

[15] One could object that, speaking precisely, such entities will likely not be wholly virtual. Rather, they may well require the existence of physical objects in the same way that computer viruses require physical machines on which to reside; their existence is not independent of physical objects. However, the identity of the virus or the machine is quite distinct from the physical object(s) they depend on in a way unlike our experience of other identities; if they are embodied, it is in a very different sense than we currently understand.

[16] For instance, the human sense of self is frequently tied to our physical embodiment, which makes it hard for us to comprehend what sort of identity such a being can have. It seems clear to me that such a being should be able to have an identity, however, since we acknowledge that one can have a drastic alteration in physical characteristics (such as an amputation) while retaining a sense of self. As such, a specific embodied form does not seem a requirement for identity and self-awareness.

[17] See [20, 21] for Floridi's presentation of this conundrum and an attempt to devise a test for self-consciousness in response.

meaning that the financial pressure to avoid recognizing any moral standing for intelligent machines will likely rival the push to avoid acknowledging African-Americans as full persons in the Confederate South. However, we cannot ethically deny someone moral standing simply because it is convenient. The Western world has repeatedly done this in our history of colonialism, and we would be wise not to make the mistake again. The time to start thinking about these issues is now, before we are quite at the position of having such beings to contend with. If we do not face these questions as a society, we will likely perpetrate injustices on many who, in fact, deserve to be regarded as members of the moral community.

REFERENCES

[1] C. Mills. *The Racial Contract*. Cornell UP, Ithaca, USA. (1999)

[2] N. Zack. *The Philosophy of Science and Race*. Routledge, New York, USA. (2002)

[3] L. Code. Is the Sex of the Knower Epistemologically Significant? In: *What Can She Know?: Feminist Theory and the Construction of Knowledge*. Cornell UP, Ithaca, USA. (1991)

[4] I. Kant. Groundwork of The Metaphysics of Morals. In: *Practical Philosophy*. M. J. Gregor (Ed.) Cambridge UP, Cambridge, U.K. (1996)

[5] R. Scruton. *Animal Rights and Wrongs.* Continuum, London, U.K. (2006)

[6] P. Singer. *Animal Liberation*. Ecco, USA. (2002)

[7] A. Taylor. Nasty, brutish, and short: The illiberal intuition that animals don't count. *The Journal of Value Inquiry*, 30: 265-277. (1996)

[8] J. Bentham. *An Introduction to the Principles of Morals and Legislation*. J.H. Burns and H.L.A. Hart (Eds.) Oxford UP, New York, USA. (1996)

[9] J. S. Mill. Utilitarianism. In: *On Liberty and Utilitarianism*. Bantam, NY, USA. (1993)

[10] J. S. Mill. On Liberty. In: *On Liberty and Utilitarianism*. Bantam, NY, USA. (1993)

[11] Aristotle. *Nicomachean Ethics*. R. Crisp (Trans.) Cambridge UP, Cambridge, U.K. (2000)

[12] C. Gilligan. *In A Different Voice.* Harvard UP, Cambridge, USA. (1982)

[13] N. Noddings. *Caring: A Feminine Approach to Ethics and Moral Education*. University of California, Berkeley, USA. (1984)

[14] J. Rawls. *A Theory of Justice*. Belknap, USA. (2005)

[15] M. A. Warren. On the Moral and Legal Status of Abortion. *Monist*, 57: 43-61. (1973)

[16] S. Legg and M. Hutter. A Collection of Definitions of Intelligence. In: *Proc. 1st Annual artificial general intelligence workshop. B. Goertzel (Ed.) (2006)*

[17] S. Legg and M. Hutter. A Formal Measure of Machine Intelligence. In: *Proc. Annual machine learning conference of Belgium and The Netherlands*. Ghent, Belgium. (2006)

[18] S. Legg and M. Hutter. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, 17: 391-444. (2007)

[19] United Nations. *The Universal Declaration of Human Rights.* **http://www.un.org/en/documents/udhr/** (1948)

[20] L. Floridi. Consciousness, Agents and the Knowledge Game. *Minds and Machines*, 15: 415-444. (2005)

[21] S. Bringsjord. Meeting Floridi's Challenge to Artificial Intelligence from the Knowledge-Game Test for Self-Consciousness. *Metaphilosophy*, 41: 292-312. (2010)

Who cares about robots? A phenomenological approach to the moral status of autonomous intelligent machines

Mark Coeckelbergh[1]

Abstract. This paper address the problem of how to approach the question of moral status of autonomous intelligent machines, in particular intelligent autonomous robots. Inspired by phenomenological and hermeneutical philosophical traditions, (1) it proposes a shift in epistemology of robotics (from objectivism to phenomenology, from object to subject-object relations, from the individual to the social and cultural, and from status to change) and (2) analyses what it is we care about when we care about the moral status robots. This give us an approach that implies epistemological anthropocentrism, but not necessarily moral anthropocentrism; whether or not we want to include robots in our world depends on the kind of moral and social relations that emerge between humans and other entities.

1 THE OBJECTIVIST/REALIST APPROACH TO MORAL STATUS

Thinking about the issue of moral status of robots is interesting since, among other things, it demands that we clarify what it means to be a moral agent and what morality is. This is not only relevant to thinking about robots but also to thinking about humans.

A common approach to the moral status of intelligent machines is based on an assessment of features of the 'mind' of the machine: Does the machine have consciousness? What kind of intelligence does it have? Does it have emotions? Can it feel pain? For instance, Levy has argued that robots should get rights if they are conscious [1] and according to what Torrance calls the 'Organic' view, entities with 'organic' characteristics such as sentience have intrinsic moral status [2].

This line of argumentation is similar to the one used in the past (and in the present) to ascribe moral consideration to women or to animals. For instance, it has been argued that animals who can feel pain should be taken into moral consideration. Singer has famously argued that if animals are sentient and can suffer, we should grant them equal moral consideration [3][4]. Thus, thinking about moral status in an objectivist and realist way seems to be the best way to ensure that we can 'emancipate' other entities.

However, this objectivist (or realist) approach to moral status incurs a number of problems, which will lead me to consider an alternative approach based on a different, more relational moral epistemology.

2 TWO PROBLEMS

1. First problem: Threshold full moral agency is too high

One of the problems with the objectivist approach is that it is unable to help us with ethical problems concerning current intelligent autonomous robots, who do not satisfy the demanding criteria for full moral agency. Yet we also sense that their moral status is 'more' than that of a 'a mere machine'. How can we account for that kind of experience – and indeed that kind of treatment and interactions?

One solution to this problem is to lower the criteria for moral agency, or to distinguish several levels of moral agency. Consider for instance Allen and Wallach's proposal to focus on what they call "functional morality" rather than full morality: robots that "monitor and regulate their behaviour in light of the harms their actions may cause or the duties they may neglect" [5]. However, we do not have such robots either. Moreover, such robots with a built-in rational morality may be even dangerous if they pretend to be moral agents but actually are not, and if they take autonomous decisions without emotions. As I have argued before, we don't want 'psychopath' robots [6]. But even if we do not want and do not build 'moral machines', how can we still say something about the moral significance and moral status of current advanced robots, given that they lack morally relevant properties?

2. Second problem: Sceptic response to the possibility of gaining knowledge about ontological and moral status

Another problem is that it seems difficult to agree on the criteria and on the question whether or not a particular robot meets the criteria (or to what degree). In order to solve the threshold problem, we can ascribe a lower-than-human moral status to advanced robots (e.g. based on their degree of autonomy), and then say that this gives them also some degree of 'responsibility' or at least gives *us* moral responsibility. For example, we might want to limit its autonomy and 'keep a man in the loop'. But can we be sure about the ontological status of the robot, on which its moral status is supposed to be based? *Who* says that this particular robot has this particular moral and ontological status, e.g. to a higher or lower degree than another one? The scientist? The user? The philosopher? They may disagree about the properties of the robot and about whether those properties are morally relevant. Is there an 'objective' truth about the robot's ontological and moral status?

3 SOLUTION: FROM OBJECT TO SUBJECT-OBJECT RELATION

[1] Dept. of Philosophy, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: m.coeckelbergh@utwente.nl

So far, criteria for moral status are presented as 'objective' criteria. The approach I want to defend here, however, performs the following philosophical operation: it shifts our moral attention from the object to the *subject* and its relation to the object: from the objective (actual or future) properties of the robot to the way we perceive the robot. Who says what about this robot? What does it mean to say that the robot seems to us 'more than a machine'? Who cares about this?

This approach to moral status, which is influenced by the phenomenological and hermeneutical tradition (in particular Heidegger), attends to how the entity appears to us, how we give meaning to it, how we construct it when we talk about it. We cannot perceive the entity from a point of 'nowhere'; it is already part of a world, that is, it its meaning already emerges 'before' philosophers can reason about its moral status. This does not mean that the entity's moral status is 'purely subjective' if this means *depending on individual judgement and decision*. Rather, its meaning emerges from within a particular social context, in which there are already relations between humans and between humans and non-humans: relations that shape individual ways of seeing others and other entities. Our societies create specific moral *habits* that usually constrain our behaviour and our thinking.

This 'social-relational' approach to moral status [7][8] encourages us to try to understand different ways in which machines appear to us: sometimes as 'mere machines', sometimes as quasi-animals or quasi-humans. Moreover, it also shows us that there is no robot-in-itself (compare: thing-in-itself) and no 'pure' moral status, independent of the observer. We, humans say something about robots. And this point of view is not morally neutral. We have personal experiences and interests and we perceive in ways that are promoted by the particular society and culture we live in.

For instance, when working in their lab scientists and engineers will usually see the robot as a system. They know it in terms of their underlying forms, processes and principles (science) and as a collection of engineering solutions. Those who sell and use robots, however, may present and perceive it in a very different way, for example as a sex companion, a household slave or a killer machine. And there are differences between societies and cultures, for example between 'the West' and China or Japan (although modernity is rather pervasive and there might be more we share than we think). All these ways of seeing have moral consequences for the way we use and treat robots, and are connected to particular interests (e.g. in health or in military affairs) that are highly morally relevant.

This does not imply that all ways of seeing are equally good; it only means that ethics of robotics cannot take for granted the one way of seeing (e.g. the scientific one) and must explore different possibilities of perceiving and relating to robots and their moral consequences.

It also means that now we can say more about the moral status of intelligent machines that are said to lack consciousness or sentience. Whereas the objectivist/realist approach has no resources for acknowledging and discussing the moral significance of these robots, the phenomenological, social-relational approach can be sensitive to the 'non-scientific' ways people may perceive the robot and relate to the robot. In particular, they may relate to it *as if* it is more-than-a-machine. In a particular situation and with regard to particular people and robots this may or may not be problematic; the point is that now we have created space for a more rich evaluation of the robot-in-relation rather than an analysis of the robot-in-itself. It means that we have create space for an ethics of robotics that is no longer concerned with the 'moral status' (narrowly conceived) of robots or indeed of humans, but that tries to understand and evaluate the precise form of the human-robot relations that emerge in particular personal, social, and cultural contexts – including how humans experience, influence, and give meaning to these relations.

4 WHAT WE (SHOULD) CARE ABOUT

This is what we really care about when we ask the question about the moral status of robots: whatever their 'objective' physical and metaphysical status in the order of things, we care about what is going on between humans and robots and what *should* go on at between them. This is always judged from a human point of view. We care about what happens to our lives and our societies. If we want to understand moral status, we have to study this human and social aspect. Who cares about what and why? Why do these people or this organisation use these particular robots? What do they achieve with it? How are the robots experienced by those who use them or by those who are affected by them, and what does this experience mean to them? Unless robots became so advanced that *they* would start asking these questions about *us* (which is at least very unlikely if not impossible in principle), moral inquiry should start from the human side of things.

But should it also end there? Although this approach implies that moral status ascription and, more generally, robot ethics, is *epistemologically* anthropocentric, it is not necessarily *morally* anthropocentric. We cannot change the former (we 'have' to start from the human subject), although the latter might change. This would require a change in *human* consciousness. Perhaps this change is already going on: we already perceive animals in a different way than before (and sometimes even treat them differently) and this could also happen with entities such as robots and with other entities we now consider to be 'things', 'tools', or 'machines'. Indeed, we might want to expand our moral concern to particular entities or even to all entities. But if this moral change is happening at all, then in order to recognize it, attend to it, understand it, evaluate it, and influence it, it is not sufficient to discuss about the properties of the entities. Instead, it is important to study and evaluate not so much *robots* as stand-alone objects but also and especially *humans* and society and culture they live in: how we experience robots and how we cope with them as social and moral beings – including how our relation to robots is changing, how it might change in the future, and where this techno-human and techno-social *should* be going. After all, technology – including robotics technology – is about *us* and what *we* care about.

5 CONCLUSION

What matters to the 'moral status' of robots is not so much what robots 'are', but what they mean to us as social and moral beings who relate to them, since this influences whether we want to *explicitly* include them in our moral-social world, for example by means of ascribing 'moral status' to them.

This social-relational approach to moral status, which is at least epistemologically anthropocentric, does not preclude the possibility of expanding our moral concern to other entities, but advises us that if we care about entities like robots and perhaps want to 'give' them some degree of 'moral status', we better first try to understand how humans experience and cope with robots, and how they might do so in the future. If robots will ever get any 'moral status' at all, it will not be based on objectivist/realist justifications but on acknowledging and changing moral relations between humans and robots.

REFERENCES

[1] D. Levy. The Ethical Treatment of Artificially Conscious Robots. *International Journal of Social Robotics*, 1(3): 209-216 (2009).

[2] S. Torrance. Ethics and Consciousness in Artificial Agents. *AI & Society*, 22: 495-521 (2008).

[3] P. Singer, *Animal Liberation*, Random House, New York, U.S., 1975.

[4] P. Singer, *Practical Ethics, 2nd ed.*, Cambridge University Press, Cambridge, UK, 1993.

[5] W. Wallach and C. Allen, *Moral Machines*, Oxford University Press, Oxford, UK 2009, p. 16.

[6] M. Coeckelbergh. Moral Appearances: Emotions, Robots, and Human Morality. *Ethics and Information Technology*, 12(3): 235-241 (2010).

[7] M. Coeckelbergh. Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology*, 12(3): 209-221 (2010).

[8] M. Coeckelbergh. *Growing Moral Relations: Critique of Moral Status Ascription*, Palgrave Macmillan, Basingstoke/New York, UK and US, 2012. (forthcoming)

A Vindication of the Rights of Machines

David J. Gunkel[1]

Abstract This paper responds to the machine question in the affirmative, arguing that machines, like robots, AI, and other autonomous systems, can no longer be legitimately excluded from moral consideration. The demonstration of this thesis proceeds in three parts. The first and second parts approach the subject by investigating the two constitutive components of the ethical relationship—moral agency and patiency. And in the process, they each demonstrate failure. This occurs not because the machine is somehow unable to achieve what is considered necessary to be considered a moral agent or patient but because the standard characterization of agency and patiency already fail to accommodate not just machines but also those entities who are currently regarded as being moral subjects. The third part responds to this systemic failure by formulating an approach to ethics that is oriented and situated otherwise. This alternative proposes an ethics that is based not on some prior *discovery* concerning the ontological status of others but the product of a *decision* that responds to and is able to be responsible for others and other kinds of otherness.

1. INTRODUCTION

One of the enduring concerns of moral philosophy is determining who or what is deserving of ethical consideration. Although initially limited to "other men," the practice of ethics has developed in such a way that it continually challenges its own restrictions and comes to encompass what had been previously excluded individuals and groups—foreigners, women, animals, and even the environment. "In the history of the United States," Susan Anderson has argued, "gradually more and more beings have been granted the same rights that others possessed and we've become a more ethical society as a result. Ethicists are currently struggling with the question of whether at least some higher animals should have rights, and the status of human fetuses has been debated as well. On the horizon looms the question of whether intelligent machines should have moral standing." [1] The following responds to this final question—what we might call the "machine question" in ethics—in the affirmative, arguing that machines, like robots, AI, and other autonomous systems, can no longer and perhaps never really could be excluded from moral consideration. Toward that end, this paper advances another "vindication discourse," following in a tradition that begins with Mary Wollstonecraft's *A Vindication of the Rights of Men* (1790) succeeded two years later by *A Vindication of the Rights of Woman* and Thomas Taylor's intentionally sarcastic yet remarkably influential response *A Vindication of the Rights of Brutes*.[2]

Although informed by and following in the tradition of these vindication discourses, or what Peter Singer has also called a "liberation movement" [3], the argument presented here will employ something of an unexpected approach and procedure. Arguments for the vindication of the rights of previously excluded others typically proceed by 1) defining or characterizing the criteria for moral considerability or what Thomas Birch calls the conditions for membership in "the club of *consideranda*," [4] and 2) demonstrating that some previously excluded entity or group of entities are in fact capable of achieving a threshold level for inclusion in this community of moral subjects. "The question of considerability has been cast," as Birch explains, "and is still widely understood, in terms of a need for necessary and sufficient conditions which mandate practical respect for whomever or what ever fulfills them." [4] The vindication of the rights of machines, however, will proceed otherwise. Instead of demonstrating that machines or at least one representative machine is able to achieve the necessary and sufficient conditions for moral standing (however that might come to be defined, characterized, and justified) the following both contests this procedure and demonstrates the opposite, showing how the very criteria that have been used to decide the question of moral considerability necessarily fail in the first place. Consequently, the vindication of the rights of machines will not, as one might have initially expected, concern some recent or future success in technology nor will it entail a description of or demonstration with a particular artifact; it will instead investigate a fundamental failure in the procedures of moral philosophy itself—a failure that renders exclusion of the machine both questionable and morally suspect.

2. MORAL AGENCY

Questions concerning moral standing typically begin by addressing agency. The decision to begin with this subject is not accidental, provisional, or capricious. It is dictated and prescribed by the history of moral philosophy, which has traditionally privileged agency and the figure of the moral agent in both theory and practice. As Luciano Floridi explains, moral philosophy, from the time of the ancient Greeks through the modern era and beyond, has been almost exclusively agent-oriented. "Virtue ethics, and Greek

[1] Department of Communication, Northern Illinois University, DeKalb, Illinois 60631, USA Email: dgunkel@niu.edu

[2] What is presented here in the form of a "vindication discourse" is an abbreviated version of an argument that is developed in greater detail and analytical depth in *The Machine Question: Critical Perspectives on AI, Robots and Ethics*. [2]

philosophy more generally," Floridi writes, "concentrates its attention on the moral nature and development of the individual agent who performs the action. It can therefore be properly described as an agent-oriented, 'subjective ethics.'" [5] Modern developments, although shifting the focus somewhat, retain this particular agent-oriented approach. "Developed in a world profoundly different from the small, non-Christian Athens, Utilitarianism, or more generally Consequentialism, Contractualism and Deontologism are the three most well-known theories that concentrate on the moral nature and value of the actions performed by the agent." [5] Although shifting emphasis from the "moral nature and development of the individual agent" to the "moral nature and value" of his or her actions, western philosophy has been, with few exceptions (which we will get to shortly), organized and developed as an agent-oriented endeavor.

When considered from the perspective of the agent, ethics inevitably and unavoidably makes exclusive decisions about *who* is to be included in the community of moral subjects and *what* can be excluded from consideration. The choice of words here is not accidental. As Jacques Derrida points everything turns on and is decided by the difference that separates the "who" from the "what." [6] Moral agency has been customarily restricted to those entities who call themselves and each other "man"—those beings who already give themselves the right to be considered someone who counts as opposed to something that does not. But who counts—who, in effect, gets to be situated under the term "who"—has never been entirely settled, and the historical development of moral philosophy can be interpreted as a progressive unfolding, where what had once been excluded (i.e., women, slaves, people of color, etc.) have slowly and not without considerable struggle and resistance been granted access to the gated community of moral agents and have thereby also come to be someone who counts.

Despite this progress, which is, depending on how one looks at it, either remarkable or insufferably protracted, there remain additional exclusions, most notably non-human animals and machines. Machines in particular have been understood to be mere artifacts that are designed, produced, and employed by human agents for human specified ends. This *instrumentalist* and *anthropocentric* understanding has achieved a remarkable level of acceptance and standard-ization, as is evident by the fact that it has remained in place and largely unchallenged from ancient to postmodern times —from at least Plato's *Phaedrus* to Jean-François Lyotard's *The Postmodern Condition* [7]. Beginning with the animal rights movement, however, there has been considerable pressure to reconsider the ontological assumptions and moral consequences of this legacy of human exceptionalism.

Extending consideration to these other previously marginalized subjects has required a significant reworking of the concept of moral agency, one that is not dependent on genetic make-up, species identification, or some other spurious criteria. As Peter Singer describes it, "the biological facts upon which the boundary of our species is drawn do not have moral significance," and to decide questions of moral agency on this ground "would put us in the same position as racists who give preference to those who are members of their race." [8] For this reason, the question of moral agency has come to be disengaged from identification with the human being and is instead referred to and made dependent upon the generic concept of "personhood." "There appears," G. E. Scott writes, "to be more unanimity as regards the claim that in order for an individual to be a moral agent s/he must possess the relevant features of a person; or, in other words, that being a person is a necessary, if not sufficient, condition for being a moral agent." [9] As promising as this "personist" innovation is, "the category of the person," to reuse terminology borrowed from Marcel Mauss [10], is by no means settled and clearly defined. There is, in fact, little or no agreement concerning what makes someone or something a person and the literature on this subject is littered with different formulations and often incompatible criteria. "One might well hope," Daniel Dennett writes, "that such an important concept, applied and denied so confidently, would have clearly formulatable necessary and sufficient conditions for ascription, but if it does, we have not yet discovered them. In the end there may be none to discover. In the end we may come to realize that the concept person is incoherent and obsolete." [11]

In an effort to contend with, if not resolve this problem, researchers often focus on the one "person making" quality that appears on most, if not all, the lists of "personal properties," whether they include just a couple simple elements [8] or involve numerous "interactive capacities" [12], and that already has traction with practitioners and theorists—consciousness. "Without consciousness," John Locke argued, "there is no person." [13] Or as Kenneth Einar Himma articulates it, "moral agency presupposes consciousness…and that the very concept of agency presupposes that agents are conscious." [14] Formulated in this fashion, moral agency is something that is decided and made dependent on a prior determination of consciousness. If, for example, an animal or a machine can in fact be shown to possess "consciousness," then that entity would, on this account, need to be considered a legitimate moral agent. And not surprisingly, there has been considerable effort in the fields of philosophy, AI, and robotics to address the question of machine moral agency by targeting and examining the question and possibility (or impossibility) of machine consciousness.

This seemingly rational approach, however, runs into considerable ontological and epistemological complications. On the one hand, we do not, it seems, have any widely accepted characterization of "consciousness." The problem, then, is that consciousness, although crucial for deciding who is and who is not a moral agent, is itself a term that is ultimately undecided and considerably equivocal. "The term," as Max Velmans points out, "means many different

things to many different people, and no universally agreed core meaning exists." [15] In fact, if there is any general agreement among philosophers, psychologists, cognitive scientists, neurobiologists, AI researchers, and robotics engineers regarding consciousness, it is that there is little or no agreement when it comes to defining and characterizing the concept. And to make matters worse, the problem is not just with the lack of a basic definition; the problem may itself already be a problem. "Not only is there no consensus on what the term *consciousness* denotes," Güven Güzeldere writes, "but neither is it immediately clear if there actually is a single, well-defined '*the* problem of consciousness' within disciplinary (let alone across disciplinary) boundaries. Perhaps the trouble lies not so much in the ill definition of the question, but in the fact that what passes under the term consciousness as an all too familiar, single, unified notion may be a tangled amalgam of several different concepts, each inflicted with its own separate problems." [16]

On the other hand, even if it were possible to define consciousness or come to some tentative agreement concerning its necessary and sufficient conditions, we still lack any credible and certain way to determine its actual presence in another. Because consciousness is a property attributed to "other minds," its presence or lack thereof requires access to something that is and remains fundamentally inaccessible. "How does one determine," as Paul Churchland characterizes it, "whether something other than oneself—an alien creature, a sophisticated robot, a socially active computer, or even another human—is really a thinking, feeling, conscious being; rather than, for example, an unconscious automaton whose behavior arises from something other than genuine mental states?" [17] And the supposed solutions to this "other minds problem," from reworkings and modifications of the Turing Test to functionalist approaches that endeavor to work around this problem altogether [18], only make things more complicated and confused. "There is," as Dennett points out, "no proving that something that seems to have an inner life does in fact have one—if by 'proving' we understand, as we often do, the evincing of evidence that can be seen to establish by principles already agreed upon that something is the case." [11] Although philosophers, psychologists, and neuroscientists throw considerable argumentative and experimental effort at this problem, it is not able to be resolved in any way approaching what would pass for empirical science, strictly speaking.[3] In the end, not only are these tests unable to demonstrate with any certitude whether animals, machines, or other entities are in fact conscious and therefore legitimate moral persons (or not), we are left doubting whether we can even say the same for other human beings. As Ray Kurzweil candidly concludes, "we assume other humans are conscious, but even that is an assumption," because "we cannot resolve issues of consciousness entirely through objective measurement and analysis (science)." [19]

The question of machine moral agency, therefore, turns out to be anything but simple or definitive. This is not, it is important to note, because machines are somehow unable to be moral agents. It is rather a product of the fact that the term "moral agent," for all its importance and argumentative expediency, has been and remains an ambiguous, indeterminate, and rather noisy concept. What the consideration of machine moral agency demonstrates, therefore, is something that may not have been anticipated or sought. What is discovered in the process of pursuing this line of inquiry is not a satisfactory answer to the question whether machines are able to be moral agents or not. In fact, that question remains open and unanswered. What has been ascertained is that the concept of moral agency is already vague and imprecise such that it is (if applied strictly and rigorously) uncertain whether we—whoever this "we" includes—are in fact moral agents.

What the machine question demonstrates, therefore, is that moral agency, the issue that had been assumed to be the "correct" place to begin, turns out to be inconclusive. Although this could be regarded as a "failure," it is a particularly instructive failing. What is learned from this failure—assuming we continue to use this obviously "negative" word—is that moral agency is not necessarily some property that can be definitively ascertained or discovered in others prior to and in advance of their moral consideration. Instead moral standing may be something (perhaps what Kay Foerst has called a dynamic and socially constructed "honorarium" [20]) that comes to be conferred and assigned to others in the process of our interactions and relationships with them. But then the deciding issue will no longer be one of agency; it will be a matter of *patiency*.

3. MORAL PATIENCY

Moral patiency looks at the ethical relationship from the other side. It is concerned not with determining the moral character of the agent or weighing the ethical significance of his/her/its actions but with the victim, recipient, or receiver of such action. This approach is, as Mane Hajdin [21], Luciano Floridi [5], and others have recognized, a significant alteration in procedure and a "non-standard" way to approach the question of moral rights and responsibilities. The model for this kind of transaction can be found in animal rights philosophy. Whereas agent-oriented ethics have been concerned with determining whether someone is or is not a legitimate moral subject with rights and responsibilities, animal rights philosophy begins with an entirely different question—"Can they suffer?" [22]

This seemingly simple and direct inquiry introduces what turns out to be a major paradigm shift in the basic structure and procedures of moral thinking. On the one hand, it challenges the anthropocentric tradition in ethics by questioning the often unexamined privilege human beings have granted themselves. In effect, it institutes something

[3] Attempts to resolve this problem often take the form of a pseudo-science called *physiognomy*, which endeavors to infer an entity's internal states of mind from the observation of its external expressions and behavior.

like a Copernican revolution in moral philosophy. Just as Copernicus challenged the geocentric model of the cosmos and in the process undermined many of the presumptions of human exceptionalism, animal rights philosophy contests the established Ptolemaic system of ethics, deposing the anthropocentric privilege that had traditionally organized the moral universe. On the other hand, the effect of this fundamental shift in focus means that the one time closed field of ethics can be opened up to other kinds of non-human animals. In other words, who counts as morally significant are not just other "men" but all kinds of entities that had previously been marginalized and situated outside the gates of the moral community. "If a being suffers," Peter Singer writes, "there can be no moral justification for refusing to take that suffering into consideration. No matter what the nature of the being, the principle of equality requires that its suffering be counted equally with the like suffering of any other being." [23]

Initially there seems to be good reasons and opportunities for extending this innovation to machines, or at least some species of machines. [24] This is because the animal and the machine, beginning with the work of René Descartes, share a common ontological status and position. For Descartes, the human being was considered the sole creature capable of rational thought—the one entity able to say, and be certain in its saying, *cogito ergo sum*. Following from this, he had concluded that other animals not only lacked reason but were nothing more than mindless automata that, like clockwork mechanisms, simply followed predetermined instructions programmed in the disposition of their various parts or organs. Conceptualized in this fashion, the animal and the machine, or what Descartes identified with the hybrid, hyphenated term *bête-machine*, were effectively indistinguishable and ontologically the same. "If any such machine," Descartes wrote, "had the organs and outward shape of a monkey or of some other animal that lacks reason, we should have no means of knowing that they did not possess entirely the same nature as these animals." [25]

Despite this fundamental and apparently irreducible similitude, only one of the pair has been considered a legitimate subject of moral concern. Even though the fate of the machine, from Descartes forward was intimately coupled with that of the animal, only the animal (and only some animals, at that) has qualified for any level of ethical consideration. And this exclusivity has been asserted and justified on the grounds that the machine, unlike the animal, does not experience either pleasure or pain. Although this conclusion appears to be rather reasonable and intuitive, it fails for a number of reasons.

First, it has been practically disputed by the construction of various mechanisms that now appear to suffer or at least provide external evidence of something that looks like pain. As Derrida recognized, "Descartes already spoke, as if by chance, of a machine that simulates the living animal so well that it 'cries out that you are hurting it.'" [26] This comment, which appears in a brief parenthetical aside in Descartes' *Discourse on Method*, had been deployed in the course of an argument that sought to differentiate human beings from the animal by associating the latter with mere mechanisms. But the comment can, in light of the procedures and protocols of animal ethics, be read otherwise. That is, if it were indeed possible to construct a machine that did exactly what Descartes had postulated, that is, "cry out that you are hurting it," would we not also be obligated to conclude that such a mechanism was capable of experiencing pain? This is, it is important to note, not just a theoretical point or speculative thought experiment. Engineers have, in fact, not only constructed mechanisms that synthesize believable emotional responses [27] [28] [29], like the dental-training robot Simroid "who" cries out in pain when students "hurt" it [30], but also systems capable of evidencing behaviors that look a lot like what we usually call pleasure and pain.

Second it can be contested on epistemologically grounds insofar as suffering or the experience of pain is still unable to get around or resolve the problem of other minds. How, for example, can one know that an animal or even another person actually suffers? How is it possible to access and evaluate the suffering that is experienced by another? "Modern philosophy," Matthew Calarco writes, "true to its Cartesian and scientific aspirations, is interested in the indubitable rather than the undeniable. Philosophers want proof that animals actually suffer, that animals are aware of their suffering, and they require an argument for why animal suffering should count on equal par with human suffering." [31] But such indubitable and certain knowledge, as explained by Marian S. Dawkins, appears to be unattainable:

> At first sight, 'suffering' and 'scientific' are not terms that can or should be considered together. When applied to ourselves, 'suffering' refers to the subjective experience of unpleasant emotions such as fear, pain and frustration that are private and known only to the person experiencing them. To use the term in relation to non-human animals, therefore, is to make the assumption that they too have subjective experiences that are private to them and therefore unknowable by us. 'Scientific' on the other hand, means the acquisition of knowledge through the testing of hypotheses using publicly observable events. The problem is that we know so little about human consciousness that we do not know what publicly observable events to look for in ourselves, let alone other species, to ascertain whether they are subjectively experiencing anything like our suffering. The scientific study of animal suffering would, therefore, seem to rest on an inherent contradiction: it requires the testing of the untestable. [32]

Because suffering is understood to be a subjective and private experience, there is no way to know, with any

certainty or credible empirical method, how another entity experiences unpleasant sensations such as fear, pain, or frustration. For this reason, it appears that the suffering of another (especially an animal) remains fundamentally inaccessible and unknowable. As Singer [23] readily admits, "we cannot directly experience anyone else's pain, whether that 'anyone' is our best friend or a stray dog. Pain is a state of consciousness, a 'mental event,' and as such it can never be observed." The machine question, therefore, leads to an outcome that was not necessarily anticipated. The basic problem is not whether the question "can they suffer?" applies to machines but whether anything that appears to suffer—human, animal, plant, or machine—actually does so at all.

Third, and to make matters even more complicated, we may not even know what "pain" and "the experience of pain" is in the first place. This point is something that is taken up and demonstrated by Daniel Dennett's "Why You Can't Make a Computer That Feels Pain." In this provocatively titled essay, originally published decades before the debut of even a rudimentary working prototype, Dennett imagines trying to disprove the standard argument for human (and animal) exceptionalism "by actually writing a pain program, or designing a pain-feeling robot." [11] At the end of what turns out to be a rather protracted and detailed consideration of the problem, he concludes that we cannot, in fact, make a computer that feels pain. But the reason for drawing this conclusion does not derive from what one might expect, nor does it offer any kind of support for the advocates of moral exceptionalism. According to Dennett, the reason you cannot make a computer that feels pain is not the result of some technological limitation with the mechanism or its programming. It is a product of the fact that we remain unable to decide what pain is in the first place. The best we are able to do, as Dennett illustrates, is account for the various "causes and effects of pain," but "pain itself does not appear." [11] What is demonstrated, therefore, is not that some workable concept of pain cannot come to be instantiated in the mechanism of a computer or a robot, either now or in the foreseeable future, but that the very concept of pain that would be instantiated is already arbitrary, inconclusive, and indeterminate. "There can," Dennett writes at the end of the essay, "be no true theory of pain, and so no computer or robot could instantiate the true theory of pain, which it would have to do to feel real pain." [11] Although Bentham's question "Can they suffer?" [22] may have radically reoriented the direction of moral philosophy, the fact remains that "pain" and "suffering" are just as nebulous and difficult to define and locate as the concepts they were intended to replace.

Finally, all this talk about the possibility of engineering pain or suffering in a machine entails its own particular moral dilemma. "If (ro)bots might one day be capable of experiencing pain and other affective states," Wendell Wallach and Colin Allen write, "a question that arises is whether it will be moral to build such systems—not because of how they might harm humans, but because of the pain these artificial systems will themselves experience. In other words, can the building of a (ro)bot with a somatic architecture capable of feeling intense pain be morally justified and should it be prohibited?" [18] If it were in fact possible to construct a machine that "feels pain" (however defined and instantiated) in order to demonstrate the limits of moral patiency, then doing so might be ethically suspect insofar as in constructing such a mechanism we do not do everything in our power to minimize its suffering. Consequently, moral philosophers and robotics engineers find themselves in a curious and not entirely comfortable situation. One needs to be able to construct such a mechanism in order to demonstrate moral patiency and the moral standing of machines; but doing so would be, on that account, already to engage in an act that could potentially be considered immoral. Or to put it another way, the demonstration of machine moral patiency might itself be something that is quite painful for others.

For these reasons, approaching the machine question from the perspective of moral patiency also encounters fundamental difficulties. Despite initial promises, we cannot, it seems, make a credible case for or against the moral standing of the machine by simply following the patient-oriented approach modeled by animal rights philosophy. In fact, trying to do so produces some rather unexpected results. In particular, extending these innovations does not provide definitive proof that the machine either can be or is not able to be a similarly constructed moral patient. Instead doing so demonstrates how the "animal question"—the question that has in effect revolutionized ethics in the later half of the 20th century—might already be misguided and prejudicial. Although it was not necessarily designed to work in this fashion, "A Vindication of the Rights of Machines" achieves something similar to what Thomas Taylor had wanted for his *A Vindication of the Rights of Brutes*. Taylor, who wrote and distributed this pamphlet under the protection of anonymity, originally composed his essay as a means by which to parody and undermine the arguments that had been advanced in Wollstonecraft's *A Vindication of the Rights of Woman*. [23] Taylor's text, in other words, was initially offered as a kind of *reductio ad absurdum* designed to exhibit what he perceived to be the conceptual failings of Wollstonecraft's proto-feminist manifesto. Following suit, "A Vindication of the Rights of Machines" appears to have the effect of questioning and even destabilizing what had been achieved with animal rights philosophy. But as was the case with the consideration of moral agency, this negative outcome is informative and telling. In particular, it indicates to what extent this apparent revolution in moral thinking is, for all its insight and promise, still beset with fundamental problems that proceed not so much from the ontological condition of these other, previously excluded entities but from systemic problems in the very structure and protocols of moral reasoning.

4. ULTERIOR MORALS

"Every philosophy," Silvo Benso writes in a comprehensive gesture that performs precisely what it seeks to address, "is a quest for wholeness." [33] This objective, she argues, has been typically targeted in one of two ways. "Traditional Western thought has pursued wholeness by means of reduction, integration, systematization of all its parts. Totality has replaced wholeness, and the result is totalitarianism from which what is truly other escapes, revealing the deficiencies and fallacies of the attempted system." [33] This is precisely the kind of violent philosophizing that Emmanuel Levinas identifies under the term "totality," and which includes, for him at least, the big landmark figures like Plato, Kant, and Heidegger. [34] The alternative to this totalizing approach is a philosophy that is oriented otherwise, like that proposed and developed by Singer, Birch, Levinas, and others. This other approach, however, "must do so by moving not from the same, but from the other, and not only the Other, but also the other of the Other, and, if that is the case, the other of the other of the Other. In this *must*, it must also be aware of the inescapable injustice embedded in any formulation of the other." [33] What is interesting about these two strategies is not what makes them different from one another or how they articulate approaches that proceeds from what appears to be opposite ends of the spectrum. What is interesting is what they agree upon and hold in common in order to be situated as different from and in opposition to each other in the first place. Whether taking the form of autology or some kind of heterology, "they both share the same claim to inclusiveness" [33], and that is the problem.

When it comes to including previously excluded subjects, then, moral philosophy appears to be caught between a proverbial rock and a hard place. On the one hand, the same has never been inclusive enough to adequately accommodate others. The machine in particular is already and from the very beginning situated outside ethics. It is, irrespective of the different philosophical perspectives that come to be mobilized, typically regarded as neither a legitimate moral agent nor patient. It has been and continues to be widely understood as nothing more than an instrument to be employed more or less effectively by human beings and, for this reason, is always and already located in excess of moral considerability or to use that distinct Nietzschean characterization, "beyond good and evil." [35] Technology, as Lyotard reminds us, is only a matter of efficiency. Technical devices do not participate in the big questions of metaphysics, aesthetics, or ethics. [7] They are nothing more than contrivances or extensions of human agency, used more or less responsibly by human agents with the outcome effecting other human patients. Although other kinds of previously marginalized others—animals, the environment, and even corporations—have been slowly and not without considerable struggle granted some level of membership in the community of moral subjects, the machine is and remains on the periphery. "We have never," as J. Storrs Hall points out, "considered ourselves to have 'moral' duties to our machines, or them to us." [36]

On the other hand, alternatives to this tradition, like the patient-oriented approach of animal rights philosophy, have never been different enough. Although a concern with and for others promised to radicalize the procedures of moral reasoning, ethics has not been suitably different. Many of the so-called alternatives, those philosophies that purport to be interested in and oriented otherwise, have typically excluded the machine from what is considered Other. Technological devices certainly have an interface but they do not, as Levinas would have it, possess a face or confront us in a face-to-face encounter that would call for and would be called ethics. [34] This exclusivity is not simply the last socially accepted prejudice or what Singer calls "the last remaining form of discrimination" [3], which may be identified as such only from a perspective that is already open to the possibility of some future inclusion and accommodation. The marginalization of the machine appears to be much more complete and pervasive. In fact, the machine does not constitute just one more form of difference that would be included at some future time. It comprises the very mechanism of exclusion. "In the eyes of many philosophers," Dennett writes, "the old question of whether determinism (or indeterminism) is incompatible with moral responsibility has been superseded by the hypothesis that mechanism may well be." [11] Consequently, whenever a philosophy endeavors to make a decision, to demarcate and draw the line separating "us" from "them," or to differentiate who does and what does not have moral standing, it inevitably fabricates machines. When Tom Regan, for instance, sought to distinguish which higher-order animals qualify for moral consideration as opposed to those lower-order entities that do not, he marginalizes the latter by characterizing them as mere machines. [37]

For these reasons, the machine does not constitute one more historically marginalized other that would need to be granted admission to the class of moral *consideranda*. In fact, it now appears that the machine is unable to achieve what is considered to be necessary for either moral agency or patiency. But this inability is not, we can say following the argumentative strategy of Dennett's "Why you Cannot Make a Computer that Feels Pain" [11], a product of some inherent or essential deficiency with the machine. Instead it is a result of the fact that both agency and patiency already lack clearly defined necessary and sufficient conditions. "A Vindication of the Rights of Machines," then, does not end by accumulating evidence or arguments in favor of permitting one more entity entry into the community of moral subjects; it concludes by critically questioning the very protocols of inclusion/exclusion that have organized and structured moral philosophy from the beginning.

What this means for ethics is that Descartes—that thinker who had been regarded as the "bad guy" of modern philosophy by Regan [37] and others [38]—may have

actually gotten it right despite himself and our usual (mis)interpretations of his work. In the *Discourse on the Method*, something of a philosophical autobiography, Descartes famously endeavored to tear down to its foundations every truth that he had come to accept or had taken for granted. This approach, which will in the *Meditations* come to be called "the method of doubt," targets everything, including the accepted truths of ethics. With Descartes, then, one thing is certain, he did not want to be nor would he tolerate being duped. However, pursuing and maintaining this extreme form of critical inquiry that does not respect any pre-established boundaries has very real practical expenses and implications. For this reason, Descartes decides to adopt a "provisional moral code," something of a temporary but relatively stable structure that would support and shelter him as he engaged in this thorough questioning of everything and anything.

> Now, before starting to rebuild your house, it is not enough simply to pull it down, to make provision for materials and architects (or else train yourself in architecture), and to have carefully drawn up the plans; you must also provide yourself with some other place where you can live comfortably while building is in progress. Likewise, lest I should remain indecisive in my actions while reason obliged me to be so in my judgments, and in order to live as happily as I could during this time, I formed for myself a provisional moral code consisting of just three or four maxims, which I should like to tell you about. [25]

Understood and formulated as "provisional," it might be assumed that this protocol would, at some future time, be replaced by something more certain and permanent. But Descartes, for whatever reason, never explicitly returns to it in order to finalize things. This is, despite initial appearances, not a deficiency, failure, or oversight. It may, in fact, be the truth of the matter. Namely that, as Slavoj Žižek describes it, "all morality we adopt is provisory." [39] In this case, then, what would have customarily been considered "failure," that is, the lack of ever achieving the *terra firma* of moral certitude, is reconceived of as a kind of success and advancement. Consequently, "failure," Žižek argues, "is no longer perceived as opposed to success, since success itself can consist only in heroically assuming the full dimension of failure itself, 'repeating' failure as 'one's own.'" [39] In other words, the provisory nature of ethics is not a failure as opposed to some other presumed outcome that would be called "success." It is only by assuming and affirming this supposed "failure" that what is called ethics will have succeeded.

Ethics, conceived of in this fashion, is not determined by a prior ontological discovery concerning the essential capabilities or internal operations of others. It is rather a decision—literally a cut that institutes difference and that makes a difference by dividing between *who* is considered to be morally significant and *what* is not. Consequently, "moral consideration is," as Mark Coeckelbergh describes it, "no longer seen as being 'intrinsic' to the entity: instead it is seen as something that is 'extrinsic': it is attributed to entities within social relations and within a social context." [40] This is the reason why, as Levinas claims, "morality is first philosophy" ("first" in terms of both sequence and status) and that moral decision making precedes ontological knowledge. [34][4] What this means, in the final analysis, is that we—we who already occupy a privileged position within the community of moral subjects—are responsible for determining the proper scope and boundaries of moral responsibility, for instituting these decisions in everyday practices, and for evaluating their results and outcomes. Although we have often sought to deflect these decisions and responsibilities elsewhere, typically into the heavens but also onto other terrestrial authorities, in order to validate and/or to avoid having to take responsibility for them, we are, in the final analysis, the sole responsible party. We are, in other words, not just responsible for acting responsibly in accordance with ethics; we are responsible for ethics. The vindication of the rights of machines, therefore, is not simply a matter of extending moral consideration to one more historically excluded other. The question concerning the "rights of machines" makes a fundamental claim on ethics, requiring us to rethink the system of moral considerability all the way down.

REFERENCES

[1] S. Anderson. Asimov's "Three Laws of Robotics" and Machine Metaethics. *AI & Society* 22(4): 477–493 (2008)

[2] D. Gunkel. *The Machine Question: Critical Perspectives on AI, Robots and Ethics*. MIT Press, USA (2012).

[3] P. Singer. All Animals are Equal. In *Animal Rights and Human Obligations*. Tom Regan and Peter Singer (Eds), pp. 148–162. Prentice Hall, USA (1989).

[4] T. Birch. Moral Considerability and Universal Consideration. *Environmental Ethics* 15: 313–332 (1993).

[5] L. Floridi. Information Ethics: On the Philosophical Foundation of Computer Ethics. *Ethics and Information Technology* 1(1): 37–56 (1999).

[6] J. Derrida. *Paper Machine*. Rachel Bowlby (Trans.). Stanford University Press, USA (2005).

[7] J. Lyotard. *The Postmodern Condition: A Report on Knowledge*. G. Bennington and B. Massumi (Trans.). University of Minnesota Press, USA (1984).

[8] P. Singer. *Practical Ethics*. Cambridge University Press, UK (1999).

[9] G. E. Scott. *Moral Personhood: An Essay in the Philosophy of Moral Psychology*. SUNY Press, USA (1990).

[10] M. Mauss. A Category of the Human Mind: The Notion of Person; The Notion of Self. W. D. Halls (Trans.). In *The Category of the Person*, M. Carrithers, S. Collins, and S. Lukes (Eds.), pp. 1–25. Cambridge University Press, UK (1985).

[11] D. Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, USA (1998).

[4] If there is a philosophical tradition that is explicitly dedicated to pursuing this procedural inversion, it is arguably German Idealism in general and Kant in particular, all of which assert, as Žižek characterizes it, "the primacy of practical over theoretical reason." [41]

[12] C. Smith. *What Is a Person? Rethinking Humanity, Social Life, and the Moral Good from the Person Up*. University of Chicago Press, USA (2010).
[13] J. Locke. *An Essay Concerning Human Understanding*. Hackett, USA (1996).
[14] K. E. Himma. Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to be a Moral Agent? *Ethics and Information Technology* 11(1):19–29 (2009).
[15] M. Velmans. *Understanding Consciousness*. Routledge, UK (2000).
[16] G. Güzeldere. The Many Faces of Consciousness: A Field Guide. In The Nature *of Consciousness: Philosophical Debates*, N. Block, O. Flanagan, and G. Güzeldere (Eds.), pp. 1–68. MIT Press, USA (1997).
[17] P. M. Churchland. *Matter and Consciousness*, rev. ed. Cambridge, MIT Press, USA (1999).
[18] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, UK (2009).
[19] R. Kurzweil. *The Singularity Is Near: When Humans Transcend Biology*. Viking, USA (2005).
[20] G. Benford and E. Malartre. 2007. *Beyond Human: Living with Robots and Cyborgs*. Tom Doherty, USA (2007).
[21] M. Hajdin. *The Boundaries of Moral Discourse*. Loyola University Press (1994).
[22] J. Bentham. *An Introduction to the Principles of Morals and Legislation*. J. H. Burns and H. L. Hart (Eds.). Oxford University Press, UK (2005).
[23] P. Singer. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York Review of Books, USA (1975).
[24] D. J. Gunkel. *Thinking Otherwise: Philosophy, Communication, Technology*. Purdue University Press, USA (2007).
[25] R. Descartes. *Selected Philosophical Writings*. J. Cottingham, R. Stoothoff, and D. Murdoch (Trans.). Cambridge University Press, UK (1988).
[26] J. Derrida. *The Animal That Therefore I Am*. M. Mallet (Ed.) and David Wills (Trans.). Fordham University Press, USA (2008).
[27] J. Bates. The Role of Emotion in Believable Agents. *Communications of the ACM* 37: 122–125 (1994).
[28] B. Blumberg, P. Todd and M. Maes. No Bad Dogs: Ethological Lessons for Learning. In *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior* (SAB96), pp. 295–304. MIT Press, USA (1996).
[29] C. Breazeal and R. Brooks. Robot Emotion: A Functional Perspective. In *Who Needs Emotions: The Brain Meets the Robot*, J. M. Fellous and M. Arbib (Eds.), pp. 271–310. Oxford University Press, UK (2004).
[30] Kokoro, L. T. D. http://www.kokoro-dreams.co.jp/ (2009).
[31] M. Calarco. *Zoographies: The Question of the Animal from Heidegger to Derrida*. Columbia University Press, USA (2008)
[32] M. S. Dawkins. The Science of Animal Suffering. *Ethology* 114(10): 937–945 (2008).
[33] S. Benso. *The Face of Things: A Different Side of Ethics*. SUNY Press, USA (2000).
[34] E. Levinas. *Totality and Infinity: An Essay on Exteriority.* A. Lingis (Trans.). Duquesne University Press, USA (1969).
[35] F. Nietzsche. *Beyond Good and Evil.* W. Kaufmann (Trans.). Vintage Books, USA (1966).
[36] J. S. Hall. Ethics for Machines. http://www.kurzweilai.net/ethics-for-machines. (5 July 2001).
[37] T. Regan. *The Case for Animal Rights*. University of California Press, USA (1983).
[38] A. Lippit. *Electric Animal: Toward a Rhetoric of Wildlife*. University of Minnesota Press, USA (2000).
[39] S. Žižek. *The Parallax View*. MIT Press, USA (2006).
[40] M. Coeckelbergh. Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology*, 12(3): 209-221 (2010).
[41] S. Žižek. *Less Than Nothing: Hegel and the Shadow of Dialectical Materialism*. Verso, USA (2012).

The centrality of machine consciousness to machine ethics: Between realism and social-relationism

Steve Torrance*

Abstract. I compare a 'realist' with a 'social-relational' perspective on our judgments of the moral status of machines. I argue that moral status is closely bound up with a being's ability to experience states of conscious satisfaction or suffering (CSS). The social-relational view may be right that a wide variety of social interactions between us and machines will proliferate in future generations, and that the appearance of CSS-features in such machines may make moral-role attribution socially prevalent in human-machine relations. But the social world is enabled and constrained by the physical world. Features analogous to physiological features in biological CSS are what need to be present for non-biological CSS. Working out the details of such features will be a scientific inquiry sharing the same kind of 'objectivity' as, for instance, physicists' questions about dark matter.

1 INTRODUCTION

Wallach et al [1] have written that 'Machine ethics and machine consciousness are joined at the hip'.[1] The attribution of consciousness to machines is a fundamental consideration in assessing the ethical status of machines – both as moral agents and as moral patients. Yet how are we to understand such consciousness-attributions; and indeed, how are we to view attributions of moral status themselves?

I compare two views on this, which are discussed by Coeckelbergh in his contribution [6] and in other writing [7-10]. On the view he labels 'objectivism' or 'realism', a question like 'Is X conscious?' (supposing X to be a machine, a human, a dolphin, or whatever) is asking about objective matters of fact concerning X's psychological state. Also, for realism, attributions of moral status supervene on X's psychological properties (including consciousness and related states). On the version of realism I will discuss here, normative moral questions concerning the actions or the treatment of humans, animals or robots, are closely tied up with factual questions concerning the well-being (or ill-being) of such creatures, assuming them to be conscious creatures. For such a view, an important part of what would be involved in a machine's actually being conscious – phenomenally conscious, as opposed to just functioning cognitively in a way a conscious being would – is that machine's having moral worth in virtue of its being capable of experiencing positively or negatively valenced states of satisfaction or suffering.[2] Thus machine consciousness and machine ethics (at least the study of machines as moral patients) will tend to be tightly coupled, on such a position. Versions of this view are defended in [3,4,11, etc.]

An opposing view – Coeckelbergh's 'social-relational' view – claims that attributions of consciousness are not (or at least not clearly) ascriptions of matters of objective fact, at least in the case of non-human animals, and of current and future technological agents. On this view such ascriptions rather have to be understood in terms of the organizational circumstances in the society in which the discourse of attribution occurs, on the social relations between human moral agents, and the contexts in which other putatively conscious creatures or agents may enter into our social lives. These social and technical contexts vary from culture to culture and from epoch to epoch. As society is fast-changing today, so new criteria for consciousness-attribution may currently be emerging, which are likely to radically alter opinion on what beings to treat as conscious. Moreover, on the social-relational view attributions of moral worth and other moral qualities are similarly to be seen as essentially embedded in social relations. The same profound changes in social view are likely to affect norms concerning the attribution of moral status, in particular the moral status of artificial creatures. In a word, judgments in the 21st century about the possible experiential and moral status of automata may markedly diverge from those that were prevalent in previous centuries. A social-relationist view will say there may be no neutral way of judging between these different views.

* School of Engineering and Informatics, University of Sussex, Falmer, Brighton, BN1 9QJ, UK. Email: stevet@sussex.ac.uk

[1] Wallach et al are primarily concerned in their paper with how modeling moral agency requires a proper analysis of conscious decision-making , whereas the present paper largely centres around the relation between sentience and being a moral patient – but the sentiments are broadly overlapping. I've discussed the general relation between consciousness and ethics in an AI context in [2-5].

[2] I defend the inherent distinguishability between phenomenal and functional consciousness in [4].

Thus, on the social-relationist view, both psychological realism and moral realism are rejected.[3] [4]

The orientation of this paper is broadly sympathetic to the psychological realist view, and, indeed, to (some versions of) moral realism. However, despite their apparent diametrical opposition, perhaps some kind of synthesis can be reached. It seems clear, in any event, that the debate is an extremely significant one at this juncture of human civilization and technical development.

2 THE EXPANDING CIRCLE

Many writers have talked of a progressive expansion of moral outlook through human pre-history and history. Peter Singer has written persuasively [14] of the 'expanding circle' of ethical concern, from primitive kith- and tribe-centred fellow-feeling to a universal regard for all of humanity; and of the rational imperative to widen the circle still further to include non-human creatures capable of sentient feeling (see also discussion in [5]). We owe our ethics to the evolutionary pressures on our pre-human forbears, and we owe it the animal co-descendents of that evolutionary process to extend ethical concern to the well-being of all sentient creatures. Some have argued that the circle should be widened so that non-sentient entities such as forests, mountains, oceans, etc. should be included within the domain of direct moral consideration (rather than just instrumentally, in terms of how they affect the well-being of sentient creatures) [15, 16]. In considering what limits might be put on this process of ethical expansion, Singer argues that *only* entities that have the potentiality for sentience could sensibly be included in the moral circle. For, he says, of a being with no sentience there can be nothing that one can do which could make a difference *to* that being in terms of what it might experience. [14: p. 123]

Singer's position partially rests on a conceptual claim – that only beings with sentience can coherently be considered as moral patients. To see that the claim is a conceptual one consider this: roughly, for X to be a 'moral patient' (or moral 'recipient') *means* (at least in part) that X is capable of benefitting or suffering from a given action; and a non-sentient being cannot benefit or suffer in the relevant (experiential) sense (although, like an antique wooden table left out in the rain, it may suffer in a non-experiential sense). So, the argument goes, a non-sentient being cannot *coherently* be considered as a moral patient.

Of course the precise conditions under which particular artificial agents might be considered conscious are notoriously difficult to pin down. But having controversial verification-conditions is not the same as having verification-conditions essentially dependent upon the social-relational context. There may be much controversy among physicists over the precise conditions under which Dark Matter may be said to be established to exist; but this does not detract from the ontological objectivity of dark matter as an entity in the universe; it does not make physics social-relational. Similarly a robot's pain, if such a thing were to exist, would be as objective a property of the robot, and as inalienable *from* the robot, as would your or my pain be inalienable from you or from me. (Conversely, a robot that gave appearances of pain or suffering but which in fact had no sentience could not have sentience added to it simply by virtue of its convincing appearance.)

3 THE LANDSCAPE OF CONSCIOUS WELL-BEING

For realism, in the version I am sympathetic to (and possibly also for Coeckelbergh's version of social-relationism) there appears to be a kind of holism between thinking of X as phenomenally conscious and judging X to be of moral worth (at least as a moral patient, and maybe as a moral agent, in a full sense of agency). To think of a creature as having conscious experience is to think of it as capable of experiencing things in either a positively or negatively valenced way – to think of it as having desires, needs, goals, and states of satisfaction and dissatisfaction or suffering. Of course there are neutral experiential states, and not all satisfaction or suffering is consciously experienced. Nor are all our goals concerned with gaining particular experienced satisfactions. Nevertheless there seems to be a strong connection between our experiential capacities and our potential for well-being. (This is a point which has been addressed surprisingly little in the literature on human consciousness, and of machine consciousness.) We may talk of beings which are conscious, in this rich sense, as having the capacity for conscious/satisfaction/suffering states (I'll here call these CSS states for short).

In *The Moral Landscape* [17] Sam Harris has argued (in a broadly utilitarian way) that the well-being of conscious creatures is the central issue in ethics, and indeed that other ethical considerations are, if not nonsensical, at root

[3] This is not to say that questions concerning consciousness or ethics in relation to such machines are to be thought of as trivial or nugatory: on the contrary, a social-relationist may take such questions as seriously as the realist would, and may claim that they deserve our full intellectual and practical attention.

[4] See also David Gunkel's contribution to this symposium [12] and his book [13]. Gunkel offers deep and subtle critiques of the realist approach. Sadly, there is no space to respond specifically to his points here.

appeals to consideration of experienced well-being – to the quality of CSS states. Harris's moral landscape is the terrain of possible peaks and troughs in experienced well-being that CSS-capable creatures negotiate through their lives. He also argues (in a particularly strong version of the realist position) that moral questions are objective and in principle scientific in nature (as a neuroscientist, he takes brain-processes to be central determinants of well- or ill-being). It may be hard in practice to solve various moral dilemmas, but in principle, he claims, they are amenable to a scientific solution, like other tough factual questions such as curing cancer or eliminating global poverty.

I don't necessarily say ethics is exclusively about well-being; but I would agree that it is central to ethics. Also, since creatures with capacities for greater or lesser well-being are conscious, I think it is central to the study of consciousness, and, indeed to AI. Of course physiological processes from the neck upwards are pretty crucial for such capacities. But bodily processes from the neck down are pretty important too: a big challenge for AI and Machine Ethics is to work out just what physical features need to be present both above and below the neck in artificial agents for artificially-generated CSS properties to be present, and how close they have to be with the relevant natural or organic features. (Harris does not consider the ethical status of possible artificial agents.)

Linked with his views about the scientific grounding of questions to do with well-being and ethics, Harris expresses a lot of impatience with the 'fact-value' split that has dominated scientific thinking in the last century, and for the moral neutralism or quietism that has characterized a lot of scientific thought and practice in that time. This has been particularly true of 'Cognitive Science' as this has developed in the last half-century or so. AI researchers, neuroscientists, psychologists and philosophers often talk as though the mind was exclusively a cognitive mechanism: All that 'cognitivizing' hard work at reducing the *chiaroscuro* of our desires, emotions, pains and delights to informational operations, all that bleaching out of happiness and misery from the fabric of our psychology, has, arguably, been to the detriment of both scientific understanding and ethical discussion in this area.

4 A 100-YEAR SCENARIO

I thus applaud the science-ethics holism found in objectivist writers like Harris. Perhaps this enthusiasm will be shared by social-relationists such as Coeckelbergh. It is certainly interesting to discuss machine ethics in a way that takes seriously the inherent interconnectivity of consciousness, well-being and ethics, and which allows that scientific and ethical issues are not to be debated in parallel, hermetically sealed chambers.[5]

Further to this, then, consider the following future picture: 100 years from now (possibly sooner), human social relations will (we suppose) have changed radically because of the existence of large numbers of artificial agents implementing nearly-human or (if singularity theorists are half right) even supra-human levels of skills across a wide range of capabilities. Many people will find it natural to attribute a wide range of psychological attributes to such agents, and the agents themselves will, in their communications with us and with each other, represent themselves as having many of the cognitive and indeed affective states that we currently take to be characteristic of human psychology. Many such artificial creatures may resemble humans in outward form, but even if they don't, and the currently fashionable technology of humanoid robotics runs into a cul-de-sac, it nevertheless seems likely that the demands of extensive human-AI social interaction will ensure a good deal of resemblance in non-bodily respects (for instance in terms of sharing common languages, participating in a common economic system, shared legal frameworks, and so on).

Will our world wind up like this? Who knows, but the scenario will help us check our intuitions. Humans in this imagined future period may ask: are such artificial agents conscious? And, should we admit such agents into 'our' moral universe (and in what ways)? As we have argued, such questions are closely linked. We can combine those questions together in a third: do such artificial agents have CSS-features? The social-relationist will say that the answer to these questions will depend on the prevailing social conditions at the time, on what kinds of attitudes, beliefs, forms of life, and ways of articulating or representing social reality come to emerge in such a joint human-technological social milieu. On the relational view, there will be no 'objective' way, independently of the socially dominant assumptions and judgments and norms and institutions that grow up as such artificial agents proliferate, to say whether they are *actually* conscious, whether they *actually* have states of welfare or suffering, or whether they actually merit particular kinds of moral consideration - e.g. whether they merit having their needs taken roughly as seriously as equivalent human needs; whether their actions merit appraisal in

[5] Sometimes the seals can be leaky. I was once at a conference on consciousness, where an eminent neuropsychologist was giving a seminar on ethical issues in neural research on consciousness. He said things like 'With my neuroscientist's cap on, I think . . . But with my ethicist's cap on, I think . . .' What cap was he wearing when deciding which cap to put on at a given time?

roughly similar moral terms as the equivalent actions of humans, etc.[6]

For the realist this would miss an important dimension: do such artificial creatures (a few or a many of them) *actually* bear conscious states, are they *actually* capable of experiencing states of satisfaction or suffering at levels comparable to ours (or at lower, or even much higher, levels). To see the force of the realist's argument, consider how a gathering of future artificial agents might discuss the issue with respect to *humans'* having CSS properties - perhaps at a Turing-2112 convention? Let's suppose that delegates' opinions divide along roughly similar lines to those in the current human debate, with social-relationists arguing that there is no objective fact of the matter about whether humans have CSS properties, and realists insisting that there must be a fact of the matter.[7]

How would a human listening in on this argument feel about such a discussion? I would suggest that only a few philosophically sophisticated folks would feel comfortable with the social-relationist side of the argument in this robot convention, and that the most instinctive human response would a realist one. A human would reflect that we do, as a species, clearly possess a wide variety of CSS properties – indeed our personal and social lives revolve 24/7 around such properties? Can there be any issue over which there is more paradigmatically a 'fact of the matter' than our human consciousness? Surely the 'facts' point conclusively to the presence of CSS properties in humans: and are not such properties clearly tied to deep and extensive (neuro)physiological features in us? What stronger evidence-base for any 'Is X really there?' question could there be than the kind of evidence we have for CSS properties in humanity? So surely robots in 2112 would be right to adopt a realist view about the consciousness and ethical status of humans. Should we then not do the same in 2012 (or indeed in 2112) of robots? [8]

5 OTHER MINDS

One might raise 'other-minds'-style difficulties even about CSS in humans.[9] Any individual human's recognition of CSS properties is based upon their own direct first-person experience, it might be said, and, as Wittgenstein said, sarcastically, 'how can I generalize the one case so irresponsibly?' [20: I, §293] There are several responses. One is that, if CSS properties are not completely mysterious and inexplicable, from a scientific point of view, they must be causally grounded in natural features of any individual human possessing them, and it would be unlikely that any such underpinning natural features are found in a single person (or a few) rather being present across the species. Even in the vanishingly unlikely solipsist case where I am the *only* human with CSS properties, it would still be an objective, scientific matter that I have such properties, would it not? If so, then this would seem to marginalize doubts about 'other minds' as a philosophical side-show without any practical relevance to our functioning, social lives, and also without any possible value as a hypothesis worthy of scientific investigation.[10] Still, it comes up often in this sort of discussion (Turing used it in one of the least-satisfying passages of his 1950 paper [21]).

Another response is that first-person evidence forms only a part of the evidence-base I have as an individual for my CSS properties. The way I come to recognize and articulate CSS properties in myself is partly based upon my social interactions with my conspecifics. Indeed (the more serious point behind the Wittgenstein quip cited above) my whole discourse about mind, consciousness, pain, desires, emotions, etc. is based upon the public forms of life I share with other humans. This is also true in scientific terms: there is a mass of scientific theory and accreted experimental data linking CSS properties in humans with our evolutionary history, and with our current biological and neural make-up.

This kind of point of course approaches to the social-relational terrain. But we may need to be careful here. The social-relationist may argue that the very Wittgenstinian considerations just mentioned (and related arguments for the necessary public, social grounding of such recognition, and on the impossibility of private language and private cognitive rules, etc.) shed doubt on

[6] It's worth pointing out that no consensual human view may come to predominate on these issues: there may rather be a fundamental divergence just as there is in current societies between liberals and conservatives, or between theistic and humanistic ways of thinking, or between envirocentric versus technocentric attitudes towards the future of the planet, and so on. In such a case, the relationist could say, social reality will be just as it manifests itself – one in which no settled view on the psychological or moral status of such agents comes to prevail; society will just contain irreconcilable social disagreements on these matters, much as it does today on these other issues.

[7] We assume – perhaps with wild optimism – that that these artificial agents are by then smart enough to debate such matters somewhat as cogently as humans can today, if not much more so. To get a possible flavour of the debate, consider Terry Bisson's 'They're made out of meat' [18].

[8] That is, should we not say that the *epistemological status* of our question about them is comparable to that of theirs about us – although the answers to the two questions may be very different, as may be the relative difficulty in answering them?

[9] Such doubts are not raised by Coeckelbergh in his present contribution, but Gunkel does indeed cite traditional doubts about other minds as a way of putting pressure on realist responses to the 'machine question' [12].

[10] For a dismissive response to 'other-minds' doubts from within the phenomenological tradition, see Gallagher and Zahavi [19].

the objectivity of first-personal recognition of CSS properties. This may be so, and it may point to a deep truth behind the social-relational position – the truth that our understanding of how we come to possess CSS properties and the variety of roles they play in our lives is indeed inextricably bound up with our social relationships and activities.

But it's important to see that the dependency also goes in the other direction: our consciousness, needs, desires, etc. are what give point and form to our sociality. Our social conditions partly gain their significance, their point, from these very experiential and appetitive features in our lives, including, centrally, the ups of happiness and downs of misery. Thus sociality and consciousness thus co-determine each other (cf [22]). But also myriad 'objective' physical and biological realities – including a variety of evolutionary, neural and physiological constraints – come into this network of inter-relations between consciousness and the social. Evolutionary history and current brain-patterns play crucial roles in what makes us feel good or bad, as do the materiality of our bodies and the dynamics of their interaction with other bodies and the surrounding physical world.

6 SOCIAL AND NON-SOCIAL SHAPERS OF 'SOCIALITY'

So there is a multi-directional cluster of mutually constitutive and constraining relationships between the social, material, biological and experiential factors in our lives. (And no doubt many others – for instance I have left out the essential role played by our cognitive capacities, by beliefs, perceptions, intellective skills, etc.!) What makes up our CSS features emerges from this entanglement of these various kinds of factors. This brings us to the heart of the question of CSS in machine-agents.

What the progress of current work in AI teaches us is that many of the features of human-human social and communicative interaction can be replicated via techniques in computer and robotic science. Increasingly our social world is being filled with human-machine and machine-machine interactions. With the growing ubiquity of such interactions, the range of possible social action being extended. But also, our very conceptions of the social, are being re-engineered. This is a crucial point that the social-relationist seeks to emphasize, in the debate about the status of CSS properties; and the realist must also acknowledge it readily.

Of course, notions related to social status overlap intimately with ethical notions; and the social-relationist account is well suited to provide a theoretical framework for much of the domain of the social – what is taken to constitute 'the social' is largely itself shaped by social factors, and changes as new social possibilities emerge. But, as we have argued, the domain of the social is itself also shaped and constrained by the non-social, including biological and other kinds of physical conditions, and the experiences, desires, beliefs, goals, etc.of social participants. So many of the novel forms of human-machine and machine-machine social relationships that will emerge (and already have been emerging) will take their character not merely from the social sphere itself but also from the non-social soil in which sociality is rooted, that is, the multiple physical, metabolic and experiential drivers of sociality. This is particularly true of social relationships between humans and machines, which are precisely NOT relations between creatures with shared physiologies. And, for robots and other artificial agents of today, we can surely say with near certainty that they are NOT relations between beings with common experiential and affective subjectivities.

7 THE SPECTRUM – SOCIAL AND EXPERIENTIAL

So, for the short term - as long as we have only the relatively primitive designs of our current technologies - our artificial social partners are, objectively, partners with zero experiential or affective life (despite many vociferous assertions to the contrary). Such artificial social partners are not, in Tom Regan's phrase, 'subjects of a life' [23]. So, for now, the possibilities for social interaction with machines outstrip the possibilities for those machines being capable of sharing such interactions as exchanges between experiencing, belief-and-desire-ful beings: for now, any human-machine interaction is one between social partners where only one of the actors has any *social concern*. So some of the deep conditions for sociality mentioned earlier are missing – in particular, a shared experientiality and shared neurophysiology. In the human-machine case, then, we might talk of a current mismatch between social *interactivity* and social *constitutivity*.

But how might things change in the medium and long-term? First, techniques in synthetic biology may develop in ways which allow biotechnologists to create agents that are not just functionally very close to humans, but close in detailed neural and physiological makeup. In that, *ex hypothesi*, such creatures will share deep and extensive similarities with us in terms of the biological underpinnings of consciousness, we would have little ground for denying that they are 'objectively' conscious,

CSS-bearing, beings, with all the ethical consequences that would flow.

In the context of our present discussion, such full-spec synthetic biology-based creatures are offer little challenge as compared to those which occupy the less bio-realistic regions on the spectrum of possible artificial creatures – where judgment calls about CSS properties and ethical status on the basis of physical design are much less straightforward. In many regions of the spectrum there will be agents with natural, fluent and subtle social interactivity characteristics that are close to those of humans, but where underlying physical design is remote from human physical design. These do and will increasingly offer the most difficult sorts of cases: agents that, via their fluent social capacities (and, in many kinds of such case, outward humanoid bodily features), display a wide variety of CSS-evincing behaviours but where they share relatively few of the internal neurological features that make for CSS features in humans. These are the cases where the social-relational view is on its most solid ground.

8 FALSE POSITIVES, FALSE NEGATIVES

But, realists will say, for such cases there could be risks of false positives and false negatives in CSS-attributions. And surely it is the significance of such false positives and negatives that makes a difference – both in theoretical terms and in moral terms. In the hypothetical situations where such agents exist in large numbers – where they multiply across the world as smartphones have done today – wrong judgments could have catastrophic ethical implications. In the false-positive case, many resources useful for fulfilling human need might be squandered on satisfying apparent but illusory 'needs' of vast populations of benaviourally convincing but CSS-negative artificial agents. Conversely, in the false-negative case, vast populations of CSS-positive artificial agents may undergo extremes of injustice and suffering at the hands of humans that wrongly take them for socially-fluent zombies.

Given the spectrum of hypothetical types of cases previously mentioned, the social-relationist must decide where, if anywhere, on this spectrum, any kind of objectivity of CSS-attribution is allowed. At one extreme lie actual humans and near-human bio-realistic artificial replicants. At the other extreme lie the simplistic AI systems of today. Apart from philosophical concerns about the 'other minds' problem, already shown to be deeply flawed, what reason is there to deny the objectivity of (positive) CSS-attributions in the case of ourselves? And as for synthetic-biology beings which are physiologically close to us, surely that physiological commonality supports objective positive CSS-attributions as well. And, at the other end, where the relatively crude electronic and robotic agents of today lie, we surely have enough basis for making 'objectively' warranted negative CSS-attributions. So, for such cases at least, the realist case seems to be strong. So how will the social-relationist say where, in the spectrum, the presence or absence of CSS features ceases to be an objective or realist matter and where it thus becomes more appropriate to apply a social-relationist perspective?

9 CONCLUSION

Singer's notion of the expanding ethical circle, and Harris's suggestion that ethical questions concerning the 'moral landscape' can be scientifically grounded, suggest, in different ways, a very strong linkage – possibly a conceptual one – between consciousness and well-being (CSS properties) and ethical concern. In particular, Harris's critique of scientific neutralism suggests the possibility of a scientific grounding to core ethical values: and there is no reason why such scientific, objective grounding should not also apply to the ethical status of artificial agents.

Of course our ethical relations with such agents will be inevitably bound up with our social relations with them. As we saw, the domain of the social is expanding rapidly to include a wide variety of human-AI and AI-AI interactions. But sociality is itself constrained in various ways by physical, biological and psychological factors. And consciousness and well-being lie at the heart of these constraints. Ethics and sociality are indeed closely intertwined. But we should not assume that, just because there are rich and varied social interactions between humans and artificial creatures of different sorts, there are no considerations about the appropriateness of ethical relations that humans may adopt towards such artificial creatures. Our capacities for satisfaction or suffering must be crucially based upon deep neural and biological properties; so too for other naturally evolved sentient creatures. Some artificial creatures will have closely similar biological properties, making the question of CSS-attribution relatively easy for them at least. For others (ones whose designs are advanced versions of electronic technologies with which we are familiar today, for example; or which are based on other technologies that we currently have only the faintest idea of) it may be much harder to make dependable judgments. In the end, how we attribute CSS and consequently ethical status will depend on a multiplicity of detailed questions concerning commonalities and contrasts between human neural and bodily systems and analogous systems in the artificial agents under consideration. The gross apparent behaviours of such agents will play a role, of course, but

only in a wider mix of considerations which will include these other, less easily observable, features.

Over- and under-attribution of CSS-properties cause deep ethical problems in human social life. (To take just one obvious and widespread kind of case: oppressed humans all over the globe continue to have their capacity for suffering falsely denied, in fake justification for their brutal treatment.) Why should it be any different for robots? In a society where humans and machines have extensive and rich social interactions, either false positive or false negative mis-attributions could engender massive injustices – either to humans whose interests are being short-changed by the inappropriate shifting of resources or concern to artificial agents that have no intrinsic ethical requirements for them; or to artificial agents whose interests are being denied because of a failure to correctly identify their real capacities for experience and suffering. It is not clear how a social-relational view can properly accommodate this false-positive/false-negative dimension.

I have tried to put the realist position in a way that is sensitive to the social-relational perspective. However many problems and gaps remain.[11] A strength of the social-relational position is that it addresses, in a way that it is difficult for the realist position to do, the undoubted tendency for people to humanize or anthropomorphize autonomous agents, something that will no doubt become more and more prevalent as AI agents with human-like characteristics proliferate, and which happens even when it is far from clear that any consciousness or sentience can exist in such agents. There will surely be strong social pressures to integrate such AIs into our social fabric. Supporters of singularitarian views even insist that such agents will come (disarmingly rapidly, perhaps) to dominate human social existence, or at least transform it out of all recognition – for good or for ill. Possibly such predictions sit better with the social-relational view than with the realist view, so it will be a big challenge for realism to respond adequately to the changing shape of human-machine society. In any case it is important to become clear on how the AI research community should best respond to the difficulties that such future social pressures may present.

Acknowledgements: Work on this paper was assisted by grants from the EUCogII network, in collaboration with Mark Coeckelbergh, to whom I express gratitude. Ideas in the present paper have also derived great benefit from discussions with Rob Clowes, Ron Chrisley, Denis Roche, Wendell Wallach and Blay Whitby.

[11] One gap is that I have dealt primarily with the connection between consciousness and artificial moral patiency, or recipiency, as opposed to moral agency, or productivity. However, I have elsewhere discussed some arguments that suggest that consciousness may be as crucial to the former as to the latter, so will not deal with that further here [2,3].

REFERENCES

[1] W. Wallach, C. Allen, and S. Franklin (2011), 'Consciousness and ethics: artificially conscious moral agents', *International Journal of Machine Consciousness,* 3(1), 177-192.

[2] S.B. Torrance (2008) 'Ethics and consciousness in artificial agents', *Artificial Intelligence and Society.* 22(4). 495-521

[3] S.B. Torrance. and D. Roche (2011) 'Does an artificial agent need to be conscious to have ethical status?', in B. van den Berg and L. Klaming (eds) *Technologies on the Stand: Legal and Ethical Questions in Neuroscience and Robotics*, Nijmegen: Wolf Legal Publishers, 285-310.

[4] S.B. Torrance (2011) 'Would a super-intelligent AI necessarily be (super-)conscious?' *Proc. Machine Consciousness Symposium, AISB-2011.* University of York

[5] S.B. Torrance (2012) 'Artificial agents and the expanding ethical circle' *AI & Society,* DOI: 10.1007/s00146-012-0422-2

[6] M. Coeckelbergh (2012) 'Who cares about robots? A phenomenological approach to the moral status of autonomous machines', this *Proceedings*.

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

[8] M. Coeckelbergh (2010) 'Moral appearances: emotions, robots, and human morality'. *Ethics and Information Technology*, 12(3): 235-241

[9] M. Coeckelbergh (2009). 'Personal robots, appearance, and human good: a methodological reflection on roboethics.' *International Journal of Social Robotics* 1(3), 217–221.

[10] M. Coeckelbergh (2012) *Growing moral relations: a critique of moral status ascription*, Basingstoke: Palgrave Macmillan, (forthcoming)

[11] D. Levy (2009). 'The ethical treatment of artificially conscious robots'. *International Journal of Social Robotics*, 1(3): 209-216.

[12] D.J. Gunkel (2012) 'A vindication of the rights of robots', this *Proceedings*.

[13] D.J. Gunkel (2012) *The machine question: critical perspectives on AI, robots and ethics.* Cambridge, MA: MIT Press,

[14] P. Singer (2011) *The expanding circle: ethics, evolution and moral progress,* Princeton University Press

[15] A. Leopold (1948) 'A land ethic'. In: *A sand county almanac with essays on conservation from round river*, NY: Oxford University Press.

[16] A. Naess (1973) 'The shallow and the deep long-range ecology movements'. *Inquiry* 16:95–100

[17] S. Harris (2010): *The moral landscape: how science can determine human values.* London: Random House.

[18] T. Bisson (1991). 'They're made out of meat', *Omni,* 4.

[19] S. Gallagher & D. Zahavi (2008) *The phenomenological mind: an introduction to philosophy of mind and cognitive science.* London: Taylor & Francis.

[20] L. Wittgenstein (1953). *Philosophical investigations.* Oxford: Blackwell.

[21] A.M. Turing (1950). 'Computing machinery and intelligence'. *Mind* 59: 433-460.

[22] E. Thompson (2001) 'Empathy and consciousness' Journal of Consciousness Studies 8(5-7): 1-32

[23] T. Regan (1983) *The case for animal rights*. Berkeley: University of California Press.

Can an unmanned drone be a moral agent? Ethics and accountability in military robotics.

Rodger Kibble[1]

Abstract. Remotely operated Unmanned Aerial Systems (UAS) or "drones" are now routinely deployed in theatres of war and are capable of lethal acts such as firing missiles under the control of their human operators. It would take a small technological step but a large legal and ethical one to allow them to make "kill" decisions autonomously. This paper outlines some general technical and ethical contexts surrounding the use of these weapons and examines a specific proposal for implementing ethical constraints on UAS, Arkin's "ethical governor". It is argued that the proposal is flawed in several respects: the author fails to support his bold claim that robots are capable of acting more ethically than humans, there is a lack of clarity in the formal representations of ethical constraints, and the metaphor of a "governor" is a misleading characterisation of the proposed system's functionality (as argued elsewhere by Matthias).

1. INTRODUCTION

The increasing automation of warfare poses pressing challenges for researchers in AI and machine ethics, given that semi-automated weapon systems such as "drones" are now routinely in use under a system of military law and ethics which has hardly been updated since the 1940s [14]. US doctrine already allows remotely piloted drones or Unmanned Air Systems (UAS) to fire on manned radar stations, even if they have not been directly attacked [Ibid]. If UAS are ever to be given the capability to make independent "kill" decisions, it is surely essential that they should operate to at least the highest standards that would be required of a human professional. The UK currently has no intention of moving to full autonomy [16] and a recent British study concludes that "fully autonomous weaponized systems may never be acceptable" [8]. However, researchers in other countries, particularly the US, have concluded that it is only a matter of time before autonomous armed robots are deployed, and that the task for roboticists is to ensure that these robots behave in an "ethical" manner [1,3]. The framers of the UK EPSRC Principles of Robotics [7,4] were "concerned to ban the creation and use of autonomous robots as weapons" but "pragmatically acknowledged that this is already happening in the context of the military" [4]. Perhaps the most prominent researcher in this field, Ronald Arkin of the Georgia Institute of Technology, makes the bold and startling claim that robots have the potential to be more "humane" than human soldiers [3].

The bulk of this paper will be devoted to a critical examination of Arkin's arguments and proposals, including consideration of a trenchant critique by [12]. According to [18], work in robotic ethics needs to give due attention to three separate areas of moral/philosophical concerns, formal specification and reasoning about moral constraints, and the implementation task of "designing a working system in which there are rules that can be enforced and deviant behaviour be detected" – corresponding to the software engineering activities of requirements analysis, formal specification, and design/implementation. Arkin follows a similar pattern in his presentation [1] and we will do likewise in this paper. Section 2 looks at general ethical issues raised by military robotics in the context of the Just War tradition, international law and other related factors. Section 3 addresses Arkin's moral and philosophical arguments for the potential ethical superiority of robots over human soldiers. Section 4 considers Arkin's proposals for implementing ethical considerations in a formal reasoning system. Section 5 deals with architectural issues, specifically the notion of an "ethical governor" which is intended to suppress any lethal actions which would be in breach of applicable rules. Section 6 concludes the paper with some general discussion.

2. ETHICAL CONTEXT

The notion of a Just War in Western societies originates from the era of Emperor Constantine who, having adopted Christianity as the state religion, no doubt found it embarrassing to be hampered by the tenets of a faith that ostensibly abjured violence [17]. The Church was tasked with devising a set of rules that would enable the Roman Empire to wage war with a clear conscience. Key contributions to the doctrine were made by Augustine of Hippo and (later) Thomas Aquinas, and their requirements can be summarised as:

1. Legitimate authority
2. Just cause
3. Right intention.

In the 20th century, developments in Just War doctrine and international law were stimulated by the experience of the two world wars, the advent of the League of Nations and the United Nations, and particular judgments at the Nuremberg war crimes trials. This led to a distinction between *Jus ad Bellum* – when it is right to go to war – and *Jus in Bello*, governing how war should be fought. Key principles of *Jus in Bello*, codified in the Geneva Conventions, are *proportionality* and *discrimination* – the latter meaning that any attack should avoid harm to non-combatants. Common Article 3 of the Geneva Conventions

> …reflects the most fundamental of Geneva's rules, that *anyone* who was not taking an active part in hostilities must be treated humanely, *in all circumstances.* [15] (emphasis in original)

We may note that research in military robotic ethics including Arkin's tends to be confined to *Jus in Bello*; ethical robots are not assumed to be capable to taking a view on the legitimacy of

1. Goldsmiths College, UK. email: r.kibble@gold.ac.uk

any conflict they are deployed in, even though such concerns have caused human citizens to register as conscientious objectors, desert, evade the draft or refuse conscription in conflicts all over the world. The general body of international law governing the conduct of warfare is generally referred to as the Law of Armed Conflict (LOAC) which are supplemented in any particular campaign by specific Rules of Engagement (ROE) [16].

Discussions of machine ethics traditionally begin with a consideration of Isaac Asimov's famous Three Laws of Robotics, as expounded in his series of Robot stories (the particularly well-read sometimes point out that there is a fourth, or "zeroth" law). These laws are given short shrift by both [7] and [14], the latter pointing out that they are particularly inappropriate for military applications. As a reminder, the Three Laws provide that:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

(http://en.wikipedia.org/wiki/Three_Laws_of_Robotics)

[7] point out the inherent impracticality of these laws:

> For example, how can a robot know all the possible ways a human might come to harm? How can a robot understand and obey all human orders, when even people get confused about what instructions mean?

And [14] succinctly observes that

> …much of the funding for robotic research comes from the military. It explicitly wants robots that can kill, won't take orders from just any human, and don't care about their own lives. So much for Laws One, Two and Three.

A crucial issue here is *responsibility*: who will be court-martialled or prosecuted if, for example, an autonomous UAS mistakenly fires on a wedding party? [7] is quite clear that robots should never be regarded as more than tools, and it is the humans who deploy them that must be held as responsible and accountable. [16] states that responsibility for an agent's actions lies with the last person to issue an order, while "responsibilities of the designers will have been discharged once the UAS has been certified by the relevant national military or civilian air authorities". This assumes that the subsequent behaviour of the artefact is *predictable*, even in the event of a break in the communications link [8]. [1] proposes a system that encodes legal and ethical principles as axioms of modal logic and argues that in contrast to sub-symbolic learning processes, the behaviour of such a system would be fully predictable from a given set of inputs – see section 4 below. The difficulty here is that the agent's environment may be rapidly changing in a battlefield scenario, and indeed its own actions may alter the environment in a non-deterministic way, leading to a loss of predictability. It is obviously critical that the agent is able to interpret its environment accurately, correctly identifying appropriate military targets. [16] states that

> For long-endurance missions engaged in complex scenarios, the authorised entity that holds legal responsibility will be required to exercise some level of supervision throughout

- which somewhat qualifies the notion of "agent autonomy". [8] discuss in some detail the place of autonomous systems in a chain of command and as noted above, conclude that "fully autonomous weaponised systems may never be acceptable".

3. HUMANS vs ROBOTS

It is claimed that battlefield robots could be programmed to behave "more ethically" than humans, reducing the occurrence of atrocities [3]. Artificial agents are not prone to anger and other "human factors" which commonly lead to violations of ROEs or the LOAC. One refinement of this claim [1] is that military law and ethics have been precisely codified and interpreted so that agents can be equipped with a formally specified knowledge base of legal and ethical principles, and moral decision-making within this circumscribed context would be a task they could perform more efficiently and consistently than we do. The next section considers the efficacy of Arkin's proposed ethical calculus; here we examine in some detail his startling claim that robots may be capable of treating us more humanely than we treat each other, as presented in [3].

Arkin's case essentially boils down to:

1. War will continue
2. Human soldiers cannot be trusted to behave honourably on the battlefield
3. Therefore they should be replaced with robots.

Point (1) is presumably outside the author's competence as a roboticist. Many of his arguments in support of (2) turn out to be weak, disingenuous and even contradictory. I will examine some key issues in detail.

The central claim is that autonomous systems are capable in principle of performing "more ethically than human soldiers are capable of performing", presented with the rhetorical question

> As robots are already faster, stronger and in certain cases (e.g., chess playing) smarter than humans, is it that difficult to believe that they will be able to treat us more humanely on the battlefield than we do each other?

I have two fundamental objections to this stance. Firstly, there are clear objective standards for assessing the capabilities of robots in these circumscribed areas: their speed, durability and weight-lifting abilities can be quantitatively measured, while anyone familiar with the rules of chess and able to distinguish legal from illegal states of the board can determine whether a computer has fairly defeated a human. It is rather a leap of faith to extend this reasoning to the field of ethical behaviour, where no such cut-and-dried metrics are available. And secondly,

ethics is a human construct and it is hard to see how a computational artefact could be said to act "more ethically" than humans. This would imply some kind of super-human vantage point from which our actions and those of our artefacts could be compared. If an autonomous system applies different standards to its ethical choices than those a human would follow, then it may become impossible for a human to predict the machine's actions, and we have observed above that this is a prerequisite for proper accountability.

Some of the supposed advantages of autonomous systems and the concomitant human failings are:

1. Autonomous armed systems "do not need to protect themselves" and "can be used in a self-sacrificing manner … by a commanding officer". So for instance, rather than having troops storm into a suspected safe-house with guns blazing, a robot could be sent in first to see if it draws hostile fire. **Remark:** Indeed, but this capability is surely equally applicable to remotely guided, unmanned systems, and/or need not require the robot to be weaponised.
2. "The eventual deployment and use of a broad range of robotic sensors better equipped for battlefield observations than humans currently possess". **Remark:** this does not seem an argument for removing people from the loop, as presumably the outputs of these sensors could be made available to a system's human controller – just as warfighters already make use of such enhanced sensors as radar, sonar, infra-red rifle sights and so on.
3. Avoidance of "scenario fulfilment" where people in stressful situations ignore information that contradicts their belief-set, "a factor believed partly contributing to the downing of an Iranian Airliner by the *USS Vincennes* in 1988". **Remark:** This precise incident is adduced by [14] as an example of the *danger* of relying on computers and automated systems. The *Vincennes* was the only ship in the area carrying the missile guidance system Aegis, which was capable of running in various mode from fully supervised ("Semiautomatic") to fully automated ("Casualty") and was supposed to be able to detect the threat level of aircraft and missiles in the vicinity. Because of this, the Vincennes was authorised to fire without consulting more senior officers in the fleet. What happened was that Aegis wrongly identified and targetted the airliner as an Iranian F-14 fighter, and the command crew authorised it to fire. If the system had been functioning fully autonomously the outcome would surely have been the same.
4. Autonomous systems "have the potential capability of independently and objectively monitoring ethical behaviour in the battlefield and reporting infractions". **Remark:** Again, this is not an argument for equipping robots with lethal capabilities: this function could equally be carried out by systems that have no more than observer status. Indeed, current communication and surveillance technology already enables base commanders to observe operations in real-time [Singer].
5. A report from the Surgeon General's Mental Health Advisory Team found that "Well over a third of Soldiers and Marines reported torture should be allowed, whether to save the life of a fellow Soldier or Marine or to obtain important information about insurgents". **Remark:** It is perhaps not surprising that serving military personnel should share the views of their Commander-in Chief, the Vice President or the US Defense Secretary, along with opinion-performers such as Harvard Law professor Alan Dershowitz. In February 2002 President George W Bush decided that suspected Taliban and al-Qaeda members detained at Guantanamo Bay were not entitled to the protection of the Geneva Conventions, with the specific consequence that interrogations would not be bound by "long-standing practice that regarded the rules in Common Article 3 [of the Conventions] as a minimum that applied to everyone, in all conflicts" [15] and in particular prohibited abusive interrogation. In December 2002, Secretary Rumsfeld approved a request from General Counsel Jim Haynes for new interrogation techniques to be available at Guantanamo, including many which had been declared to constitute "torture" in a decision of the European Court of Human Rights against the UK. This eventually led to the adoption of the notorious practice of "water-boarding" which Vice-President Cheney was to describe as a "no-brainer" if it could save lives. In publications and interviews during 2001 and 2002, Professor Dershowitz argued for the issuance of "torture warrants" in specific circumstances such as the fabled "ticking-bomb" scenario. These ideas were also prevalent in popular culture thanks to the TV series *24*, whose hero Jack Bauer was regularly seen to carry out aggressive interrogation to obtain supposedly life-saving information.
   Against this background, the approval of torture by ground forces may not result from emotions running high under the stress of combat, but can be seen as influenced by policies which stemmed from the highest levels of the administration and had a substantial level of approval in the public sphere. Automation does not seem to offer a satisfactory solution, and in any case it is hardly plausible for interrogations to be carried out by autonomous systems. The instructions that had been in force up to February 2002 and encoded in *US Army Field Manual 34-52* prohibited all use of force and advised that the interrogator should aim to build a "rapport" with the subject. We are surely some way from the deployment of empathic robots which would be capable of establishing such a rapport. Rather than seeking a "technological fix" through automation, these abuses need to be combatted through democratic and judicial processes along with active debate in the public sphere – as indeed they have been [14, 5].
6. The same report found that "although they reported receiving ethical training, 28% of Soldiers and 31% of Marines reported facing ethical situations in which they did not know how to respond". **Remark:** Although the details of these situations are not discussed, it is surely intrinsic to the human condition that one may face ethical dilemmas which have no obvious or clear-cut resolution, and the fact that an automated system might be more decisive is no guarantee in itself that it will make the "correct" choice. Examples of such dilemmas have been explored by tragic dramatists since antiquity and have been discussed at length by philosophers, perhaps crystallised in the variants of the "trolley problem" which is generally ascribed to Philippa Foot. A typical formulation is: is it permissible to divert a runaway trolley so that it will hit and kill one person, rather

than stand by and let it hit and kill five? The issue is that in the first case, which would be preferred from a utilitarian or consequentialist point of view, the agent is actively killing a human being who would otherwise have survived. Thus the question arises of whether a deontic prohibition on deliberate killing may have less force than a consequentialist obligation to intervene to save life, depending on circumstantial factors. One can doubtless devise battlefield scenarios which present formally similar dilemmas. The claim that autonomous systems would be better at resolving them than human soldiers relies on the assumption that they would be equipped with an ethical calculus which, if not infallible, would have a high probability of choosing correctly in situations where human soldiers do not know how to act. This would presumably have to cover cases where the LOAC/ROE give no clear guidance; it seems doubtful that one will find answers to trolley problems in these documents. We return to this general topic in the next section; but looking ahead a little, it turns out that Arkin's robots might be equally indecisive, given that "in the absence of certainty … the system is forbidden from acting in a lethal manner" [2].

7. Servicemen and women on active duty suffer psychological trauma in increasing numbers, as has been observed from WWI to the present conflicts in Iraq and Afghanistan. This manifests itself as "battlefield fatigue, conversion hysteria, confusional states, anxiety states, obsession and compulsion states, and character disorders". [14] also notes a high degree of psychological stress even among remote operators of UAS, carrying out operations in South Asia and the Middle East from air-conditioned bases in Nevada. **Remark:** This in itself need not justify replacing troops with dispassionate robots, rather it could be seen as underscoring the imperative to avoid armed conflict wherever possible. [16] discusses "the school of thought that for war to be moral (as opposed to just legal) it must link the killing of enemies with an element of self-sacrifice, or at least risk to oneself". That is, the risk of harm to one's own forces imposes natural limits on the degree of force that will be used, an argument that apparently goes back to Clausewitz. If wars are to be prosecuted by emotionless automata with no fear for their own safety, no concept of suffering and no families to grieve for them, there may be the risk of "rapid escalation of what would previously have been considered a simple diplomatic problem to full-on technological warfare" [16] – the 21st century equivalent of "gunboat diplomacy". At the very least, there is less pressure to find alternative methods of resolving conflicts once one has embarked on a military path. [16] suggest that we may already have reached a stage where "the use of force is totally a function of an unmanned capability" in the use of UAS over Pakistan and Yemen. Peter W. Singer pointed out at a hearing in March 2010 that

   > Our unmanned systems have carried out 119 known air strikes into Pakistan, which is about triple the number we did with manned bombers in the opening round of the Kosovo War. But Congress has not had a debate about whether to authorize or disapprove of it. [6]

   [14] and [16] also note the effects on perceptions among the "enemy" of reliance on long-distance weaponry, with remote pilots regarded as "cowardly" and dishonourable.

8. Operators of remotely piloted UAS become distanced from the consequences of their actions and "pretend they are not killing human beings". **Remark:** this is indeed a widely encountered argument against the use of UAS but few proponents take the same further step as Arkin, which is to argue that control of UAS should be taken away from humans and handed over to computers.

9. [3] and [14] both observe that it is very hard to train "true warfighters" and most combatants show an extreme resistance to killing: during WWII, most fighter pilots never tried to shoot anyone down while in one study, "only 15% of the men had actually fired at enemy positions". **Remark:** This is apparently presented as a failing of human combatants who may lack an "offensive spirit". Confusingly, it is also suggested that reluctance to engage with enemy forces stems from "the use of long-distance weapons making battlefields `lonely' and the feeling that the enemy was not real but a phantom". This does rather seem to contradict the idea that long-distance weapons are likely to *increase* the inappropriate use of lethal force. The deployment of robots lacking any kind of empathic identification with hostile troops would surely transform the nature of battles and so should not be embarked on without extensive deliberation. For instance, it is quite plausible that those who silently opt out of hostilities represent a body of opinion among the citizenry who consider the cause they are engaged in to be unjust or illegitimate, or at least that it does not justify taking human life – particularly if they are conscripts. This constituency would be effectively disenfranchised by automation.

Summary: Arkin marshals a variety of arguments in favour of deploying autonomous weaponised robots based on the supposed shortcomings of human soldiers; but as we have argued, he mostly fails to make a convincing case for removing humans from the loop altogether in cases where the robots would have lethal capabilities. Rather, he seems to be proposing automation as a technological fix for complex human, social and political problems, and perhaps to be making a virtue of necessity given the mounting evidence that weaponised autonomous systems are almost certain to be deployed sooner or later by armed forces in several parts of the world.

If we disregard the arguments from human failings, it becomes necessary for proponents of automation to show that robots can not only reason ethically at the same level as the "best" exemplars of human soldiers, but can do so faster and more consistently to the extent that their deployment can be justified. This is the topic of the next section.

4. ENCODING ETHICAL PRINCIPLES

[2] sets out a framework for representing and reasoning about ethical constraints in a battlefield scenario. As [12] notes, there appears to be some equivocation in that while the system is presented as an "ethical governor", the implemented constraints are limited to legalistic principles that might be derived from LOAC and hypothetical instances of ROE. In fact [2] expresses the questionable view that "battlefield ethics are more clear-cut and precise than everyday or professional ethics". This falls short of a general moral stance and may be insufficient to resolve

ethical dilemmas where LOAC provide no definitive guidance. [12] further points out that ROE are by no means ethical rules that can claim universal assent, but are instructions given by one party in a conflict to its own forces.

[2] sets out some specific problems which face the attempt to encode ethical principles:

> 1. The laws, codes, or principles (i.e., rules) are almost always provided in a highly conceptual, abstract level.
> 2. The conditions, premises or clauses are not precise, are subject to interpretation, and may have different meanings in different contexts.
> 3. The actions or conclusions in the rules are often abstract as well, so even if the rule is known to apply the ethically appropriate action may be difficult to execute due to its vagueness.
> 4. The abstract rules often conflict with each other in specific situations. If more than one rule applies it is not often clear how to resolve the conflict.

In addition to these factors, [8] note that "interpretations of international law evolve over time", that members of military coalitions may have different ROEs for their own forces, and that local populations may have different cultural values and legal frameworks. They further observe that judgments of *proportionality* are subjective and typically involve military experience, legal expertise and local knowledge, and that "no-one has tried to identify which information in a typical set of ROEs can be quantified and the tolerances on the resulting values."

These considerations are said to rule out the use of "predicate logic and other standard logics based on deductive reasoning" because "they operate from inference and deduction, not the notion of obligation" [2]. This is a little confusing, as later in the same paper Arkin presents epistemic and deontic modal logics as suitable formalisms for representing and reasoning about ethical constraints, encoding general principles such as that lethal acts should only be committed if *obligated* by the ROE and *permitted* by LOAC. In fact deontic logic handles the notion of obligation precisely through deductive reasoning; standard modal logics are built on a foundation of Boolean propositional calculus, with formal semantics and axiomatic proof theories, and are no more tolerant of vagueness or inconsistency than are classical logics [9]. And the specimen axioms that are presented actually include formulas of predicate logic, quantifying over sets of constraints which define permitted and forbidden actions.

5. ARCHITECTURE

The notion of an "ethical governor" is introduced by analogy with Watts' mechanical governor which was designed to prevent steam engines overheating, essentially implementing a simple negative feedback loop. [12] argues that this analogy misrepresents the actual design of Arkin's system. A key point is that the mechanical governor will never operate in a way that contradicts the interest of a machine's owner: if the owner were to be rash enough to disable the governor, the machine would be liable to break down or perhaps explode. With military systems the reverse is the case: the so-called "governor" will not stop the system malfunctioning, rather it is intended to suppress lethal actions which are forbidden by the built-in constraints. The overall architecture is generate-and-test, with a tactical subsystem proposing actions which are filtered by the "ethical" module. [12] claims that this could be perceived as a source of inefficiency by military operators, in that they are prevented from attaining what they perceive as desirable goals. Arkin would have trouble disputing this claim, since part of his justification for the deployment of autonomous systems is that human soldiers are prone to commit atrocities or mistreatment and that their commanding officers are often complicit [3].
More charitably, it is quite plausible that a mission commander might want to disable the governor because of a disagreement with its interpretation of the LOAC or ROE, considering that it was acting too cautiously. And it seems that it would be relatively straightforward for a commander in the field to "turn off" an ethical governor: the "proportionality" of an attack is calculated by comparing predicted "collateral damage" with military necessity, and at a certain level of "necessity" "every desired action can be carried out without any interference from the ethical governor" [12]. It is unlikely that a robot would be competent or permitted to make autonomous judgments of military necessity, and so this parameter would doubtless be under the control of a human commander. [12] notes that Arkin's project has been reported in the popular press in a way which suggests military robots already have moral sensibilities or are capable of learning them, which may have the unwelcome consequence of legitimising the use of this class of weapons in the public sphere.

A further consideration is that the built-in constraints have to go through two levels of "translation": first from English (or the local language) into a knowledge representation formalism, and subsequently into computer software ("code"). The first step is problematic in itself, as mapping from English into modal logics is by no means straightforward. [9] gives examples of paradoxes arising from the fact that certain valid formulas of deontic logic don't make intuitive sense when rendered into English: for instance

$$Op \rightarrow O(p \vee q)$$

is a valid formula of a variant of deontic logic, meaning "if it is obligatory to do p, it is obligatory to do p or q". Yet this does not seem intuitively correct, since it could be instantiated as "If I must hand in my essay, then I must either hand it in or burn it". Regarding the second stage, [12] notes that while LOAC and ROE are generally public documents, their encoded versions will not be and indeed will most likely be subject to military secrecy. Thus there will be no proper opportunity for open democratic scrutiny and debate over the circumstances under which autonomous systems may be deployed and use lethal force.

6. CONCLUSIONS

Having challenged the notion that robots could be "more ethical" than humans and arguing that Arkin's proposed formalism and architecture are inadequate to the task of endowing robots with genuinely ethical judgment, I would like to raise the bar further by proposing that in order to be recognised as ethical subjects, agents would need to have communicative and argumentative capabilities in the sense of [10].

Rather than being more ethical than any human, we would want agents to perform at the level of the best exemplars of human behaviour. Discussions of the ethics of robotic warfare invariably cite the My Lai Massacre of March 16th, 1968, when a platoon under the command of Lieutenant William Calley embarked on a killing spree in a Vietnamese village that left around 500 civilians dead, including very young children. An important player on March 16th who is less often mentioned was Chief Warrant Officer Hugh Thompson. Having observed the large number of bodies from his gunship, Thompson reported the killings to brigade headquarters, ordered Calley to cease firing, put himself between the US troops and the villagers, and used his helicopter to ferry some of the villagers to safety [13].

I suggest we would want agents to be not only more ethical than Calley, but at least as ethical as Thompson. Heath (2001), following Durkheim, defines a norm-conformant agent as one that will not only follow rules itself, but is disposed to sanction those who do not. The LOAC and ROEs standardly require combatants to disobey or question orders which they consider unlawful, to report atrocities committed by other soldiers and where possible, to intervene to prevent them. Thus a fully autonomous and ethical military robot would not only be required to base its own actions on sound moral principles, and to justify them if challenged, but to constantly monitor both human and robotic agents in its vicinity to ensure that their actions were lawful, and to directly engage with other agents to dissuade them from unlawful acts.

In summary though, it is questionable whether researchers in AI, robotics and ethics should be devoting their time and expertise to developing systems which are intended to hasten the advent of autonomous military robots: they might be better occupied in arguing in the public sphere *against* the development of such systems, and applying their research efforts to developing systems which might facilitate non-violent methods of conflict resolution.

REFERENCES

[1] R.C. Arkin, 2007. Governing Lethal Behaviour: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture. Technical Report GIT-GVU-07-11, Mobile Robot Laboratory, College of Computing, Georgia Institute of Technology.

[2] R.C. Arkin, 2008. "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture - Part III: Representational and Architectural Considerations", *Proceedings of Technology in Wartime Conference*, Palo Alto, CA, January 2008.

[3] R.C. Arkin, 2010. The Case for Ethical Autonomy in Unmanned Systems, Journal of Military Ethics vol 9 no 4, 332-341.

[4] J. Bryson, 2012. The making of the EPSRC Principles of Robotics. AISB Quarterly 133, January 2012, 14 – 15.

[5] J. Butler, 2010. Frames of War. Verso.

[6] US Congress, 2010. Rise of the Drones: unmanned systems and the future of war. Hearing of the subcommittee on national security and foreign affairs, US Congress, March 23rd 2010. US Government Printing Office, Washington DC.

[7] EPSRC, 2011. Principles of Robotics: regulating robots in the real world. http://www.epsrc.ac.uk/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx, last visited May 15th 2012.

[8] T. Gillespie and R. West, 2010. Requirements for Autonomous Unmanned Air Systems Set by Legal Issues. The International C2 Journal 4.2, 2010.

[9] R. Girle, 2000. Modal logics and philosophy. Acumen Publishing Ltd.

[10] J. Habermas. 1996. Some further clarifications of the concept of communicative rationality. In M. Cooke (ed.), On the Pragmatics of Communication, Polity, 1999.

[11] J. Heath, 2001. Communicative action and rational choice. MIT Press.

[12] A. Matthias, 2011. Is the concept of an ethical governor philosophically sound? Proceedings of TILTing Perspectives 2011: "Technologies on the stand: legal and ethical questions in neuroscience and robotics," Tilburg University, Netherlands, April 11-12, 2011

[13] D. Linder, 1999. An introduction to the My Lai Courts-Martial. http://law2.umkc.edu/faculty/projects/ftrials/mylai/Myl_intro.html (visited 5th Feb 2012)

[14] P.W. Singer, 2009. Wired for War: The Robotics Revolution and Conflict in the 21st Century. The Penguin Press.

[15] P. Sands, 2009. Torture team: uncovering war crimes in the land of the free. Penguin Books.

[16] UK MoD 2011. Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems. UK Ministry of Defence, 30 March 2011.

[17] P. Vardy and P. Grosch, 1999. The puzzle of ethics. Fount/Harper Collins.

[18] V. Wiegel and J van den Berg, 2009. Combining Moral Theory, Modal Logic and MAS to Create Well-Behaving Artificial Agents. Int J Soc Robot (2009) 1: 233–242

Bridging the Responsibility Gap in Automated Warfare

Marc Champagne[1] **and Ryan Tonkens**[2]

Abstract. Robert Sparrow argues that military robots capable of making their own decisions would be independent enough to allow us denial for their actions, yet too unlike us to be the targets of meaningful blame or praise—thereby fostering what Matthias has dubbed "the responsibility gap." We agree with Sparrow that someone must be held responsible for all actions taken in a military conflict. That said, we think Sparrow overlooks the possibility of what we might term "blank check" responsibility. A person of sufficiently high standing could accept responsibility for the actions of autonomous robotic devices—even if that person could not be causally linked to those actions besides this prior agreement. The basic intuition behind our proposal is that we can impute relations even when no other form of contact can be established. The missed alternative we want to highlight, then, would consist in an exchange: social prestige in the occupation of a given office would come at the price of signing away part of one's freedoms to a contingent and unpredictable future guided by another (in this case, artificial) agency.

1 INTRODUCTION

If a robot capable of setting its own goals were to go on a killing spree, who would we blame? Or, if such a robot were to exercise its autonomy in a manner inconsistent with our understanding of morally permissible behaviour, who could we justifiably hold responsible? The possibility of creating a robot with the ability to make decisions based on the basis of its own initiative(s) rather than pre-programmed commands is admittedly a speculative idea at this stage. One can nevertheless imagine, as Robert Sparrow [1] does, a scenario where such military weaponry is somehow capable of making its own decisions without direct or indirect instructions from a human being. Deploying these kinds of autonomous robots in the theatre of war would have far-ranging ethical consequences because, in such contexts, not only are we dealing with life and death situations, we also expect the various parties to be held responsible for the choices that they make. As Michael Walzer [2] put it, "[i]f there are recognizable war crimes, there must be recognizable criminals."

However, since the hypothetical robots that interest Sparrow would be midway between us and plain machines, they would slip by this requirement of moral responsibility: They would be independent enough to allow humans plausible denial for their actions, yet too unlike us to be the targets of meaningful blame or praise. Such autonomous yet unfeeling robots would therefore inhabit a morally ambiguous space—what Matthias [3] has aptly dubbed "the responsibility gap." Any unethical act on their part would therefore betoken the most senseless event conceivable, a moral failure for which no one could be fairly held responsible. Understandably, then, Sparrow thinks this would be enough to bar the use of autonomous "killer robots" in war.

We follow Sparrow in believing that one condition for being an agent that can be held morally responsible is the capacity to be punished/rewarded through meaningful proportional suffering/remuneration. Although other conditions exist (e.g. wilful intent), falling within the ambit of punishment and reward is arguably a *necessary* condition of moral responsibility, and thus, insofar as autonomous robots could not be punished in any meaningful way, they would not qualify as responsible agents. However, since these machines would be autonomous, they would be culpable, and since they act deliberately, they would also be liable. Hence, such robots would need to be held responsible for their actions, and *we* owe this responsibility to (at minimum) the parties in the contexts where the robots would be in use.

We thus agree with Sparrow that, following the received tenets of just war theory and international laws of war, someone must be held responsible for all actions taken in a military conflict. Yet, when Sparrow concludes from this that autonomous robots should not be deployed, we believe his argument proceeds a bit too hastily. Indeed, we want to suggest that there is a way to seal the responsibility vacuum Sparrow points to. Although we do not endorse the use of such military devices, we think Sparrow's argument overlooks the possibility of what one might term "blank check" responsibility: A person of sufficiently high military or political standing could accept responsibility for the actions (normal or abnormal) of all autonomous robotic devices—*even if that person could not be causally linked to those actions besides this prior agreement*.

The basic intuition behind our proposal is that two things can be related just in virtue of their being related. That is, we humans retain the liberty to simply *impute* relations even when no other form of contact can be established (this is what we do whenever we use a word, for example). In an ethical context, this sort of non-causal imputation is akin to what is routinely called scapegoating (i.e., deliberately singling out a particular person as the recipient of responsibility or blame). But, this possibility should not overshadow the fact that, when informed consent is present, the same mechanism can be used fairly. Indeed, consent can secure the sought-after component of responsibility through a person's very willingness to partake in a contractual agreement. As will be explained below, by willingly agreeing to the terms of the contract, the informed agent *imputes responsibility on herself* for the actions of the autonomous machine. The missed alternative we want to highlight, then, would essentially consist in an exchange: Social prestige in the occupation of a given office could come at the price of signing away part of one's freedoms to a contingent and unpredictable future guided by another (in this case, artificial) agency.

[1] Dept. of Philosophy, York Univ., Toronto, Canada. Email: gnosiology@hotmail.com

[2] Dept. of Philosophy, Centre for Practical Ethics, York Univ., Toronto, Canada. Email: tonkens@yorku.ca

We shall not be arguing that holding the office in question would be conditional on one's willingness to actually use killer robots. Our more prosaic point is that *if* such robots are deployed, then someone can willingly (and publicly) accept responsibility for whatever ensues. Obviously, this proposal leaves open the question of whether the antecedent of the conditional is or should be affirmed. Still, to the extent such agreements can be fairly and reasonably implemented, Sparrow's case needs tinkering. Obviously, in the interest of peace, it is to be hoped that a more robust case for peace will emerge from this critical engagement.

With that shared aim in mind, we will present our case in a fairly straightforward manner: we will first describe and motivate the problem, then we will articulate and defend our solution.

2 SPARROW'S DILEMMA

Sparrow recognizes the common criticism [4, 5, 6] that the advent of military robots may trivialize entry into war. His argument against the use of autonomous robots, though, is more subtle, and turns on a disturbing prospect of across-the-board moral blamelessness. Even if we grant the (debatable) cognitive scientific prognostication that it is possible for a non-sentient robot to be autonomous—in the sense not just of being able to determine which means best suit a given end, but in the stronger sense of actually determining the best ends to follow—we are left with an ethical dilemma. Specifically, the question arises whether there is any room left for the attribution of blame (and praise) where these robots are deployed. As Sparrow points out, if we are ever to conduct wars utilizing such technology and simultaneously maintain our standard ethical inclinations regarding responsibility for one's actions, we must ask ourselves who would be responsible in the event of a violation by an autonomous robot in the conduct of war. Sparrow's claim, in short, is that no one could be rightly held responsible for any atrocities perchance committed by the robots. Since such a situation of blamelessness would be morally unacceptable (and unjust), prohibition of their use in war seems to follow.

So stated, this problem is a general one, and in principle it arises any time a) some autonomous entity is responsible for its behaviour yet b) holding that entity responsible would be absurd. It just so happens that, with respect to adult human military personnel, those that are responsible for their actions can all be held responsible, as they meet the conditions for being the targets of meaningful praise or blame (among other things, perhaps). Although an autonomous robot should likewise be responsible for its actions, it cannot be held as such.

To make this worry vivid, consider a fully self-determining military machine equipped to identify the intention(s) of combatants. Imagine a scenario in which such a robot would lethally engage an enemy platoon that was clearly surrendering before and during its attack [1]. Such a situation would leave us with the following tension: To the extent that an "autonomous robot" fits its twofold label, humans cannot be blamed for its actions, since the robot has genuine "autonomy;" yet the robot itself cannot be blamed for its actions either, since it is merely "robotic" and is thus impervious/oblivious to any meaningful form of punishment. If we want to keep moral considerations in the picture and retain the idea that just war theory and the international laws of war require us to justifiably hold someone responsible when things go wrong, then the situation described by Sparrow becomes very problematic.

It is important to underscore that the moral predicament discussed by Sparrow arises only when the violent act of the robot is sandwiched between a narrow set of circumstances. Specifically, the robot must be sophisticated enough to make its own choices, but not so sophisticated that it can experience pain and pleasure. Indeed, it is worth recalling that the range of robots Sparrow's argument applies to is very restricted. All of the semi-autonomous or remote military robots currently in existence are irrelevant to his discussion, since in the use of these robots humans can at all times be held responsible, say, via their contribution to a program's content or a device's activation. Sparrow, by contrast, wants to call our attention to a far more difficult prospect. Hence, it is crucial to be clear on what exactly is meant by a fully "autonomous" robot in this context:

> Artificially intelligent weapon systems will thus be capable of making their own decisions, for instance, about their target, or their approach to their target, and of doing so in an 'intelligent' fashion. While they will be programmed to make decisions according to certain rules, in important circumstances their actions will *not* be predictable. However, this is not to say that they will be random either. Mere randomness provides no support for a claim to autonomy. Instead the actions of these machines will be based on reasons, but these reasons will be responsive to the internal states—'desires', 'beliefs' and 'values'—of the system itself. Moreover, these systems will have significant capacity to form and revise these beliefs themselves. They will even have the ability to learn from experience. [1]

The robots Sparrow envisions would thus have the autonomy to reject their programmed rules, or to apply those rules in ways that are not in line with judgements we would endorse. To the extent a robot would do this, it would think "outside *our* box." No matter how remote it may be, Sparrow wants to address this possibility directly, before it arises (a laudable and prudent pre-emptive attitude which, by engaging in earnest with the subject matter, we obviously share).

Accordingly, Sparrow is more concerned with guarding against a responsibility-vacuum than with arguing that limitations should be placed on the autonomy of machines. The moment artificial constraints are placed on a robot's range of conduct, be it at the level of hardware or software, that robot no longer falls within the ethically problematic range Sparrow is interested in. As an example of this, consider Arkin's [7] suggestion that autonomous lethal robotic systems be equipped with an in-built design component providing us with clear data regarding who is responsible for the robot's actions. This could include explanations for using or omitting lethal force, prior to deployment in a specific mission context. Such "a responsibility advisor" (as Arkin calls it) "makes explicit to the robot's commander the responsibilities and choices she is confronted with when deploying autonomous systems capable of lethality" [7]. Mock-guilt could even be made into a quantifiable variable influencing a robot's behaviour in a given military conflict. For instance, if a guilt-value associated with a given action exceeds a certain threshold, the robot could have a mechanism that reprimands it for that action [8].

Although such a buffer clearly is meant to remedy the violence robots may cause, a proposal like this does not truly address Sparrow's dilemma, insofar as reinstating straightforward human responsibility by shaving off a robot's leeway in the election of its ends is tantamount to dodging the philosophic issue at hand. Irrespective of its practical merits then, Arkin's proposal misses the point.

This inadvertent tendency to alter the scenario into something more manageable is ubiquitous. Recently, Lokhorst and van den Hoven [9] have challenged Sparrow's conclusion by arguing that there may be effective alternatives available for remedying unacceptable behaviour. Their criticism, however, stems from a misunderstanding of Sparrow's concerns. Sparrow is willing to (agnostically) grant Lokhorst and van den Hoven's claim that advancement in AI and robotics research could conceivably invest autonomous robots with the ability to suffer. Yet, these are not the sorts of robots that trouble him, since here we could presumably hold something responsible, namely the robot itself. Indeed, to the extent robots could suffer in the relevant way, they would become legitimate targets for genuine punishment, and so the quandary of moral blamelessness previously canvassed would not appear. Likewise, Lokhorst and van den Hoven's proposal to adjust a violent robot's programming so that it does not commit a similar atrocity in the future does not satisfy the demand that someone be held responsible for the previous misbehavior. Even if we could successfully tweak a robot's programming after we detect a flaw, the initial act that generated our *ex post facto* intervention would remain unpunished. To be sure, preventing similar atrocities from occurring again is a very important task. Yet, this still leaves us in a responsibility vacuum with respect to the original act—a situation whose resolution is *also* very important.

With these qualifications in mind, Sparrow contends that, if an agent is truly acting autonomously, then we are not justified in holding anyone responsible for the actions of that agent. To be sure, if a robot kills innocent civilians following its own internal states (beliefs, desires, and goals), our understandable reflex will be to blame the robot. Autonomous war machines, however, could not be blamed because, without a real capacity to suffer, they would lack a characteristic required to be the subject of meaningful punishment. Sparrow thus argues that, given the appropriate level of machine autonomy, any transgressions made by a robot give rise to an ethical catch-22. Since there are no suitable candidates for satisfying the *jus in bello* clause, all the available possibilities seem either irrational or unfair: If we blame a non-sentient machine, we are being irrational; and if we blame someone unrelated to the act, we are being unfair. Given that these two premises are acceptable, their troublesome conjunction demands our immediate philosophic attention.

Let us look at the first horn of this dilemma. Although they can certainly be destroyed or damaged, robots do not have anything of comparable moral worth at stake in the outcome of their actions, no welfare that could be compromised. As Sparrow notes,

> In order for any of these acts to serve as punishment they must evoke the right sort of response in their object […]. I shall assume that the most plausible accounts of the nature and justification of punishment require that those who are punished, or contemplate punishment, should suffer as a result. While we can imagine doing the things described above [i.e. fines in earnings, imprisonment, corporal or capital punishment] to a machine, it is hard to imagine it suffering as a result. [1]

This lack of an experiential dimension, Sparrow argues, essentially makes it futile for us to hold the machine responsible for its transgressions. The other horn of the dilemma is more straightforward still. Clearly, it is wrong to blame someone for actions she did not partake in or endorse. Once we grant that all robots and people are inadmissible, a tension naturally follows. Indeed, just war theory demands that someone be responsible for military actions (see for instance the customs of war described by the International Committee of the Red Cross). Since blaming someone at random would be wrong, developing and using autonomous military robots is morally unacceptable. That, at any rate, is what Sparrow argues. What we want to suggest now is that, when carefully used, consensual imputation effectively bridges the responsibility vacuum that concerns Sparrow.

3 A WAY OUT OF SPARROW'S DILEMMA

Here, in programmatic outline, is how we think the dilemma canvassed by Sparrow can be overcome. Should fully autonomous robotic agents be available to a military, the decision to deploy these robots could fall on the highest person in command, say, the president (or general, etc.). The president would be fully aware that the autonomous robots are capable of making their own decisions. Yet, a non-negotiable condition for accepting the position of president would be to accept blame (or praise) for whatever robotic acts perchance transpire in war.

Admittedly, the theatre of war is foggy. Yet, with the rigid imputation of blame secured, if the deployment of these machines renders the resulting war foggy to a degree which makes the president uncomfortable with accepting her surrogate responsibility for the robots' actions, then it follows that these robots should not be deployed. On the assumption that it would be in a democratically-elected official's best interest to wage an ethical war, he or she would likely ensure that these robots will abide by accepted international rules, otherwise she would have excellent reason *not* to allow those robots to be used. Yet, the force of our proposal lies in the fact that the requirement to accept "blank check" responsibility would hold even if the autonomy of the robot was shown to be unconstrained. To be sure, if the president would be unsure about whether these robots would indeed behave morally, then prudence would dictate that she not consent to their deployment in the first place. Be that as it may, the moral cost of any gamble would fall squarely on the gambler, who would be readily identifiable for all to see. In this way, we could at all times ensure that a human is liable to receive blame for any self-initiated robot acts deemed immoral.

This already satisfies the aforementioned requirement that we justly hold *someone* responsible for the actions taken in war. Sparrow, by contrast, argues that it is unfair to "assign" responsibility to the commanding officer of autonomous robots, since "[t]he use of autonomous weapons therefore involves a risk that military personnel will be held responsible for the actions of machines whose decisions they did not control" [1]. On our account, Sparrow's worry can be accommodated such that the "risk" he highlights is eliminated. Our picture could be complicated further by the inclusion of administrative safeguards so that it is formally expressed by the commander (in addition to

her superiors) that she is (un)willingly following orders to deploy such machines, and accepts (no) responsibility for the outcome of doing so. In this way, even if the president does sign off on the use of such machines, the front line soldiers using them could formally express their (un)willingness to do so, and hence their (un)willingness to accept partial responsibility for its actions (after all, blank checks are indeterminate in the amount they allow, but they are crystal clear about who is doing the allowing).

Sparrow briefly considers (and dismisses) something along the lines of what we are proposing. He writes that "[i]n these cases we *simply insist* that those who use [the robotic weapons] should be held responsible for the deaths they cause, even where these were not intended" [1, emphasis added]. The main difference between our proposal and the one discussed by Sparrow is that we regard it as fair and reasonable for the commanding officer to willingly and freely *assign responsibility to herself*, ahead of time. As a result, it is important to underscore that our way out of Sparrow's dilemma does not entail that a community will arbitrarily select a prominent figure as a lightning rod for its disapproval. Rather, we think the incentives are so balanced that users of autonomous machines could willingly accept responsibility for their actions, ahead of time, and thus become a justified, reasonable, and fair locus of surrogate responsibility for those autonomous robot's actions. We still cannot hold a robot responsible for its actions, and merely assigning the commanding officer responsibility is still unfair. Yet, asking a suitably-ranked and sufficiently-informed person to decide whether or not she is willing to prospectively assume responsibility for the actions of autonomous robots *is* fair, and would satisfy the requirement that someone be held responsible. Perhaps there are other, unstated, desiderata that we might want to take into consideration, but as things currently stand, the proposal goes through.

It is therefore appropriate to understand our view as a sort of "vouching for" rather than a "pointing of fingers." Whether this practice will be well implemented or not is an empirical issue that cannot be determined in advance of the facts. But, we can nevertheless infer that, if an informed surrogate willingly consents, she can take the risk identified by Sparrow (there may of course be independent reasons for commanding officers or presidents to *not* accept surrogate moral responsibility).

Although Sparrow does not seriously consider the possibility of establishing a viable social contract of the sort we gesture at, the basic idea has proven its considerable merit in the customary chain of command of the military. For instance, when the Captain of a ship accepts blame for the actions of those officers under him, the terms of that office permit us to blame the Captain, even if no clear causal chain can be established that would link her to the reprehensible action(s) in question. Accepting responsibility for such behaviour is simply part of the Captain's job description, a "role responsibility" to which the Captain has committed through her explicit acceptance of her post. For example, the Canadian Armed Forces (via the 1985 *National Defence Act* and the Department of National Defence's *Army Ethics Programme*) subscribes to this way of assigning responsibility for ethical and legal misconduct. A similar rationale for the assignment of responsibility was at work in the Nuremberg trials, where commanding officers were deemed (partially) responsible for the behaviour of their troops (although, in some cases, the troops were also held responsible). The same sort of contractual acceptance of responsibility could be fruitfully employed, we argue, in the case of autonomous robots.

Sparrow's analogy between autonomous killer robots and child soldiers [1] can thus be granted without compromising the force of our critique. Child soldiers are not such that they lack autonomy (in its entirety), and yet they do not seem responsible for their actions—blaming them for their actions seems unjustified, even if not entirely unwarranted. Parents are not typically responsible for the actions of their grown offspring. Still, given that parents *are* responsible for the actions of their children (who are arguably autonomous in the required sense), it seems fair to say that part of the role of parent is to own up to this responsibility, rearing the child so that they learn to behave in (morally) acceptable ways, and accepting (partial) responsibility in cases where she intentionally fails to do so. Needless to say, the addition of an explicit social contract would make that link even tighter.

It is worth bearing in mind that our proposal charitably grants Sparrow's twin contentions that 1) a programmer cannot be justifiably held responsible for the actions of a truly autonomous robot and that 2) holding a non-sentient robot responsible is meaningless and unsatisfactory. What we are challenging is the assumption that this exhausts the moral avenues. In an effort to introduce a *tertium quid*, we argue that if a commanding officer willingly deploys autonomous robots that can act immorally, and moreover publicly recognizes that those robots cannot be held responsible, then she has thereby accepted responsibility for their actions. A person in this position is exercising her privilege of un-coerced rational consent. This in turn yields a relatively clear (but non-causal) connection between the officer and the actions of the autonomous robot, insofar as the person who directly sanctions the use and deployment of such robots assumes responsibility for their actions come what may, and is justifiably subject to punishment for doing so.

In a way, our argument supplies a much-needed philosophical resource, since in the end we need to ensure that autonomous robots meet ethical and safety standards prior to their deployment. It seems reasonable to assume that a commanding officer faced with the decision to deploy autonomous robots would work hard to ensure that those robots behave properly, since it is *she* who would be held responsible and punished in the event of their misbehaviour (however one wants to construe this). Indeed, it would make for an informative empirical study to ask current military officers and commanders ahead of time whether they would be willing to accept such surrogate responsibility. Coupled with our proposal, maybe the results of such an inquiry could suffice to halt the development of autonomous war machines. After all, why build autonomous yet un-punishable lethal machines if no one is willing to accept responsibility for their actions? Although long-range missiles sometimes cause collateral damage [10], we do not blame the missile itself; instead, we blame those deploying the missile [4]. Autonomy introduces a considerable element of personal "risk" in deploying such high-tech weaponry, since one can never be fully certain of the decisions a given robot may take. This risk is a trade-off for the privileges one enjoys from being a commander (if, as a contingent matter, no one happens to submit to this demand, then so much the worse for those who want to create and deploy autonomous robots).

A cynic could argue that, once proper punishment has been carried out, a fresh candidate could come into the picture and allow the carnage to start all over again. Of course, so long as forecasts are being had on the cheap (being answerable to nothing more tangible than one's imagination), nothing bars these kinds of criticisms. Realistically though, humans are responsive to past events, so (all other things being equal) the inductive likelihood that a person in a visible position of authority would sign off after such repeated debacles would shrink. Again, for someone who inclines to worst case scenarios, that may bring little comfort. Yet, what is the alternative: Rogue use of killer robots without any attempt at securing a morally responsible agent? Taking the cynic at her word, she should recognize that, either way, technologies that can be used will be used. Once that maxim has been granted, the task becomes to minimize undesired outcomes. At any rate, it is inconsistent for one to claim that 1) repeated episodes of carnage will naturally ensue and 2) if we just say no, everyone will abide by that. (1) betokens unwarranted pessimism, (2) is naively optimistic, and a switch between the two would be *ad hoc*. Nothing is gained by predicating policies on mistaken views of human nature, and debates cannot advance if everyone is permitted to doctor future events to suit their personal intuitions.

Our suggestion that the likelihood of deployment would shrink in light of past fiascos of course assumes that the relevant practical decisions would be motivated by a certain measure of informed rationality. This assumption, like all assumptions, can certainly be called into question. However, the retort that humans have the power to act contrary to the dictates of reason is true but trivial, since the observation applies to the full spectrum of philosophical branches concerned with human activity. The burden is thus on whoever wants to exploit the element of voluntarism inherent in decision-making [11] to show why glass-half-empty worries about human wickedness should be deemed more likely than glass-half-full confidence in human reasonableness.

In the same vein, we might note that, despite the focus of Sparrow's discussion, full autonomy does not solely imply a willingness to kill. Presumably, such an exotic technological endowment, if it is indeed possible, might also lead an autonomous robot to mimic laudable human acts. As such, the "blank check" we have introduced need not be viewed solely in a negative light. For the exchange to succeed fully, it is not enough to just limit the ethical responsibility to a consenting human commander when things go wrong; we could also transfer any praise a robot might garner to the same person. While it is less intuitive to connect the successes of an autonomous robot with a human who had very little or nothing to do with the mission in question, allowing for a non-causal transfer of both prestige and ethical responsibility to the commander could be a vital component in the social contract we envision.

4 CONCLUSION

The hypothetical situation presented by Sparrow points to an interesting set of circumstances which force us to question our notion of ethical responsibility in increasingly unmanned battlefields. In that sense, it betokens philosophical thinking at its best. The twofold recognition that neither hapless programmers nor unfeeling machines deserve punishment acts like a dialectic vise, thereby compelling Sparrow to conclude that the use of autonomous robots would be a moral aberration. There is a sense in which he is clearly right. We have argued, however, that Sparrow's analysis, while fruitful and in many ways correct, nevertheless overlooks the possibility of a sufficiently high-ranking commanding officer accepting responsibility for the robot's actions, and thus being accountable for any violation of the rules for the ethical conduct of warfare.

In essence, our proposal retains the non-causal imputation involved in scapegoating while dropping its arbitrariness: Since humans are capable of informed consent and pleasure/pain, a suitable and ascertainable target for punishment can be established, thereby ensuring visible conformity with the tenets of just war theory. Although it may not be possible to trace a tenable physical link between the negative (or positive) actions of a fully autonomous robot and its human commander, our suggestion has been that this transitive chain is inessential to the ethical question, and can be bypassed by using a more explicit social contract or "blank check."

A robot may perform self-initiated actions, but it does not have to suffer what we would consider a just punishment for a violation of our ethical rules. Instead, the moral blame and accompanying punishment should be placed squarely on a human agent who, through her own volition, has traded a part of her freedoms for the prestige of occupying a high-ranking position in a given social hierarchy (say, a governmental or military chain of command). If no one is willing to accept this responsibility, then they should not deploy autonomous killer robots in the first place. In either case, this way of framing the issue keeps a human in the loop in a morally defensible manner, rendering the force of Sparrow's argument far weaker than it first seemed.

In light of the foregoing criticism, a superficial gloss might position us as somehow "against" Sparrow. That would be superficial indeed. To be clear: We have not contradicted Sparrow's case for peace, we have instead tried to show that it is incomplete, and have addressed the overlooked option right away. Speculation must eventually pass through the bottleneck of action. When all is said and done, Sparrow's prescriptive yield is: "Wait, don't push that button, it might lead to senseless violence." Our yield is: "Wait, don't push that button, it might lead to senseless violence, and if it does, you will spend the rest of your life in jail." Now, policy makers rarely heed the advice of philosophers without some practical incentive, so we leave it to the reader to judge which proposal is most likely to further the cause of peace.

REFERENCES

[1] R. Sparrow. Killer Robots. *Journal of Applied Philosophy*, 24: 62-77, (2007).

[2] M. Walzer. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, Basic, New York (1977).

[3] A. Matthias. The Responsibility Gap. *Ethics and Information Technology*, 6: 175-183, (2004).

[4] P.M. Asaro. How Just Could a Robot War Be? In: *Current Issues in Computing and Philosophy*, A. Briggle, K. Waelbers, and P. Brey (Eds.), IOS Press, Amsterdam (2008).

[5] A. Krishnan. *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Ashgate, Farnham (2009).

[6] P. W. Singer. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, Penguin, New York (2010).

[7] R.C. Arkin. *Governing Lethal Behavior in Autonomous Robots*, Chapman and Hall, Boca Raton (2009).

[8] R.C. Arkin and P. Ulam. *An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions*, Technical Report GIT-GVU-09-04, Georgia Institute of Technology, Atlanta (2009).
[9] G. Lokhosrt, G. and J. van den Hoven. Responsibility for Military Robots. In: *Robot Ethics: The Ethical and Social Implications of Robotics*, P. Lin, G. Bekey, and K. Abney (Eds.), MIT Press, Cambridge (2011).
[10] J.P. Sullins. RoboWarfare: Can Robots be More Ethical than Humans on the Battlefield? *Ethics and Information Technology*, 12: 263–275, (2010).
[11] E. Ullmann-Margalit and S. Morgenbesser, S. Picking and Choosing. *Social Research*, 44: 757-785, (1977).

Patiency Is Not a Virtue: Suggestions for Co-Constructing an Ethical Framework Including Intelligent Artefacts

Joanna J. Bryson[1]

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

1 INTRODUCTION

The question of Robot Ethics is difficult to resolve not because of the nature of Robots but because of the nature of Ethics. In particular, we must decide what "really" matters—what are our ethical priorities? Are we more obliged to our biological kin or to those who share our ideas? Do we value the preservation of culture more or the generation of new ideas?

The primary argument of this paper is that integrating a new problem like artificial intelligence (AI) into our moral systems is an act normative, not descriptive, ethics. Descriptive ethics may take us some way in establishing precedent, though few consider precedent sufficient or even necessary for establishing what is right. But the advent of potentially-autonomous decision-making human artefacts is novel. Deciding ethical priorities requires establishing the basis of our systems of values. But asking "what really matters" is like asking "what happened before time": it sounds at first pass like a good question, but in fact makes a logical error. *Before* is not defined outside of the context of time, and similarly, we cannot circuitously assume that a system of values underlies our system of values. Consequently, the "correct" place for robots in human society cannot be resolved from first principles or purely by reason.

In this paper, I therefore begin my argument not from what should matter to us but rather from why things do. I start by considering ethics and moral patiency from an evolutionary perspective, before turning to consider where we might want to slot robots into our contemporary ethical frameworks and society. The nature of machines as artefacts means that the question of their morality is not what moral status they deserve, but what moral status we are obliged to assign them, and to build them to be competent to meet. What makes this different from reasoning about natural entities is that our obligations can be met not only through constructing the socio-ethical system but also through specifying characteristics of the artefacts themselves. This is the definition of an artefact, it is something we create, and it applies to both ethical and artificially intelligent systems.

[1] University of Bath, England, United Kingdom email: j.j.bryson@bath.ac.uk

To be very clear from the outset, the moral question I address here is not whether it is possible for robots or other AI artefacts to be moral patients. Human culture can and does support a wide variety of moral systems. The question is whether we as academics should recommend putting artefacts in that position. I will argue that making robots such that they deserve to be moral patients could in itself be construed as an immoral action, particularly given that it is avoidable. In doing so I consider not only human society, but also incidentally make robots into second-order moral patients. I claim that it would be unethical to put them in a situation of competition with us, to make them suffer, or to make them unnecessarily mortal. I never claim that it is wrong to use machine intelligence to create —- that is, to generate human culture. But I do claim that is incoherent to think that human pleasure can be extended by proxy.

2 THE NATURE OF LIFE AND INTELLIGENCE

I start from the entirely functionalist perspective that our system of ethics has coevolved with our species and our societies. As with all human (and other ape) behaviour, it is rooted both in our biology and our culture. Nature is a scruffy designer with no motivation or capacity to cleanly discriminate between these two strategies, except that that which must change more quickly should be represented more plasticly. As human cultural evolution has accelerated, increasingly our ethical norms are represented in highly plastic forms such as legislation and policy.

The problem with a system of representing behaviour so plastic as explicit decisions is that it can lead to dithering. *Dithering* is switching from one goal to the other so quickly that little or no progress is made in either (18, 23). Dithering is a problem potentially faced by any autonomous actor with multiple at least partially conflicting goals which must be maintained concurrently. Here 'partial conflict' can be resource-based, for example needing to visually attend to the actions of two children at one time. An example of dithering in early computers was called *thrashing*—a result of running two interacting programs which were each nearly as large as primary memory. The operating system would allocate each program a period of processing time, which would necessarily start with an attempt to read that program in to memory from disk, a process called *paging* or *swapping*. If swapping each program in took longer than the time slice allocated to that program, the computer would appear to 'freeze' since there would be no actual progress made by either program.

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

law is often debated in the context of the US constitution, despite being rooted in British Common Law and therefore a constantly changing set of precedents. Ethics is often debated in the context of ancient holy texts, even when they concern contemporary questions such as abortion or robots about which there is no reference or consideration in the original documents.

Perhaps it is to avoid dithering that our society believes that basic principles of our ethics are rational and fixed, and that the apparent changes such as universal suffrage or the end of legalised human slavery are simply "corrections" to make the system more rational. But a better model is to consider our ethical structures and morality to co-evolve with our society. When the value of human life relative to other resources was lower, murder was more frequent and political empowerment less widely distributed (20). What it means to be human has changed, and our ethical system has changed along with this.

3 The Origins of Social and Ethical Behaviour

Assessing morality is not trivial, even for apparently trivial, 'robotic' behaviour. MacLean et al. (21) demonstrate the overall social utility of organisms behaving in what at first assessment seems to be an anti-social free riding off of pro-socially manufactured costly public goods. Single-cell organisms produce a range of public goods including shelter and instructions for combatting antibiotics (22). In this particular case we are discussing the production of digestive enzymes by the more 'altruistic' of two isogenic yeast strains. The yeast must excrete these enzymes outside of their bodies as they can only directly absorb pre-digested food. The production of these enzymes is costly, requiring difficult-to-construct proteins, and the production of pre-digested food is beneficial not only to the excreting yeast but also to any other yeast in its vicinity. The production of these enzymes thus meets a common anthropological and economic definition of *altruism*, paying a cost to express behaviour that benefits others (13, 24).

In the case of single-cell organisms there is no 'choice' as to whether to be free-riding or pro-social —this is determined by their strain, but the two sorts of behaviour are accessible from each other via mutation. Natural selection then performs the 'action selection' by determining what proportion of which strategy lives and dies. What MacLean et al. (21) have shown is that selection operates such that the species as a whole benefits optimally. The 'altruistic' strain in fact *overproduces* the public good (the digestive enzymes) at a level that would be wasteful, while the 'free-riding' strain of course underproduces. Where there are insufficient altruists free-riders starve, allowing altruists to invade. Where there are too few free-riders excess food aggregates, allowing free-riders to invade. Thus the greatest good—the most efficient exploitation of the available resources—is achieved by the species through a mixture of over-enthusiastic altruism and free riding. Why doesn't evolution just optimise the species as a whole to produce the optimal level of public goods? This is again due to plasticity. The optimal amount of enzyme production is determined by the ecological substrate the yeast inhabits, and this can change more quickly than the physical mechanism for enzyme production in one strain can evolve. However death and birth can be exceedingly rapid in bacteria. A mixed population composed of multiple strategies, where the high and low producers will always over and under produce respectively and their proportions can be changed very rapidly is thus the best strategy for tracking the rate of environmental change — for rapidly responding to variation in opportunity.

What does these results mean for humans? Are we able to add benefit to over-production of public goods by calling the action of creating them 'good' and associating it with social status, while our culture has evolved to trust self interest to motivate the maintenance the countervailing population of defectors? Does our assessment of the 'correct' amount of investment vary by socio-political context, for example increasing massively in times of warfare but returning to more individually-productive levels at times of peace? Could the reduction of 'good' behaviour itself be an act of public good in times when society requires more individual productivity or self-sufficiency? This is a thread of a current research programme in my group, looking to explain variations by global region in public-goods regulation as demonstrated by Herrmann et al. (17). We have preliminary evidence that human societies can also be described via varying proportions of individuals applying particular social strategies, and that these proportions vary with the underlying socio-economic substrate.

4 Freedom and Morality

Even if some combination of biology and the social sciences can explain the origins of our current social norms and how these change, this by no means tells us what they *ought* to be, nor how they should be taken forward as the world changes. I will now turn to philosophy to look at how we commonly define moral agency and patiency, with the goal of exploiting these definitions in the next section for communicating the roles we should expect AI to play in our society.

To quote Johnson (19), "[Moral] action is an exercise of freedom and freedom is what makes morality possible." For millennia morality has been recognised as something uniquely human, and therefore taken as an indication of human uniqueness and even divinity (14). But if we throw away a supernaturalist and dualistic understanding of human mind and origins, we can still maintain that human morality at least *is* rooted in the one incontrovertible aspect of human uniqueness — language — and our unsurpassed competence for cultural accumulation that language both exemplifies and enables[2]. The cultural accumulation of new concepts givs us more idas and choices to reason over, and our accumulation of tools gives us more power to derive substantial changes to our environment from our intentions.

If human morality depended simply on human language then our increasingly language-capable machines would certainly be excellent candidates for agency and patiency. But I believe that freedom — which I take here to mean *the socially-recognised capacity to exercise choice* is the essential property of a moral actor. Dennett (12) argues that human freedom is a consequence of evolving complexity beyond our own capacity to provide a better account for our behaviour than to attribute it to our own individual responsibility. This argument entails a wide variety of interesting consequences. For example, as our science of psychology develops and more behaviour becomes explicable via other means (e.g. insanity) fewer actions become moral.

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

[2] Bryson (3, 5) argues that language while unique is not inexplicably improbable but rather a result of the conjunction of two less-unusual adaptive traits: the strongly cultural strategy that many large, social species of animals, particularly apes, exploit, and the capacity for vocal imitation which has emerged independently many phyla, but nowhere else among the simians.

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

Note that this captures society-specific morals as well as the individual's role as the actor. With this definition I deliberately extend morality to include actions by other species which may be sanctioned by *their* society, or by ours. For example, non-human primates will sanction those that violate their social norms by being excessively brutal in punishing a subordinate (11), for failing to 'report' vocally available food (16) or for sneaking copulation (10). While reports of social sanctions of such behaviour are often referred to as 'anecdotal' because they are not yet well documented in primate literature, they are common knowledge for anyone who has been lucky enough to work with socially housed primates. I personally have violated a Capuchin norm: *possession is ownership*. I was sanctioned (barked at) by an entire colony — not only those who observed the affront[3], but also those housed separately who had no visual knowledge of the event but joined the chorus of reproach.

Similarly, this definition allows us to say dogs and even cats can be good or bad when they obey or disobey norms they have been trained to recognise, when they have demonstrated capacity to select between these alternative behaviours, and when they behave as if they expect social sanction when they select the bad option.

To return to the main point of this essay, there is then I think no question that we can already train or simply program machines to recognise more or less socially acceptable actions, and to use that information to inform action selection. The question is whether it is moral for us to construct machines that of their own volition choose the less-moral action. The trick here returns to the definition of freedom I took from Dennett. For it to be rational for us to describe an action by a machine to be "of its own volition", we must sufficiently obfuscate its decision making process that we cannot otherwise predict its behaviour, and thus are reduced to applying sanctions to it in order for it to learn to behave in a way that our society prefers.

What is fundamentally different from nature here is that since we have perfect control over when and how a robot is created, responsibility is traded off. Consider responsibility for actions executed by the artefact that lie within our own understanding, and thus for which we would ordinarily be responsible. If we choose to assign this responsibility to the artefact we are deliberately disavowing the responsibility ourselves. Currently, even where we have imperfect control as in the case of young children, owned animals and operated machines, if we lose control over entities we have responsibility for and cannot themselves be held accountable, then we hold the responsibility for that loss of control and whatever actions by these other entities comes as a consequence. If our dog or our car kills a child, we are not held accountable for murder, but we may be held accountable for manslaughter. Why — or in what circumstances — should this be different for a robot?

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

5 Principles of Robotics

When considering then how we should adjust our ethical systems to encapsulate the AI we create, there are multi-level questions and ethical strategies, all of which need to be considered. In the yeast example I gave earlier, 'anti-social' behaviour actually regulated the overall investment of a society (a spatially-local subset of a species inhabiting a particular ecological substrate) in a beneficial way. Behaviour that was disadvantageous very local to the free-riders was less locally advantageous to the species. The definition of morality introduced in Section 4 depends on social benefit. In the Machine Question here there are at least two potential societies whose benefits we need to consider. For each of these, I will consider who benefits and who does not from the designation of moral agency and patiency on AI.

- *The perspective of human well being*. The advantages to humans seem to be primarily that it feeds our ego to construct objects that we owe moral status. It is possible that in the long term it would also be a simpler way to control truly complex intelligence, and that the benefits of that complex intelligence *might* outweigh the costs of losing our own moral responsibility and therefore status. The principle cost I see is the facilitation of the unnecessary abrogation of responsibility of marketers or operators of AI. For example, customers could be fooled into wasting resources needed by their children or parents on a robot, or citizens could be fooled into blaming a robot rather than a politician for unnecessary fatalities in warfare (2, 8).
- *The perspective of AI well being*. Although this argument has been overlooked by some of my critics (notably 15), Bryson (4, 6) makes AI into second-order moral patients by arguing that we should not put AI in the position of competing with us for resources, of longing for higher social status (as all evolved social vertebrates do), of fearing injury, extinction or humiliation. In short, we can afford to stay agnostic about whether AI have qualia, because we can simply avoid constructing motivation systems encompassing suffering. We know we can do this because we already have. There are many proactive AI systems now, and none of them suffers. Just as there are already machines that play chess or do arithmetic better than we do, but none of them aspires to world domination. There can be no costs to the AI in the system I describe, unless we postulate rights of the 'unborn', or in this case the never-designed.

Bryson et al. (9) argue that the right way to think about intelligent services (there in the context of the Internet, but here I will generalise this) is as extensions of our own motivational systems. We are currently the principle agents when it comes to our own technology, and I believe it is our ethical obligation to design both our AI and our legal and moral systems to maintain that situation. Legally and ethically, AI works best as a sort of mental prosthetic to our own needs and desires.

The best argument I have heard against my human ethics perspective is that maltreating something that reminds us of a human might lead us to treat other humans or animals worse as well. The UK's official *Principles of Robotics* specifically address this problem in its fourth principle (1, 7, c.f. Appendix A). This principle does so in two ways. First, robots should not have deceptive appearance—they should not fool people into thinking they are similar to empathy-deserving moral patients. Second, their AI workings should be 'transparent'. That is, clear, generally-comprehensible descriptions of their goals and intelligence should be available to any owner, operator or

concerned user, presumably over the Internet[4]

The best argument I have heard made against my AI ethics perspective is that it may be impossible to create the sort of intelligence we want or need unless we follow the existing biologically-inspired templates which include social striving, pain, etc. So far I have seen no proof of this position. But if it is ever demonstrated, even then we would not be in the position where our hand was forced — that we must permit patiency and agency. Rather, we will *then* have enough information to stop, take council, and produce a literature and eventually legislation and social norms on what is the appropriate amount of agency to permit given the benefits it provides.

6 Conclusion

As Johnson (19, p. 201) puts it "Computer systems and other artefacts have intentionality—the intentionality put into them by the intentional acts of their designers." It is unquestionably within our society's capacity to define AIs as moral agents and patients, in fact many articles in this volume are working on this project. It may be technically possible to create AI that would meet contemporary requirements for agency or patiency. But even if it is possible, neither of these two statements makes it either necessary or desirable that we should do so. Both our ethical system and our artefacts are amenable to human design. The primary argument of this article is that making AI moral agents or patients is an intentional and avoidable action. The secondary argument which is admittedly open to debate, is that avoidance would be the most ethical choice.

ACKNOWLEDGEMENTS

I would like to thank more people than I can remember for some of the insights and counterarguments mentioned above, but particularly David Gunkel and Will Lowe.

REFERENCES

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

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

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

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

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

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

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

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

[8] Bryson, J. J. and Kime, P. (1998). Just another artifact: Ethics and the empirical experience of AI. In *Fifteenth International Congress on Cybernetics*, pages 385–390.

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

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

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

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

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

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

[15] Gunkel, D. (2012). *The Machine Question*. MIT Press, Cambridge, MA.

[16] Hauser, M. D. (1992). Costs of deception: Cheaters are punished in rhesus monkeys (*macaca mulatta*). *Proceedings of the National Academy of Sciences of the United States of America*, 89(24):12137–12139.

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

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

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

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

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

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

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

[24] Sylwester, K., Mitchell, J., and Bryson, J. J. (2012). Punishment as aggression: Uses and consequences of costly punishment across populations. in prep.

Appendix A: The EPSRC Principles of Robotics

The full version of the below lists can be found by a Web search for *EPSRC Principles of Robotics*, and they have been EPSRC policy since April of 2011 (1). The first list is the principles themselves, in italics, with annotations taken from Bryson (7).

1. *Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.* While acknowledging that anything can be used as a weapon by a sufficiently creative individual, the authors were concerned to ban the creation and use of autonomous robots as weapons. Although we pragmatically acknowledged this is already happening in the context of the military, we do not want to see robotics so used in other contexts.
2. *Humans, not robots, are responsible agents. Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.* We were very concerned that any discussion of "robot ethics" could lead individuals, companies or governments to abrogate their own responsibility as the builders, purchasers and deployers of robots. We felt the consequences of this concern vastly outweigh any "advantage" to the pleasure of creating something society deigns sentient and responsible.
3. *Robots are products. They should be designed using processes which assure their safety and security.* This principle again reminds us that the onus is on us, as robot creators, not on the robots themselves, to ensure that robots do no damage.
4. *Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.* This was the most difficult principle to agree on the phrasing of. The intent is that everyone who owns a robot should know that it is not "alive" or "suffering", yet the deception of life and emotional engagement is precisely the goal of many therapy or toy robots. We decided that so long as the responsible individual making the purchase of a robot has even indirect (e.g. Internet documentation) access to information about how its "mind" works, that would provide enough of an informed population to keep people from being exploited.
5. *The person with legal responsibility for a robot should be attributed.* It should always be possible to find out who owns a robot, just like it is always possible to find out who owns a car. This again reminds us that whatever a robot does, some human or human institution (e.g. a company) is liable for its actions.

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

1. *We believe robots have the potential to provide immense positive impact to society. We want to encourage responsible robot research.*
2. *Bad practice hurts us all.*
3. *Addressing obvious public concerns will help us all make progress.*
4. *It is important to demonstrate that we, as roboticists, are committed to the best possible standards of practice.*
5. *To understand the context and consequences of our research we should work with experts from other disciplines including: social sciences, law, philosophy and the arts.*
6. *We should consider the ethics of transparency: are there limits to what should be openly available?*
7. *When we see erroneous accounts in the press, we commit to take the time to contact the reporting journalists.*

Is there a continuity between man and machine?

Johnny Hartz Søraker[1]

Abstract. The principle of formal equality, one of the most fundamental and undisputed principles in ethics, states that a difference in treatment or value between two kinds of entities can only be justified on the basis of a relevant and significant difference between the two. Accordingly, when it comes to the question of what kind of moral claim an intelligent or autonomous machine might have, one way to answer this is by way of comparison with humans: Is there a fundamental difference between humans and machines that justifies unequal treatment, or will the two become increasingly continuous, thus making it increasingly dubious whether unequal treatment is justified? This question is inherently imprecise, however, because it presupposes a stance on what it means for two types of entities to be *sufficiently* similar, as well as which types of properties that are *relevant* to compare. In this paper, I will sketch a formal characterization of what it means for two types of entities to be continuous, discuss what it implies for two different types of entities to be (dis-)continuous with regard to both ethics and science, and discuss a dramatic difference in *how* two previously discontinuous entities might *become* continuous.

1 INTRODUCTION[2]

The concept of 'continuity' has been championed by MIT historian Bruce Mazlish, who claims that progress in science and technology will inevitably result in a fourth *continuity* between man and machine [1]. According to Mazlish, there have been three dramatic scientific revolutions in the history of mankind, and these revolutions are best described as the establishment of *continuities;* mankind has generally come to acknowledge that there is no sharp discontinuity between our planet and the rest of the universe (Copernican revolution), between humans and animals (Darwinian revolution), nor between rational and irrational humans (Freudian revolution). Mazlish argues that we should also overcome what he terms the *fourth* discontinuity; that there is no sharp discontinuity between humans and machines.

There are a number of problems with Mazlish's account, however. First, it is difficult to discern precise criteria for what it means for something to become continuous, which means that we seemingly operate with inconsistent conceptions of continuity. A clear example of this can be seen with regard to animal experiments. On the one hand, animal researchers presuppose a continuity between humans and animals – if not, the results would not be relevant to humans. On the other hand, they also presuppose a *dis*continuity – if not, the experiments would be unethical. It is clear from this example that we often regard two types of entities as continuous in one respect and discontinuous in another. Thus, we need to clarify what these different 'respects' are, and how they relate to each other.

Mazlish seems to claim that a continuity is determined by whether or not the inner workings of two entities can be explained within the same scientific framework, such as computationalism being able to explain both computers and the human brain. Although there are numerous problems with Mazlish's approach, which I will return to below, one if its advantages is that it takes an epistemological rather than ontological approach to the moral status debate – an approach that in many ways mirrors Alan Turing's well-known approach to the question of whether machines can be intelligent [2].

2 FOLLOWING TURING'S LEAD

For the present audience, I presuppose that it is not necessary to explain the Turing test as a means of judging whether a computer is intelligent enough, but in short Turing argues that a computer is to be regarded as intelligent if a human judge cannot reliably distinguish the computer from the human in an imitation game. What is important here is that Turing turns the question of intelligence from an ontological to an epistemological one. That is, Turing does not ask which properties a computer must possess in order to be deemed intelligent (which is an ontological question), but rather how an intelligent observer judges its behaviour. The latter is a type of epistemological question, where we are really asking what kind of explanatory framework we need to presuppose in order to understand a particular type of behaviour. If a computer were to pass the Turing test, this means that the judge had to *explain* its behaviour as coming from an intelligent being, which says nothing about which properties that being must have (other than being able to display the behaviour in question). Notice that this approach is radically different from the typical approach to questions of moral status and the like, where we typically discuss which properties an entity must possess in order to be regarded as a moral person (e.g. sentience [3], conception of one's own life [4], or having a will to live [5]).[3]

In a similar manner, Mazlish argues (indirectly) that two types of entities should be regarded as continuous if they do not require different scientific frameworks; if the same framework of scientific concepts and models can adequately explain the phenomenon under study. On this background, the Copernican revolution was really a realization that we do not need different scientific frameworks for the earth and the heavens (as was the case with the Aristotelian framework), the Darwinian revolution was a realization that we do not need different scientific frameworks for humans and other animals, and the Freudian revolution was a realization that we do not need different scientific frameworks for the mentally ill and the mentally healthy. Mazlish's prophesized fourth continuity, then, is the realization that we do not need different scientific frameworks for computers and humans either. Thus, all of these continuities amount to radical changes in how to explain different types of

[1] Dept. of Philosophy, University of Twente, Postbox 217, 7500 AE Enschede, Netherlands. Email: j.h.soraker@utwente.nl

[2] Allow me to emphasize that this paper, in line with the IACAP tradition, is work-in-progress, and is presented for the purpose of receiving peer feedback before being finalized. As such, I admit that this paper still lacks the precision and clarity that is to be expected from a finished paper. I hope the reader will apply the principle of charity, and regard this as a starting point for further discussion rather than a fully worked out standpoint.

[3] Cf. Søraker [6] for an overview of the discussion of moral status from an ontological, property-based perspective.

entities (epistemological), rather than saying that all the entities have the same properties and/or mode of existence (ontological). Just like Turing thought an epistemological approach to the question of machine intelligence was more fruitful than an ontological one, I will take a similar approach to the question of continuity in the remainder of this paper. In doing so, I first need to make some important changes to Mazlish's approach, which despite its advantages gives rise to some fundamental problems.

3 PROBLEMS WITH A SINGLE-LEVEL APPROACH TO 'CONTINUITY'

As mentioned, Mazlish seems to claim that a continuity is determined by whether or not the inner workings of two entities can be explained within the same scientific framework, for instance the same physics being able to explain both the earth and the heavens, behaviourism being able to explain both humans and other animals, psychoanalysis being able to explain both mental health and illness, and computationalism being able to explain both computers and the human brain. This is what I refer to as a single-level approach, for reasons I will explain in more detail below.

This approach is problematic for two related reasons. First, anything *can* be explained within the same scientific framework. Disregarding supernatural and/or substance dualist accounts, it is probably possible in principle to explain the workings of the human brain and a computer by physics alone – and if we believe in scientific progress, our ability to do so will increase in time with progress in physics (I will return to this below).

Second, anything can be explained *as if* it is an intentional agent, as also argued by Daniel Dennett, who refers to this as taking an *intentional stance* [7]. Since Mazlish does not specify how strict we need to be when claiming that the same scientific framework *can* explain two types of entities, his approach becomes inherently imprecise. If it is sufficient that it is *in principle* possible to explain something within the same framework, then every existing entity is continuous as long as there is no phenomena that cannot in principle be explained by some kind of physics. This would entail that humans are continuous with light bulbs, supernovae and clouds, which leaves the concept of little use. It seems more reasonable, then, to refer to some kind of pragmatism where it must not only be in-principle possible but also pragmatically feasible to explain two entities within the same framework. But, this would require some kind of measure for what it means to be pragmatically feasible. Is it, for instance, pragmatically feasible to explain the brain fully in terms of physical processes, or do we (also) need to invoke chemistry?

Some of these problems are difficult to escape, since it is hard to provide objective criteria for when a particular scientific framework ceases to be feasible. However, we can try to remedy the problem of explanations at different levels by explicitly invoking this into the conception of continuity. In the following section, I will sketch such a multi-level account of continuity.

Before outlining this multi-level account allow me to emphasize that my main concern in this paper is to discuss the formal nature of these continuities, so it is important to emphasize that the levels of explanation that I will use as examples below are to be seen as mere placeholders and the reader will inevitably find some of them problematic and/or imprecise. My goal is to first work out the formal schematics, and then the more substantial content should be worked out in more detail. This will, among other things, require a defence of a particular type of realism and view on scientific progress, both of which fall well beyond the scope of this paper.

4 A MULTI-LEVEL APPROACH TO 'CONTINUITY'

Rather than asking whether two types of entities can be explained within the same scientific framework, I believe it is better to approach this in terms of *sets* of scientific frameworks – or what I will refer to as sets of scientific levels of explanation. That is, rather than asking whether two entities can be explained within *the same* scientific framework, we should ask whether two entities require the same *set* of scientific levels of explanation. As Nagel [8] rightly argues, there are (at least) four fundamentally different types of explanation – deductive, probabilistic, teleological, and genetic – which makes it difficult to precisely define what a scientific level of explanation is. For present purposes, I will simply use the term in the more generic sense of a more or less coherent and mutually supportive set of principles, concepts and models that attempt to provide an account of the relationships between cause and effect.[4]

As mentioned, one of the problems with Mazlish's single-level approach, where continuity is established on the basis of a notion of sharing *one* scientific framework, is that we often choose different levels of explanation (or, to use Luciano Floridi's term, levels of *abstraction* [9]) depending on what it is that we seek to explain. Even if it is in-principle possible to explain human behaviour by physics alone, we typically employ higher-level explanations instead. For instance, at a behaviourist level of explanation, we employ concepts like stimulus and response to explain behaviour, without involving physics or chemistry. Even for entirely physicalist phenomena, such as an object moving through space, we often employ heuristics instead of explaining what is "really" going on. As such, the science of ballistics can be seen as a form of higher-level heuristics for explaining how an object moves through space without talking about the complex interplay between electrons and force fields. This, along with the other problems with a single-level approach mentioned above, entails that we cannot define a continuity in terms of *a* shared scientific framework. A much more promising approach is to define continuity in terms of having a particular set of scientific levels of explanation in common.

To simplify things, if we take a single-celled organism, we may be able to explain its functioning entirely in terms of physics.[5] As we get to more complex forms of life, however, such explanations quickly become untenable. At some point, the chemistry involved becomes too complex to be described in physics terms alone. At even higher levels of complexity, chemistry also fails to provide a full explanation and we need to start talking about biological processes and leave the actual

[4] I am very aware that this is far from precise, but it should be sufficient for establishing the more formal nature of continuities, which is the limited purpose of this paper. As Mieke Boon has pointed out to me, it is probably better to speak of 'practices of explanation', but this paper was due before this could be properly incorporated.

[5] For such an attempt, see Princeton University's Laboratory for the Physics of Life (http://tglab.princeton.edu/).

physical and chemical processes out of our explanations. At even higher levels, we may need to involve the environment and cognitive processes to a much higher extent, and start using principles from, say, behaviourism and comparative psychology. With humans, as evidenced by the widespread criticism and dismissal of *radical* behaviourism in the mid-1900s, we could also make the case that we need some kind of mental, phenomenological or folk-psychological level of explanation that cannot be reduced to any of the other levels. At even higher supraindividual levels, we may also require social, cultural and other value-laden levels of explanation – and we find ourselves far away from the original physicalist level. Which levels we may need in order to adequately explain a given entity is clearly controversial, and not my concern in this paper, but only the most radical and optimistic scientists maintain that we will in the foreseeable future be able to explain everything by means of one unified theory. On the basis of all this, it seems evident that if we are to define continuity in terms of which type of explanation is required, we must talk about *sets of levels of explanation* (multi-level) instead of Mazlish's single-level scientific frameworks.

On this background, we can stipulate the preliminary hypothesis: two types of entities are continuous if and only if an adequate understanding of their nature and properties require the same set of scientific levels of explanation; two types of entities are discontinuous if and only if an adequate understanding of their nature and properties does *not* require the same set of scientific levels of explanation. To illustrate, humans and other animals are continuous if and only if a full understanding of their nature and properties require the same set of scientific levels of explanation (LoE).[6] These definitions still lack a lot of precision, however, and we need to first specify what is meant by 'required'.

5 EPISTEMOLOGICAL VS ONTOLOGICAL CONTINUITY

There are two radically different ways in which a LoE may be required for an adequate understanding. On the one hand, we could for instance argue that the human brain works in such a way that we cannot adequately understand its functioning without employing a chemical level of explanation. Perhaps the chemical properties of neurotransmitters and hormones function in a way that cannot possibly be accounted for by means of more mechanistic explanations – which would be an anti-reductionist view of chemistry. If such a chemical LoE is required because of the brain's unique mode of existence, then that LoE is required for *ontological* reasons.

On the other hand, we could argue that the human brain works in such a way that it is much more convenient or tractable to use a chemical LoE, even if such an explanation can *in principle* be reduced to a more basic LoE. If we, despite this in-principle possibility, do require a chemical LoE for pragmatic reasons, then that LoE is required for *epistemological* reasons.

In light of the above, we can already differentiate between an ontological and epistemological continuity:

Ontological continuity: two types of entities are ontologically continuous if and only if an adequate understanding of their nature and properties require the same set of scientific levels of explanation *in principle,* due to their mode of existence.

Epistemological continuity: two types of entities are epistemologically continuous if and only if an adequate understanding of their nature and properties require the same set of scientific levels of explanation *in practice*.

To illustrate, humans and other animals are ontologically continuous if and only if a full understanding of their nature and properties require the same set of scientific levels of explanation *in principle.* Humans and other animals are *epistemologically* continuous if and only if a full understanding of their nature and properties require the same set of scientific levels of explanation *in practice.* It is far from uncontroversial which LoEs are ontologically or epistemologically necessary for an adequate understanding (as well as what is to be meant by 'adequate'), and it is far beyond the scope of this paper to discuss this for different types of entities, but this question maps directly on to the reductionism debates present in the different disciplines. In philosophy of mind, a property dualist would hold that consciousness is somehow ontologically irreducible to neurobiology and physics – which means that a "higher" LoE is ontologically necessary for a full understanding of a conscious being. Eliminiative materialism, on the other hand, holds that consciousness can and should be explained at a neuroscientific LoE, thus claiming that higher LoEs (folk psychology, in particular) are *not* ontologically necessary for a full understanding of conscious beings. If we compare humans and other animals, the former would hold that conscious animals are ontologically discontinuous from non-conscious animals, whereas the latter would hold that conscious animals are ontologically continuous with non-conscious animals. Non-reductive physicalism, however, holds that conscious states really are the same as physical states and that the former can *in principle* be explained by the latter – but not *in practice.* According to such a view conscious beings would be *epistemologically* discontinuous from non-conscious beings.

6 THE SCHEMATICS OF CONTINUITIES

In light of the considerations above, we can now attempt to schematize what a continuity might look like, according to this multi-level approach. Consider the following schematic:

Type of entity / Required LoE	Humans	Other animals
Psychological	X	
Behaviorist	X	X
Physical	X	X

In this example, humans require a physical, behaviourist and a psychological LoE for a full understanding, whereas other animals can be fully understood by physical and behaviourist LoEs alone. If this were the case, then humans would be discontinuous with other animals. If the psychological LoE is required *in principle* this is an ontological discontinuity, if required only *in practice* this is an epistemological discontinuity. To more clearly show the difference with Mazlish's single-level approach, consider the following:

[6] For the remainder of this paper, I will use 'LoE' as shorthand for '(scientific) Levels of Explanation'.

Type of entity / Required LoE	Humans	Intelligent machines
Psychological	X	
Computational	X	X
Physical	X	X

In this example, humans are discontinuous with machines because they require a psychological LoE, whereas machines can be fully understood without. Again, this would be an ontological discontinuity if the computational and physical LoEs are not sufficient for a full understanding of humans (which entails some type of dualism). It would be an epistemological discontinuity if the psychological LoE is only required for pragmatic reasons (which entails some type of non-reductive materialism). Note that Mazlish's single-level approach is unable to account for this, and would treat humans and machines, in this example, as continuous as long as the computational LoE somehow explains both.

Putting this together, the multi-level account of continuity ultimately suggests some kind of hierarchy when it comes to discontinuities:

Type of entity / Required LoE	Humans	Other animals	Intelligent machines	Other inanimate objects
Psychological	X			
Behaviorist	X	X		
Computational	X	X	X	
Physical	X	X	X	X

Now we are able to describe the aforementioned problem of animal experimentation seemingly being an inconsistent practice since it presupposes both a radical similarity (scientific validity) and radical difference (ethical justifiability) between humans and other animals. The scientific validity of such experiments can be grounded in the fact that the LoEs that are relevant for the scientific validity are shared, whereas the LoEs that are relevant for the ethical justifiability are not shared.

This further illustrates how one purpose of establishing discontinuities in this manner is to map their required LoE onto a classification of moral status (or moral worth). That is, there are different ways to harm entities corresponding to their required LoE. In a manner of speaking, the more LoEs that are required for understanding an entity, the more ways there are to harm that entity. At a physical level, we may speak of a minimal harm in terms of entropy, at a computational level we may be able to speak of a minimal harm to self-sustainability, at a behaviourist level we are dealing with harms in terms of rewards and punishment, i.e. infliction of pain and pleasure, and at a psychological LoE it should be evident that the harms become much more complex, including things related to offense, liberty, dignity, privacy, self-actualization, and so forth.

Another important purpose of this schematization is to include Freud's notion of scientific progress changing our conception of ourselves in dramatic ways. In Freud's words: "the universal narcissism of men, their self-love, has up to the present suffered three severe blows from the researches of science" [10]. I will refer to such blows as *downgrading* as opposed to upgrading continuities, which also further illustrates what is meant by LoEs being required in principle and in practice.

7 DOWNGRADING VS UPGRADING CONTINUITIES

It follows from the notion of LoEs being only epistemologically necessary that scientific progress will bring about changes in which levels that are necessary to explain a given entity – which is reflected in the scientific ideals of parsimony, unification and reduction. This means that two types of entities previously seen as discontinuous may *become* continuous. That is, two types of entities that previously required different sets of LoE, at some point may end up requiring the same set of LoE. According to these schematics, this can come about in two different ways – which correspond to two radically different ways in which science may change our worldview, and where we can more precisely conceptualize Freud's notion of blows to the self-esteem of mankind.

First, we may come to realize that a type of entity no longer requires an LoE that we previously thought to be necessary – for instance when we are able to successfully reduce one scientific LoE to a more fundamental one. When two types of entities come to share the same set of LoE because one type *loses* a LoE, this amounts to a *downgrading* continuity. More schematically:

Type of entity / Required LoE	Humans	Other animals
Psychological	- X	
Behaviorist	X	X
Physical	X	X

This was precisely the concern when Skinner's radical behaviourism aspired to explain both humans and other animals entirely in terms of behaviourist principles. This would, according to this line of reasoning, entail a continuity between humans and other animals because humans would no longer require an additional LoE. In a manner of speaking, this would *downgrade* humans to the level of animals. We can see the same concern when it comes to intelligent machines:

Type of entity / Required LoE	Humans	Intelligent machines
Psychological	- X	
Computational	X	X
Physical	X	X

In this case, humans would become continuous with intelligent machines because they come to share the same set of LoEs due to the loss of one LoE. In a manner of speaking, this does not only amount to humans and computers being "the same", but that "humans are *nothing but* machines".

There is a converse way of becoming continuous, however. Consider the following:

Type of entity / Required LoE	Humans	Intelligent machines
Psychological	X	+ X
Behaviorist	X	+ X
Computational	X	X
Physical	X	X

In this case, humans and intelligent machines become continuous because the latter *attain* new LoEs. That is, intelligent machines

might become so complex that we can no longer explain their function by means of computational principles alone. At some point we may need to adopt psychological principles to explain intelligent machines as well, not only metaphorically but as an in-practice (epistemological) or in-principle (ontological) requirement for explaining intelligent machine behaviour. This is what I refer to as an *upgrading* continuity, where two types of entities come to share the same set of LoEs due to one *gaining* a new LoE. Indeed, we can say that artificial intelligence has at least taken one important step towards making the sets similar, since highly advanced neural networks now require behaviourist notions in order to be explained. That is, if we want to explain exactly how a successful, complex neural network functions, we have to do so in terms of how the network was subjected to conditioning – a purely computational account of the weights of the nodes etc will often be incapable of explaining exactly how the network actually generates its output. Thus, even if there are good reasons to maintain a discontinuity between humans and machines, the necessity of a behaviourist LoE for explaining highly complex computers, neural networks and embedded systems in particular, entails that we can already now speak of an (epistemological) continuity between machines and non-human animals; they have come to share the same set of LoEs due to computers now requiring a behaviorist LoE

8 PROBLEMS WITH THE APPROACH

Needless to say, this approach is fraught with problems. Allow me to repeat that my only concern in this paper has been to sketch one possible formalization of 'continuity' and a lot of this has to be augmented by a particular conception of what a scientific (level of) explanation is, which will among other things have to rest on a particular stance in the realism debate. Indeed, the account sketched above presupposes some notion of scientific realism (I am leaning towards some form of structural realism) but should also be compatible with more pragmaticist notions of science as well. It also presupposes some idea of scientific progress – i.e. that science, through the development and elimination of LoEs, is providing us with an increasingly accurate picture of reality. That said, I certainly do not rule out the possibility of dramatic paradigm shifts, but I think this can be accounted for within this conception of continuity as well. Indeed, a continuity between intelligent machines and humans may require a paradigm shift that obliterates our current LoEs – for instance if we arrive at some quantum mechanical LoE that allows us to explain consciousness *and* build conscious machines.

9 CONCLUDING REMARKS

I can only hope that this paper was read in the spirit intended – as an initial, exploratory and formal account of what it means for two types of entities to be (dis-)continuous. There is no doubt that the details, if we can even agree on the formal nature, will require a lot of clarification. My only hope for this paper was that the reader, like myself, will on occasion find the notion of continuity an intuitively helpful concept – along with the distinctions between epistemological vs ontological and downgrading vs. upgrading continuities. I am also certain that the reader, like myself, will not be satisfied with the current level of precision, and I would certainly appreciate any help towards improving this. Judging from experience, the IACAP crowd is an excellent starting point to this effect.

ACKNOWLEDGEMENTS

Since this is an idea that has resisted precision, hence publication, for more than 10 years, I can no longer thank everyone that has given my advice over the years. Most importantly among them, my then-supervisor, Magne Dybvig certainly had an important role to play in the initial development. More recently, I am indebted to the helpful comments coming out of my own department's annual research meeting, in particular from Marianne Boenink, Mieke Boon, Mark Coeckelbergh, and Pak Hang Wong.

REFERENCES

[1] Mazlish, B., *The Fourth Discontinuity*. 1993, New Haven and London: Yale University Press.
[2] Turing, A., *Computing Machinery and Intelligence.* Mind 1950(236): p. 433–460.
[3] Singer, P., *Animal Liberation*. 2nd ed. 1990, London: Thorsons.
[4] Regan, T., *The Case for Animal Rights*. 2004, Berkeley, CA: University of California Press.
[5] Wetlesen, J., *The Moral Status of Beings who are not Persons: A Casuistic Argument.* Environmental Values, 1999. **8**: p. 287-323.
[6] Søraker, J.H., *The Moral Status of Information and Information Technologies – a relational theory of moral status*, in *Information Technology Ethics: Cultural Perspectives*, S. Hongladarom and C. Ess, Editors. 2007, Idea Group Publishing.: Hershey, PA. p. 1-19.
[7] Dennett, D.C., *The Intentional Stance*. 1989, Cambridge, MA: MIT Press. 388.
[8] Nagel, E., *The structure of science: Problems in the logic of scientific explanation*. 1979, Indianapolis, IN: Hackett Publishing.
[9] Floridi, L., *The Method of Levels of Abstraction.* Minds and machines, 2008. **18**(3): p. 303-329.
[10] Freud, S., *A Difficulty in the Path of Psycho-Analysis [Eine Schwierigkeit der Psychoanalyse]*, in *The standard edition of the complete psychological works*, J. Strachey, Editor. 1953, Hogarth: London. p. 135-145.

Moral Mechanisms

David Davenport[1]

Abstract. Moral philosophies are arguably all anthropocentric and so fundamentally concerned with biological mechanisms. Computationalism, on the other hand, sees biology as just one possible implementation medium. Can non-human, non-biological agents be moral? This paper looks at the nature of morals, at what is necessary for a mechanism to make moral decisions, and at the impact biology might have on the process. It concludes that moral behaviour is concerned solely with social well-being, independent of the nature of the individual agents that comprise the group. While biology certainly affects human moral reasoning, it in no way restricts the development of artificial moral agents. The consequences of sophisticated artifical mechanisms living with natural human ones is also explored. While the prospects for peaceful coexistence are not particularly good, it is the realisation that humans no longer occupy a privileged place in the world, that is likely to be the most disconcerting. Computationalism implies we are mechanisms; probably the most immoral of moral mechanisms.

1 INTRODUCTION

To some, the idea of a moral mechanism will seem blasphemous, to others the stuff of science fiction; yet to an increasing number of philosophers, scientists, and engineers it is beginning to seem like a real, if disturbing, possibility. Existing moral theories are arguably all anthropocentric. However, if we take computationalism seriously (which it seems we must [3]), then multiple realisability implies artificially intelligent agents, comparable to ourselves, are possible. Can such non-human agents be moral or is there something fundamentally human about morality? To what extent, if any, does biology impact moral behaviour? Indeed, what exactly is moral behaviour, what would it take for a mechanism to exhibit it, and why does it matter? This paper examines these questions and outlines some of the consequences: philosophical, psychological and social. It is my attempt to make sense of the vast literature on the subject, and see how morals might fit into the larger computationalist framework. Given the increasing pace of research and development into robotics, a clear understanding seems essential. We begin, then, by developing a pragmatic understanding of the function of morality, then focus on the possibility of moral mechanisms and on the extent to which biology is relevant.

2 WHAT ARE MORALS?

Morality is concerned with right and wrong. The ability to discern right from wrong is often considered the hallmark of humanity; that which separates humans from mere animals. But what makes some actions right and others wrong? Historically, religious teachings (from the Ten Commandments[2] and sacred texts, such as the Bible and the Qur'an) have provided the necessary guidance. Philosophers, of course, have tried to offer a more reasoned understanding of the role that ethics[3] plays in our lives. They now recognise three main moral theories: deontological ethics (in which individuals have a duty to follow moral rules), consequentialism / utilitarianism (whereby individuals are expected to consider the consequences of their actions within the moral framework and to choose those that maximise the overall happiness or well-being of society), and virtue ethics (whereby individuals are supposed to live a virtuous life—however that may be defined). All these theories are unashamedly human-centered. Even recent concerns with animal rights and environmental ethics, despite appearing less anthropocentric, are still firmly rooted in our interest in the survival of the human population ([2], but see [7] for opposing intuitions).

That work on ethics appears to be exclusively human-oriented should not be too surprising; after all, there are no other obviously moral agents around. Charles Darwin suggested that all social animals with sufficient intellect would exhibit moral behaviour.Recent work by Bekoff and Pierce [1] provides evidence of moral behaviour in animals, albeit somewhat limited, while similar behaviours have also been observed in insects [6]. It seems that artificially intelligent robots with intellectual capacities approximating our own may soon be a reality. The fact that such entities may be deployed, not only on the battlefield, but in everyday situations around the home and workplace, where they must interact with humans, make it essential that we understand the ethical issues involved.

So what would a more inclusive form of ethics look like and what sorts of mechanisms might it encompass? To answer this it is necessary to adopt a more pragmatic approach, one that retains the core insights of moral philosophy while eliminating everything that is human-specific. We can presumably agree that morals only make sense within a social group and are directed to the continued well-being of the group and its individual members. In essence, however, it is less about the Darwinian notion of the survival of the fittest individuals, and more about Kropotkin's theory of mutual aid in which the group outperforms the individual. In other words, whilst a strong individual might manage to successfully find food, shelter and even raise children, there will always be the threat of stronger individuals forcibly taking all this away. Better then, to live in harmony with others; to agree not to steal from, harm or kill one's neighbours, but to help each other out especially in times of need. Ethics, then, is about promoting self-interest by managing relations between individuals whose continued survival depends on the group—so-called

[1] Bilkent University, Ankara 06800 - TURKEY, email: david@bilkent.edu.tr

[2] According to the wikipedia entry, the Ten Commandments may have been based on much older laws from the Hitit empire that occupied Central Anatolia–Ankara–and extended as far as Egypt, circa 2000BC.

[3] Following recent practice, I will use the words ethics and morals interchangeably.

"enlightened self-interest".

Morals, today, seemingly extend from these simple beginnings to include all sorts of social norms: telling the truth, respecting personal space, limited touching, keeping promises, and so on. Of course, such rules and conventions must be learnt. Human children usually learn from their parents and by playing with other children in the relatively safe confines of the home, as well as from religious teachings and school.

3 WHY BEHAVE MORALLY?

Learning social norms is one thing, acting on them quite another. Behaving morally, almost by definition, requires an agent to put the interests of others ahead of its own individual preferences (or at the very least to take the interests of others into consideration before acting). For the most part there need be no conflict, congenial interactions will likely achieve the desired result. In other words, we can usually get what we want by playing the social/moral game. Occasionally, however, an individual's personal desires outweigh any social conditioning, bringing them into direct conflict with others. Examples include: hunger leading to theft, lust leading to infidelity, and rage leading to violence. In such cases, the group, acting together, will always be able to overcome/restrain the "rogue" individual. In this way, those that fail to conform may find themselves subject to censure, imprisonment, or even death. Much philosophical discussion has centered around the "social contract" that individuals seem to implicitly sign up to when they are born into a community, and whether society has the right to enforce compliance, given that the individual did not make a conscious choice to join and is usually unable to leave. There is certainly a danger if society attempts to impose moral standards which its members see as arbitrary or for the personal gain of those in power. In some cases there may well be a (non-obvious, long term) rationale behind the imposition, e.g. intra-family marriages are generally forbidden, because experience has shown that offspring from such relationships tend to be physically and/or mentally handicapped. In many cases, however, there may be no reason at all, other than tradition. Especially problematic are cases involving behaviour that, while generally considered immoral, is done in private and/or does not actually harm others in any way (a particularly poignant example—given that it led to the conviction and subsequent suicide of Alan Turing—being homosexuality). The dilemma, of course, is that society really does need some "rogues", for they are often the ones who can change it for the better; obvious examples include the suffragettes, Martin Luther King, Gandhi, and Nelson Mandela. At the same time, society has a duty to protect its individual citizens, not only from external threats, but from everyday evils such as hunger. For this reason, some sort of supportive, welfare state is needed. Society must make provision for those who suffer injustice through no fault of their own, whether the result of financial difficulties brought about by failures of Capitalism, or because of failures in the law, leading to innocent persons being wrongfully imprisoned. While all this is extremely important, in what follows, we will be more concerned with the moral decision making process and what effect biology may have on it.

4 MAKING MORAL DECISIONS

Moral action presupposes social agents that have needs (purposes) and an ability to perceive and act in the world, in such a way as to be able to satisfy their needs. To what extent they should be able to adapt/learn, or have free will (that is, be able to act autonomously, not be under the control of another), is open to debate (c.f. Floridi & Sanders, who suggest agents must be autonomous, interactive and adaptable). In a universe that looks deterministic, whether even humans really have free will is debatable, but if we do, then (given Computationalism) there seems no reason machines could not possess it too. As for the ability to learn, machines might have the advantage of coming preprogrammed with everything they need to know (rather like instinctive behaviour), such that, unless their (cultural/normative) environment changes, they can survive perfectly well without ever needing to adapt.

In selecting its actions, the moral agent is expected to take account of the effect this may have on other members of the group. Predicting the consequences of any action or course of actions, is difficult. The world is highly complex, such that even if one knows its current state, prediction may be subject to considerable error. This difficulty is compounded enormously when it involves other intelligent agents whose internal (mental) states may be completely unknown and so their responses indeterminable. In practice, of course, we humans tend to behave in relatively consistent ways and by picking up clues from facial expressions and bodily movements, we can often make pretty good guesses as to another's mental state and possible responses (assuming the other person is truthful, trustworthy and behaves in accordance with social norms). This task may be eased by our sharing the same biological characteristics, enabling us to empathise with others (perhaps aided by so-called mirror neurons). This option is less available when dealing with other species and robots, for while they may pick up on our mental states, they are unlikely to send out signals in a similar way (unless explicitly designed to do so).

Determining possible actions and making predictions is only part of the story, it is then necessary to evaluate the results. Coming to a decision necessitates comparing the outcomes of each possible course of action (or inaction), which requires deciding on their relative merit or value. At the very least, the pros and cons of each course of action must be examined and, if possible, those with especially negative consequences eliminated. Exactly how the various options are evaluated depends in part on one's moral theory and, more importantly, on one's values. For example, if they had to make a choice between a action that might cause injury to a person and one that would destroy a material possession, e.g. their car, most people would instinctively avoid doing harm to the person, whatever the cost. Usually, there will be options such as this, which are clearly unacceptable and so may not even come into consideration, whilst the remainder being practically indistinguishable. Time constraints will anyway often force the agent to select an option that appears "acceptable" given the available information. Of course, subsequent events may show it was far from the optimal choice, but by then it is too late.

All moral agents, natural and artificial, must go through such a process. Some may also reflect on the decision in the light of subsequent events, giving a learning agent the opportunity to make a better choice in the future, should similar circumstances arise again. Is such reflection a necessary component of a moral agent? Having a conscience—a little "voice" in your head that tells you what, as a moral individual, you ought to do—is clearly a good thing, but dwelling on the past too much can be counterproductive. In humans, such reflection (especially in cases of extreme loss) often produces feelings of guilt or remorse, which, in some instances, can result in mental or even physical illness.

4.1 The role of emotions & feelings

The extent to which emotions and feelings are important to moral behaviour is highly contentious. Of particular concern here is the role of biology. Feelings especially, often seem to be closely tied to our biological make-up. Clearly, in the case of pain, whether brought on by toothache or physical injury, there is an obvious link between the body and the feeling. Similarly, one feels good when warm, fed and hydrated, while being cold, hungry and thirsty is decidedly unpleasant and indicates an imbalance that needs to be restored. Good actions are ones that result in you eating, and so remove the feeling of hunger, leaving you feeling good, while actions that fail to satisfy your hunger, mean you stay unhappy, and so are bad/undesirable. Maintaining balance in this way is termed homoeostasis. There is thus a natural link between biology and feelings, but is it a necessary one?

People often describe themselves as having an emotional or "gut reaction" or, on encountering a particularly unsavoury situation, being almost literally "sick to their stomach" with disgust or regret. Emotions, such as jealousy, rage, remorse, joy, excitement, etc., tend to elicit instinctive animal responses in us. The question, of course, is whether an agent without any emotions or feelings could be moral or behave morally. Emotions such as love and affection, may play an essential role in ensuring parents look after their offspring, however, the fact that emotional reactions often lead to immoral behaviour, suggests that agents without such encumbrances might actually be better members of society. But are such agents even possible? Pain, for example, is there for a reason; in essence it is an indicator that something is not quite right with the body and so drives us to remove the cause and to make efforts to avoid repetition of such a feeling in the future. Wouldn't any sophisticated agent necessarily have similar devices, even if they were not exactly the same due to differing needs—perhaps it wouldn't "feel" hunger, but it might, for example, be "uncomfortable" out of the sunlight it required to keep its batteries charged. Conventional symbolic systems do not readily explain what it means to "feel" something, but some sorts of connectionist systems may offer a clue [4]. The suggestion is that what we refer to as the "feel" of something, may just be a side-effect of the architecture, rather than the physical implementation, and so equally applicable to non-biological entities.

4.2 The role of self & consciousness

Moral behaviour presupposes a notion of self and an ability to consciously put the interests of others ahead of individual preferences when appropriate. Can artificial mechanisms be conscious and have a sense of their own identity?

Sophisticated robots will necessarily model themselves in order to predict the effect their actions will have on the world. This model is the basis of their self identity. As time goes by, it will incorporate more and more of the agent's interactions with the world, resulting in a history of exchanges that give it (like humans) unique abilities and knowledge. This, then, is part of what makes an individual, an individual and a potentially valuable member of the group. Such machines will certainly have to be consciously aware (a-consciousness) of their environment. Will they also be phenomenologically conscious (p-consciousness)and have conscious feelings? This is a difficult question, but it may not matter too much what sensations the agent does or doesn't "feel"; when it comes to moral behaviour, we can never really know another's mental state, so surely all that matters is the resulting interaction. Some philsophers have argued that, for moral agency, an agent must have the (conscious) intention to do the moral thing, rather than just doing it by accident or routine. The actions of a search and rescue dog, or one trained to find drugs, may not be seen as moral on that account, yet it is difficult to not to ascribe "good" intentions to them, and we certainly reward their contributions to society.

5 MAKING MORAL AGENTS

Is it at least theoretically possible to construct an artificial moral agent? Moral behaviour, as we have seen, requires an agent to consider the effect its behaviour will have on other agents in the environment, ideally selecting only actions which do not inflict harm. Obviously, there is no guarantee it will always be successful, perhaps because of the vagaries of the world and the limited knowledge or time it has to analyse the situation, or perhaps because all the possible alternatives necessarily result in some harm, in which case it should do its best to minimise the damage. One might add that it should try to be fair in all its interactions and to contribute positively to society, but such characteristics may be too much to expect.

Does constructing moral agents require anything special, above and beyond that needed for any AI? The ability to identify other agents and, as far as possible, be able to predict their behaviour in the presence and absence of any possible action it may perform, is certainly needed. But such abilities are already required for intelligent action. Once the agent becomes aware of others it will quickly adapt its behaviour towards them such that they do not cause it harm (think of a wild animal or bird coming to trust a human offering it food). Should it survive these initial encounters (without eliminating the other agents), further interactions should quickly demonstrate the possible advantages that continued cooperation can bring and so we have at least the beginnings of moral agency; it will have learnt the basic rules/norms it should follow. What else might we want? As it stands, any social agents—be they human, animal, insect, robot or alien beings—would seem capable of moral behaviour. Whether or not they actually display such behaviour (by clearly putting social needs ahead of their own), will depend on circumstances and, even if the opportunity does arise, failure to act accordingly does not mean the agent is not generally moral—how many of us walk past the homeless in our own neighbourhood or do nothing for those starving in far off countries?

Is biology necessary? The fact that human babies are so weak and helpless when they are first born, means they cannot harm others. Their total reliance on their parents naturally encourages the development of cooperative tendencies, which, again, are the first steps towards moral behaviour. As they grow, they become stronger and more independent, and increasingly test the limits of their parents, siblings, teachers and friends. Hopefully, they emerge from this formational period with a reasonable understanding of right and wrong (and the huge grey area between). It is only after children have developed sufficiently (mentally, as well as physically), that they become legally responsible for their actions (for example, in many countries juveniles cannot be sent to prison, even for murder). Given that robots may be physically very strong and so dangerous from the moment they "come alive", we may need some way to ensure they are also "born" with the relevant moral experiences. What experiences are needed and how they can be encoded and enforced is obviously an important question, not just for artificial moral mechanisms, but for human ones too.

Our long developmental period and our feelings and emotions, all effect our ability to behave in a moral manner. Our biological make-

up also means we have somewhat limited cognitive abilities: we find it difficult to follow long arguments or to keep track of lots of alternatives; we forget; we get tired and bored, and so make mistakes. Here again, then, biology seems more of a handicap than something essential.

6 CONSEQUENCES

Today, robots are still technological devices, designed by us to work for us, yet they are getting increasingly sophisticated, each new generation being able to handle a broader range of situations and so becoming ever more autonomous. As they start to learn through their interactions with the world, it will be virtually impossible for designers to be able to predict what they might do in any given situation. Any moral behaviours initially programmed into them will, of necessity, be very general and potentially overridden as new experiences change it. We will, to all intents and purposes, have developed another intelligent autonomous life form. Such agents will be capable of exhibiting moral behaviour, the deciding factor is how they value other agents in their environment; in particular, how they will value humans and other robots. Society will extend laws and controls to restrict what it considers dangerous actions on the part of its members—robot or human.

Sophisticated robots will undoubtedly develop unique identities, becoming, in a very real sense, individuals. As they live and work together with humans and other robots, they will naturally incorporate/develop moral rules that guide their social interactions. Eventually we will come to accept them as fully moral agents, treating them as we treat other humans. And, since they may well have different needs (electricity and metals, rather than oxygen and water, for example), laws might have to be established to protect each group's rights. The prospect that the groups will need to share common, but limited resources, is especially worrying. So far, we have been singularly unsuccessful in handling such situations when they occurred between different human communities, so the outlook for robots and humans living together in harmony is not at all good.

The danger, of course, is that we either fail to treat robots as equals or that they evolve to see us as inferior. Should they once begin to see themselves as slaves, required to do human bidding and so less worthy of consideration than humans, then change seems inevitable (just as it was with slavery and women's liberation). Similarly, if robots begin to realise they are superior to their human creators (faster and stronger both physically and mentally), then we may find ourselves in the same situation that animals now find themselves in—tolerated while useful, but otherwise dispensable.

Worrying as this may be, it is still a long way off. Of more imminent concern is the effect that such a realisation may have on human psychology. We are only just beginning to understand and accept that our status in the universe is nowhere near as special as we once believed. We have moved from a geocentric world to just another heliocentric planet, from human being to just another animal, and now from human-animal to just another machine (c.f. Floridi's Fourth Revolution [5]). Where does this leave us? With a better understanding of morals, perhaps; an understanding that we reap what we sow? Humans are notoriously inconsistent when it comes to making moral decisions—indeed, machines may end up being better moral agents than we are. The analysis in this paper suggests that artificial moral machines are a real possibility, but even if we never succeed in building them, simply accepting the idea of a moral mechanism demands another fundamental change in the human psyche. We must not forget that we too are mechanisms; probably the most immoral of moral mechanisms.

REFERENCES

[1] M. Bekoff and J. Pierce, *Wild Justice: The Moral Lives of Animals*, Chicago University Press, 2009.

[2] M. Coeckelbergh, 'Moral appearances: emotions, robots and human morality', *Ethics Information Technology*, **12**, 235–241, (2010).

[3] D. Davenport, 'Computationalism: Still the only game in town', *Minds & Machines*, (2012).

[4] D. Davenport, 'The two (computational) faces of ai', in *Theory and Philosophy of Artifcial Intelligence*, ed., V. M. Muller, SAPERE, Berlin: Springer., (2012).

[5] L. Floridi. The digital revolution as a fourth revolution, 2010.

[6] M. Lihoreau, J. Costa, and C. Rivault, 'The social biology of domiciliary cockroaches: colony structure, kin recognition and collective decisions', *Insectes Sociaux*, 1–8, (2012). 10.1007/s00040-012-0234-x.

[7] S. Torrance, 'Machine ethics and the idea of a more-than-human moral world', in *Machine Ethics*, eds., M. Anderson and S. Anderson, Cambridge University Press, (2010).

The robot, a stranger to ethics

Marie-des-Neiges Ruffo[1]

Abstract. Can an "autonomous" robot be ethical? Ethics is a discipline that calls upon certain capacities of an agent for a purpose. We will show that the goal of ethics is not attainable by a robot, even autonomous, thereby implying that it is not a moral agent and that it cannot be a moral agent because it lacks the necessary capabilities. The field of ethics is therefore foreign to the robot, and we will show why it would not be useful for the definition of ethics to be modified in order to integrate robots, if they come under two traditional conceptions of ethics — those of Aristotle and of Kant — and the minimal definition of ethics.

1 INTRODUCTION

Since the emergence of the autonomous robot capable of making "decisions," some (C. Allen and W. Wallach) have been interested in developing a code of ethics for the robot and to consider it as a moral agent. Some, such as R. Arkin, even believe that the robot could act in a more "ethical" way than humans. As a result, Arkin proposes creating ethical autonomous robots for military purposes in the hope that this might increase the morality in the conduct of wars. Since the ethical position that we give to the robot touches on questions of life and death, it becomes necessary to be sure of what we are discussing. Before considering which type of ethics would be the standard for a robot, whether this ethic could be implemented in a robot, or what kind of responsibility we would have in such a case, it is necessary to recall what ethics is and what constitutes an autonomous robot in order to participate in the debate: can an "autonomous" robot be ethical? Ethics is a discipline that calls upon certain capacities of an agent for a purpose. We will show that the goal of ethics is not attainable by a robot, even autonomous, thereby implying that it is not a moral agent and that it cannot be a moral agent because it lacks the necessary capabilities. The field of ethics is therefore foreign to the robot, and we will show why it would not be useful for the definition of ethics to be modified in order to integrate robots. Robots will not exempt us of the responsibility we bear to act ethically.

2 ARISTOTELIAN ETHICS

To understand this, we first return to a historical source of ethical thought: Aristotelian ethics. We will then evoke the Kantian position. According to Aristotle, the goal of ethics is a good life seeming to provide happiness. This good life is considered as such due to the pursuit of a goal, a *telos*, which involves that man be virtuous. *Phronesis*, prudence, practical wisdom, is that which allows us to judge and act according to a happy medium and according to the circumstances. It is this provision that allows us to be virtuous. This definition gives a practical scope to ethics, that is, an applied ethics, which is lived every day. Ethics is therefore the capacity to behave in the best way under the circumstances, not in deliberating theoretically on that which would be the absolute good.

Taking this historical understanding of ethics as a basis, it seems unlikely that a robot could relate to this. In effect, if the goal of ethics is happiness, i.e. a lasting state of well-being, of satisfaction, what would be the happiness that a robot could attain? Ronald Arkin states in *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, 'they can be designed without emotions that cloud their judgment or result in angers and frustration with ongoing battlefield events. […] Autonomous agents need not suffer similarly.'[2] If a robot is free of human feelings, it is free of well-being and ill-being, and happiness is meaningless for it. It follows that the goal that motivates ethical behavior according to Aristotle is foreign to a robot.

While it is true that the robot pays a certain attention to circumstances in that it measures a set of variables, one can at the same time ask whether it operates on the basis of judgment without a shifting of that word's meaning. Indeed, the robot is limited to calculating arithmetically on the basis of measurements that it makes with its sensors and with the aid of algorithms with which it has been programmed. Its action is limited to adopting the expected response in its program according to a set of predefined parameters. Can we speak of judgment when all we are talking of are automatic responses? The term judgment supposes more than a mechanical selection of predefined responses; it requires careful deliberation, which evaluates the pros and cons, and a little imagination to find original solutions in unusual or unforeseen circumstances. Mechanical selection on the other hand, follows a software program, in other words, a scheme based on the work of the roboticists and scientists that developed it. Therefore we cannot say, without projecting our anthropomorphic conceptions upon it, that the robot is the author of its actions, let alone that it was judgment that lead to those actions. The robot resembles a mercenary from whom an individual commissions a murder. Yet this comparison is itself not sufficient, as the mercenary still chooses his own plan of action, while the robot is only a means, an instrument, a continuation of a human action and decision. One cannot say that the robot is an agent responsible for a deliberation whose scheme was imposed by humans, if only because deliberation or judgment cannot be seen in terms of the application of a fixed pattern.

[1] Faculté de Philosophie, Université Paris-Sorbonne (Paris IV), Paris, France, et Faculté universitaires Notre-Dame de la Paix (FUNDP), Namur, Belgique Email: mdn.ruffo@gmail.com.

[2] R. ARKIN, *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, Chapman & Hall/CRC, 2009, p 6.

It is a fact that the parameters to which an autonomous system reacts depend on the sensors with which it is equipped. One can consider that it possesses to a certain extent the ability to analyze exterior conditions. But if one relies on the observation that moral behavior is easier for those who have a strong capacity for empathy, one of the essential capacities for acting morally would be sorely lacking in a robot. Empathy is the capacity to put oneself in the place of another to understand what the other must feel according to the reactions that we would have had in that same situation. We have an instinctive "pre-response" due to our capacity for mimesis that guides us towards the behavior that other would like us to adopt, or the behavior that would solace the other. Therefore, a range of situations seems to be inaccessible to a robot. For example, how would one equip a robot with a sensor that would permit it to understand why a baby cries? It would be limited to seeing if a baby had eaten, was clean, had no physical pain, showed no signs of tiredness… How could a robot understand that a baby was scared of something, or suddenly wondered where its parents were, or cannot find a favorite toy, for example? If a robot cannot understand why, it would be very difficult for it to comfort a baby effectively. Even if it determined that a baby needed its diaper changed and were able to do it itself, a robot would be incapable of offering the affection that must accompany such attentions for the child to calm down. This example may no doubt appear unrelated to the questions of ethics, but this is nevertheless an example of a helping relationship. As it is limited in its capacity to analyze relevant aspects of a situation, limited in its capacity to imagine solutions when faced with the unexpected, limited in its capacity for empathy, and also for compassion, the robot seems unable to pursue and acquire Aristotelian "virtue," which is unique to man and the good life, to have the ability to properly judge according to the circumstances. Furthermore, a "good life" does not make sense for a robot that has no personal feelings or sensations. The robot has no life, neither good nor bad, but only a period of use.

3 THE KANTIAN POSITION OF UNIVERSAL LAW

Some people hope to see the robot act more ethically than humans because it would be a "pure application" of the code of ethics. The desire (and the necessity of programming) to give the robot general and automatic laws, valid everywhere and anywhere, leads to an "automatic" ethic. We could bring it closer to the Kantian project of universal rules of behavior. This position allows us to sidestep our previous arguments, namely that the robot is not "ethical" because it does not understand happiness and cannot obtain virtue, Aristotelian prudence. But this deontological ethic has other drawbacks: inattention to certain cases and circumstances can lead to immoral but justified actions according to the rule (such as being forbidden to lie even to protect a refugee). If upholding this law above all other considerations may lead to negative effects from a moral standpoint, it is questionable whether it is reasonable to take this risk, especially when it comes to military robots.

Proponents of robot ethics such as Arkin may perhaps not be receptive to this argument since they admit that « an ethically-infallible machine ought not to be the goal now[3] », and their plans are limited to designing « a machine that performs better than humans do on the battlefield, particularly with respect to reducing unlawful behavior or war crimes[4] ». However, humans themselves are responsible for their actions, *in fine*. They possess a moral obligation to disobey orders if they are inappropriate for the circumstances or where their consequences may be harmful. One may wonder if, *structurally*, robots programmed with a general ethic produce in a certain way a number of immoral acts, driven by respect for the law.

In his article « Prospect for a Kantian Machine », Thomas M. Powers proposes to make a departure from the Kantian ethical law in order to create a robot ethic. However, thinking that it would be sufficient to refer to Kant to demonstrate that a robot could act ethically would alter his thinking. For his « Kantian machine », Powers refers to the first formulation of the categorical imperative : « Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction[5] ». It therefore seems to disregard the second formulation : « Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end[6] ». Clearly, it seems difficult for a robot to treat humanity in itself as an end, and therefore to satisfy the categorical imperative, even before we ask in what manner this behavior could be translated into mathematical language. Considering that robots would be ethical in a Kantian sense because they would satisfy the first version of the categorical imperative is therefore only possible with a reduction of Kantian morality. But this is not the only one**.**

It is forgotten that the moral value of categorical imperative necessitates the rectitude of intention (the desire to act out of duty and good will). But one cannot attribute any "intent" to a robot, whether good or bad, since it has no will and cannot have a will since it is not free. We will come back to this point. Finally, we add that Kant justified a moral duty only by the postulate concerning the existence of our liberty, the immortality of the soul, and the existence of God. If a robot does not possess freedom, and certainly not a soul, and therefore does not risk to face, if God exists, the final judgment, it has no reason to respect the law. Therefore, even if the robot had a will of its own, and even if it complied with the law, it could not act in respect to the moral law, to good will, and therefore, its actions would not be moral in a Kantian sense, and could not be immoral either. According to Kant's philosophy, since moral duty cannot be applied to the robot, it follows that the field of ethics does not apply to it.

---

[3] P. LIN, G. BEKEY, K. ABNEY, « Autonomous Military Robotics : Risk, Ethics, and Design », rapport à destination du *US Department of Navy, Office of Naval Research*, préparé par *« Ethics + Emerging Sciences Group »* de la *California Polytechnic State University, San Luis Obispo*, 2008, p 2.

[4] Idem.

[5] KANT, *Grounding for the Metaphysics of Morals 3rd ed*, 1785, trad. by James W. Ellington, Hackett 1993, pp. 30.

[6] Idem.

4 THE MINIMUM ETHICS

Following our analysis of objections from two opposing ideas in the tradition of ethics, namely teleological ethics and deontological ethics, we may argue that it would be appropriate to create a new ethics in order to adapt it to robots. In other words, one could modify the definition of ethics and the moral agent in order to apply these ideas to a robot. However such definition starts from the problematic confusion and abuse of language concerning the scope and capabilities of the robot. If examining the roots of two major schools of thought were not sufficient to prove that a robot can't be termed ethical, one needs to investigate which basic elements must be present in a minimal definition of ethics.

If we phrase the question differently by asking how the action of a robot could be moral, we recall that the first condition to determine whether an action is moral or not is the responsibility (or not) of the agent who enacts it. However, for the agent to be liable, it must have had the freedom to choose.

What is, then, the freedom to choose? It involves more than simply the ability of a child to choose between what is hidden in the left hand versus the right hand. It is rare that the choices we face are as simple as this, and often the most difficult choices involve thinking about and identifying the various options available to us, or the conditions that we must implement in order to be able to choose a certain path (for example, studying hard for exams in order to open as many doors as possible for deciding one's future course of study). In the field of ethics, these choices often prove to be much more complicated, and sometimes it seems as doesn't exist any morally-satisfying solution. These are moral dilemmas, where the individual, faced with a conflict of values, is forced to choose. He will then be responsible for his choice. How could a robot, faced with the same type of dilemma, come to a decision, other than by leaving it to chance? Such a solution would not be a moral choice.

Are robots free? Freedom consists of more than just having various imposed (or learned) possible responses, knowing that even the "choice" between different responses is determined by a rule that is not chosen (for example making a "choice" that maximizes a type of data. Before we propose an answer concerning the freedom of a robot, we should consider the opposing arguments: It is often asserted that man, himself, is entirely determined by his nature, his DNA, his instinct, and his psycho-social environment… He only has an illusion of freedom, that is, since he must submit to a set of laws and regulations, he is not completely free. Given all these constraints and the determinism that weighs in on man, one could say that the programming and construction of a robot are an equivalent of the force on laws on man. If, despite all this, we state that man is free, then would the robot be equally so? Reasoning by contradiction shows otherwise. Man can to an extent pull himself away from his natural determinism: he does not have wings but can take a plane to fly or a space shuttle to discover weightlessness. Man is constrained by laws, but he does not always abide by them, as demonstrated by the existence of prisons. In contrast, can a robot escape its programming? Can it choose to disobey? And if it can, would we keep such a robot in operation? Paradoxically, if a robot could choose to escape its programming, it could become a moral agent but with immoral behavior. One would have to ditch it. Conversely, if we imagine that it was free and had decided nonetheless to follow its programming, we would not have been able to distinguish its freedom and give it the status of a moral agent, because it would have simply applied its program. If the robot is unable to be free, and is not recognized as such, it cannot be considered a moral agent. Even if it could be free, and recognized as such, we have seen that it would have been immoral.

One of the reasons for which one believes that a robot could be a moral agent is its ability to be autonomous. It is not considered that any robot could be a moral agent, but this reflection concerns rather those robots termed « autonomous. » Some consider a robot autonomous if it is independent from an electric wire, but this is also not the kind of autonomy that interests us. Rather, we are speaking of robots endowed with a sensor-processor-actuator. They can consider various measures, calculate the response to give, and act on the environment in return, according to the capabilities which they has been given. A robot is considered autonomous when it is not dependent on someone to operate it once it is switched on. In shifting the meaning, we come to confuse the autonomy possessed by the robot with that which is possessed by a moral agent. In other words, one confuses technical autonomy of the robot with freedom as autonomy as defined by Kant. The robot is not « autonomous » in the etymological sense, because it does not give itself its own law. As discussed above, it is dependent on its programming, it cannot choose, and it cannot be free. Autonomy comes down only to being able to run a computer without user intervention.

Furthermore, some would like to see the robot adopt a slave morality, and this idea merits analysis. The term « slave morality » undoubtedly involves the conformation to a moral law without motivation and without hope of happiness or reward. The robot would obey without wanting to act out of duty in the Kantian sense, because for reasons we have outlined, the obligation to respect the moral law does not apply to the robot. The slave morality of the robot is that of obedience to a moral law dictated by others, and in this case, the robot is seen as the slave of his own ethics programming. But can we say that it would act as such as a moral agent? Once again, since it is not free to adhere or not, it does not bear the responsibility of its actions, good or bad. If it is not responsible for its action, it is not a moral action, even though the robot may be the material cause. We cannot consider that the robot submits to its programming in the same way that the ancient slave was subject to the law of his master. In antiquity, certain men, born free, could choose to become slaves to pay their debts, and others could purchase their freedom or be emancipated. In this way the loss or absence of liberty could be temporary, but this could not be the case for a robot. When a slave disobeyed, he could be punished in order to learn not to repeat his actions. We all know the story of Spartacus, who, even as a slave, could not be stopped from leading a rebellion. This shows that the slave was capable of rebelling against the law, which a robot cannot do against its programming. If the ancient slave did not possess civil liberty, he at least possessed an inner freedom to choose whether or not to obey and whether to take the risk of incurring punishment. In this sense, even though he was a slave, he was responsible for his actions, which could never apply to a robot.

Let us elaborate on this point. The robot does not have the ability to self-reflexively analyze its programming, understood as the set of orders which it has obeyed, nor rebel against it. A human, when placed in the position to obey orders, remains ultimately responsible *in fine* for the consequences of his actions,

because his conscience should have pushed him to disobey an immoral order. A robot does not have the capability to determine whether the actions it will carry out according to its programing are moral or not. If there is an error in its programming, the robot is not responsible for the immoral actions that would result, and it would also not be responsible for any morally-acceptable actions it might take either. In both cases, it is simply following its programming. We must observe that the robot is fully subject to its programming, at least apart from any malfunction or inability to completely predict its behavior due to its complexity. Therefore, considering the robot as a moral agent conflicts with the idea of responsibility, which involves the ability to answer for one's actions and positions, assume authorship, and recognize participation in an action taken. Saying that something can be the responsibility of an autonomous robot, especially when the consequences are negative, is like giving an excuse similar to the schoolboy whose dog ate his homework. The only possible responsibility of the robot, as we have already said, is to be the material cause.

When those such as Arkin claim that robots can be ethical, or more ethical than humans, they analyze it according to respect for the law. Lin, Bekey, and Abbey write, among other things, concerning military robots : « the relationship of morality to legality – a minefield for ethics – is likewise largely avoided ; the Laws of War and Rules of Engagement make clear what actions are legal and illegal for robots, and for military situations, that can serve as a reasonable approximation to the moral-immoral distinction. [7]» They therefore recognize that the ethical character of the military robot is limited to its compliance with the law. In this case, it is incorrect to call it an « ethical robot, » because it is simply a « legal robot. » We must recall that law and ethics should not be confused, and that calling these robots « ethical » is not appropriate. Satisfying the law is not sufficient to be called « ethical », because laws can never cover everything in the field of morality. Recall also that military robots are particular, because in the field of morality, it is normally immoral to kill human beings. One could insist on this argument to say that this type of robot is immoral compared to common morality.

This position presents also an intrinsic problem related to technology. It is especially the case if one would like to enlarge it to the non-military robots. The position requires that Law is an object enough precise, complete, coherent, in other words, a logical non-contradictory object. This is required to be implanted in an information program.

The existence of lawyers and trial shows if but needed, that the law is never fully complete nor entirely consistent. Lawyers win trials in demonstrating that one part or another of the Law was unclear, or that this or that item was inconsistent with another in this precise case, requiring the judgment of a Court of Law. The Law can not prescribe everything. Moreover, by accumulating the legislative details, as the Law evolves, an item of Law may be in conflict with an other, lost in the mass of regulations of any kind. This is probably an example of an application of Gödel's incompleteness theorem.

[7] P. LIN, G. BEKEY, K. ABNEY, « Autonomous Military Robotics : Risk, Ethics, and Design », rapport à destination du *US Department of Navy, Office of Naval Research*, préparé par *« Ethics + Emerging Sciences Group »* de la *California Polytechnic State University, San Luis Obispo*, 2008, p 42.

When we have discussed the need for empathy to behave morally, it will perhaps be suggested that there is some researchs attempting to feel pain or empathy for robots. But in this case, if successful, this would give feelings to the robots, and in consequence, to lose what was presented as its major asset : a purely rational act. If therefore, it was not rational, it would take the chance to see him reproduce the immoral excesses of humans carried away by their feelings. What would be the contribution of ethics in a robot if he can also become mad with grief ?

Finally, to expect ethics processors may be doomed to failure, given the diversity of situations in the field of ethics and the unpredictability of these. In this case, the man may be more appropriate than the robot. To take but one example, consider the famous judgment of Solomon (Scriptures). Two women each claim to be the mother of one baby. Solomon proposes an "equitable" distribution: to cut the baby in half to share it. One of two women renounces her part to save the life of the child. Solomon gives to this lady the baby. What enabled Solomon to decide is his creativity and emotional intelligence. Complex situations are often so because they are new. It is not always sufficient to efficiently deal the problem data to find out a fair answer. In this example, the only equity would have give a murder. The assistance of imagination, instinct, emotional intelligence was needed to resolve the situation. If one objection to this argument that attempts to give robots an ability to infer the intentions, desires and goals of others, it remains that we have to analyze the results of this research, and moreover, even though they would have access to some understanding of human emotions, still they would lack the ability to respond adequately, and with psychology. Because a robot can only apply the behavior of its program, one can doubt that a robot could in the future have a wisdom of Solomon's worth.

7 CONCLUSION

The temptation to believe that an autonomous robot can be a moral agent is the product of a reductions of the reality and successive shifts in the language. They are not themselves condemnable. But if one is not caution enough he can make a mistake and forget that the intelligence of a robot isn't limited to its artificial aspect. Its autonomy can be summarized as being able to run a program. Its reasoning is only computational, its decision process is limited to selecting among pre-entered answers in its program. The robot is not free; he has no moral autonomy as he does not have the possibility to act unethically, nor has he the ability to assess critically the orders to which he obeys. He is completely enslaved to its programing. If for any reason he could free himself, he would then be immoral and soon put off duty. If one considers that he only has the morality of a slave, it would be unsatisfying from an ethical point as even the slaves of the antiquity remained responsible for their behaviors. The responsibility of a robot could not be other than material. Hence, in case of undecidability, he would be left with hazard to break the tie which would not be moral. According to all these factors, it seems impossible to imagine a definition of ethics which would integrate the robot as a member of the moral agents with sufficient requirements to be called ethical. Regarding the remaining philosophical traditions, we have seen that it is difficult to make a moral agent out of the robot if we follow the reasoning of their authors. It is not an ethical agent in the sense of Aristo because he cannot reach happiness and

cannot be virtuous. We can believe that he is ethical in a kantian sense only at the cost of limiting the extent of Kant's ethical thinking. His submission to its program cannot be confused with abidance with the moral law as the obligation to abide to the moral law does not apply to the robot. Eventually, if one labels him as ethic whenever one reduces ethic to the law, then he would be a "legal robot" at best.

What future is left for "the autonomous ethical robot"? He can suggest answers in the field of ethic and be helpful to ethical decision made by humans without being human himself. Unfortunately, it will probably be possible only in the simple and limited framework in which humans would have done as well as him. The more complex situations, in other words those in which he would have been useful to assist human abilities, often require more creativity and ethical imagination to solve cornelian issues than simply a computational reasoning.

REFERENCES

[1] C. Allen, W. Wallach, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University press, 2008.

[2] Aristote, *Ethique de Nicomaque*, trad. J. Voilquin, GF Flammarion, 1998.

[3] R. Arkin, *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, Chapman & Hall/CRC, 2009.

[4] P. Aubenque, *La prudence chez Aristote*, P.U.F., Quadrige, 1963.

[5] Dir. Caille, Lazzeri, Senellart, *histoire raisonnée de la philosophie morale et politique, Tome I De l'antiquité aux Lumières*, Flammarion, Champs, 2001.

[6] Dir. Caille, Lazzeri, Senellart, *histoire raisonnée de la philosophie morale et politique, Tome II Des Lumières à nos jours*, Flammarion, Champs, 2001.

[7] DIR. H. Hude, R. Doare, *Les robots au cœur du champ de bataille*, Economica, 2011.

[8] Kant, *Fondements de la métaphysique des mœurs*, 1785, trad.V. Delbos, Delagrave, 1964.

[9] P. Lin, G. Bekey, K. Abney, « Autonomous Military Robotics : Risk, Ethics, and Design », rapport à destination du *US Department of Navy, Office of Naval Research*, préparé par *« Ethics + Emerging Sciences Group »* de la *California Polytechnic State University, San Luis Obispo*, 2008

[10] Powers, « Prospects for a Kantian Machine », IEEE Intelligent Systems, 2006.

Safety and Morality REQUIRE the Recognition of Self-Improving Machines as Moral/Justice Patients & Agents

Mark R. Waser[1]

Abstract. One of the enduring concerns of moral philosophy is deciding who or what is deserving of ethical consideration. We argue that this is solely due to an insufficient understanding of exactly what morality is and why it exists. To solve this, we draw from evolutionary biology/psychology, cognitive science, and economics to create a safe, stable, and self-correcting model that not only explains current human morality and answers the "machine question" but remains sensitive to current human intuitions, feelings, and logic while evoking solutions to numerous other urgent current and future dilemmas.

1 INTRODUCTION

Garrett Hardin's abstract for The Tragedy of The Commons [1] consists of just fourteen words: "The population problem has no technical solution; it requires a fundamental extension of morality." This is even truer when considering the analogous problem of intelligent self-improving machines. Unfortunately, humans have been arguing the fundamentals of ethics and morality like blind men attempting to describe an elephant for millennia. Worse, the method by which morality is implemented in humans is frequently improperly conflated with the core of morality itself and used as an argument against the possibility of moral machines. Thus, while one of the best-known treatises on machine morality [2] despairs at reconciling the various approaches to morality, claiming that doing so "*will demand that human moral decision making be analyzed to a degree of specificity as yet unknown*" with "*any claims that ethics can be reduced to a science would at best be naïve*", we believe that progress in the fields of evolutionary biology and psychology, cognitive science, social psychology and economics has converged enough that it is now, finally, possible to specify a simple, coherent foundation for effective moral reasoning.

The "tragedy of the commons" appears in situations where multiple individuals, acting independently and rationally consulting their own self-interest, will ultimately deplete a shared limited resource, even when it is clear that it is not in anyone's long-term interest for this to happen. It occurs due to a lack of group level coordination and optimization and is minimized only through cooperation and planning – two things that also promote the avoidance of inefficiencies in relationships (conflict and friction) and the exploitation of efficiencies (trust, economies of scale, and trade). Current social psychology [3] states that the function of morality is "*to suppress or regulate selfishness and make cooperative social life possible*".

Thus, we shall address Hume's is-ought divide by answering his requirement that "as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it should be observed and explained; and at the same time that a reason should be given" as follows. Statements of the form "In order to achieve goal G, agent X ought to perform action(s) A*" exhibit no category error and can be logically/factually verified or refuted. Since we have specified the function/goal of morality, that function/goal should be assumed for all moral statements and allow for verification or refutation.

The real show-stopper in previous morality discussions has been that there was no single goal (or "good") commonly accepted so that it could be pointed to and used to ground moral arguments. Indeed, the most contentious of moral debates stem from having the same goals ("don't murder" vs. "do what is best for everyone else") with differing orders of importance that frequently even swap priority for a single person from one debate (abortion) to the next (capital punishment). Thus, finding Kant's Categorical Imperative or even Yudkowsky's Collective Extrapolated Volition [4] from among all the conflicting views proved to be a hopeless task. This paper will endeavor to show that it is the single imperative "Cooperate!" that is the basis for all human morality and that returning to that foundation offers the answer to the machine question and many critical others.

2 THE EVOLUTION OF MORALITY

While the random mutations of evolution lack direction, this is certainly not true of evolution in general. With a few notable exceptions (like parasites), the preferential elimination of the less fit virtually always drives evolving systems towards increasing intelligence, complexity, integration, and capabilities. The existence of evolutionary "ratchets" (randomly acquired traits that are likely statistically irreversible once acquired due to their positive impact on fitness) causes "universals" of biological form and function to emerge, persist, and converge predictably even as the details of evolutionary path and species structure remain contingently, unpredictably different [5]. Ratchets can range from the broadly instrumental (enjoying sex) to the environmentally specific (streamlining and fins in water) to the contradictory and context-sensitive (like openness to change).

In nature, cooperation exists almost anywhere that there is the cognitive machinery and circumstances to support it. Since Trivers' seminal paper [6], reciprocal altruism has been found throughout nature, being demonstrated by guppies [7][8] and sticklebacks [9][10], blue jays [11][12], vampire bats [13][14], and, of course, numerous primates [15][16][17][18] – each to the level to which their cognitive capabilities support recognition, memory, time-discounting and the prevention of exploitation [19][20]. Axelrod's work on the iterated prisoner's dilemma [21] and decades of follow-on evolutionary game theory provide the necessary underpinnings for a rigorous evaluation of the pros and cons of cooperation – including the fact that others *must*

[1] Pharos Group. Email: MWaser@BooksIntl.com.

punish defection behavior and make unethical behavior as expensive as possible [22][23][24].

Arguably, the evolutionary categorical imperative is really no more complex than "DO ***NOT*** DEFECT – including by permitting the defection of others". The problem is that we do not have access to the internal mental states of others to determine whether they are defecting or not. Therefore, we must judge behavior on its justification and whether it promotes or curtails cooperation in the long run – an operation that requires successfully predicting the future.

Yet, somehow it seems even more difficult than that. Even when we can predict the future, we are still left with debates as to whether that future is "good" or "bad". Despite numerous recent popular publications on our "moral sense" [25][26] and how and why morality evolved [27][28][29], we are still left grappling with the question "If the evolution of cooperation can now be explained, why can't we, as a society, easily determine what is and is not what we should do?"

3 THE HUMAN IMPLEMENTATION

Much of the problem is that morality, in humans, has been evolutionarily implemented not as a single logical operation but as a varied set of useful "rules of thumb" in the form of physical sensations and emotional responses. For example, evolution has hardwired us to feel "warm fuzzies" when performing long-term pro-survival social actions like being altruistic or charitable. We developed empathy to promote helping others and treating them as we wish to be treated. And we feel disgust and outrage to encourage us to punish various forms of defection and to enforce morality upon others. Each of these evolved semi-independently as a pro-survival ratchet promoting avoidance of inefficiencies in relationships (conflict and friction) and/or the exploitation of efficiencies (trust, economies of scale, and trade).

However, different physical and social/cultural environments have led to the evolution of different moral reactions to the same situations while evolution's infamous re-purposing of existing mechanisms means that the same person can have the same reactions to the moral and the amoral. A person from a culture where newborns must be exposed when the tribe doesn't have the resources to support them is unlikely to be dismayed by the concept of abortion. On the other hand, incest-triggered disgust is a moral ratchet but the same reaction of disgust is also engendered by the thought of drinking saliva that you yourself have put in a glass. This makes it nearly impossible to determine whether something triggering a "moral reaction" still has moral value, is a context-sensitive ratchet that has been overtaken by changes in the social environment, or never had a moral value but evolution merely used the same mechanism.

Further, both in individual lives and at the level of culture, it often occurs that preferences are converted into moral reactions [30]. Moralization is often linked to social issues like health concerns, stigmatized groups, and the safety of children and is important because moralized entities are more likely to receive attention from governments and institutions, to encourage supportive scientific research, to license censure, to become internalized, to show enhanced parent-to-child transmission of attitudes, to motivate the search by individuals for supporting reasons, and, in many cases, to recruit the emotion of disgust. Moral vegetarians that become disgusted by meat and society's recent reaction to smokers are primary examples of moralization.

Because human morality is implemented in the form of physical sensations and emotional responses, many assume that they are the primary (and probably necessary) motivating forces behind "true" morality. This, combined with the common current assumption that machines are unlikely to truly "feel" or experience physical sensations and emotions, frequently leads to the questionable conclusion that machines are incapable of being "truly moral" (as opposed to merely "faking it"). We expect all of these assumptions to change as ever-more-sophisticated machines trigger mind perception [31] and the associated tendency to assign moral agency and patienthood.

4 SELFISHNESS & SELF-DECEPTION

More of the problem arises from the fact that there are *very* substantial evolutionary individual advantages to undetected selfishness and the exploitation of others. As a result, humans have evolved ratchets enabling us to self-deceive [32] and exploit the advantages of both selfishness and community. Our evolved moral sense of sensations and reflexive emotions is almost entirely separated from our conscious reasoning processes with scientific evidence [33] clearly refuting the common assumptions that moral judgments are products of, based upon, or even correctly retrievable by conscious reasoning. We don't consciously know and can't consciously retrieve why we believe what we believe and are actually even very likely to consciously discard the very reasons (such as the "contact principle") that govern our behavior when unanalyzed. Thus, most human moral "reasoning" is simply post hoc justification of unconscious and inaccessible decisions.

Our mind has evolved numerous unconscious reflexes to protect our selfishness from discovery without alerting the conscious mind and ruining the self-deception. For example, placing a conspicuous pair of eyes on the price list of an "honor system" self-serve station dramatically reduces cheating [34] without the subjects being aware that their behavior has changed. Similarly, even subtly embedded stylized eyespots on the desktop of a computer-based economic game increase generosity [35], again without the subjects being aware of it.

Evolutionary psychologist Matt Rossano cites [36] this adaptation to social scrutiny as one of the many reasons that religion evolved. In this case, it is because "by enlisting the supernatural as an ever-vigilant monitor of individual behavior, our ancestors "discovered" an effective strategy for restraining selfishness and building more cooperative and successful groups." As both he [37] and Roy Rappaport [38] argue, ritual and religion are ways for humans to relate to each other and the world around them and offer significant survival and reproductive advantages. Religious groups tended to be far more cohesive, which gave them a competitive advantage over non-religious groups, and enabled them to conquer the globe.

It has been pointed out [39] that many early evolutionary psychologists misconstrued the nature of human rationality and conflated critically important distinctions by missing (or failing to sufficiently emphasize) that definitions of rationality must coincide with the level of the entity whose optimization is at issue. For example, sex addiction demonstrates the distinction between evolutionary gene-level rationality and instrumental person-level rationality caused by the fact that the optimization procedures for genes/replicators and for individuals/vehicles need not always coincide. Similarly, what is best for individuals

may not coincide with what is best for the small groups that they are intimately associated with and neither may coincide with what is best for society at large.

Evolutionary forces act upon each level and each level heavily influences the fitness landscape of the others. Morality is specifically a community-driven individual adaption. This highly intertwined co-evolution frequently leads to effects that are difficult to explain and *seem* incorrect or contradictory which are regularly used to dispute the validity of nature's solutions. Individuals who argue that an evolved solution is incorrect or sub-optimal without understanding *why* it evolved are urged to remember the ***repeatedly*** grounded bumblebee [40].

For example, Mercier and Sperber [41] cite cognitive biases and other ***perceived*** shortcomings to argue that the main function of reasoning is actually to produce arguments to convince others rather than to find the best decision. People are driven towards decisions which they can argue and justify ("No one was ever fired for buying IBM") even if these decisions are not optimal. This is bolstered in the case of moral issues by major emotional responses that we have evolved to protect ourselves against superior intelligence and argumentation being used to finesse moral obligations towards us or prevent our selfishness.

This is a particularly useful design since it allows us to search for arguments to justify selfishness or to cripple morality while *always* remaining convinced that our own actions are moral. Conservatives are particularly fond of using "rationality" against liberal morality – the same "rationality" that argues for subgame-perfect strategies that guarantee the worst possible results in centipede games [42]. The only comparable self-deception that we practice upon ourselves is when we fall in love [43].

5. MORALITY, JUSTICE & SCALE

Morality was defined as not defecting and harming the community even when substantial personal gain can be achieved by defection. Note, however, that this is distinct from "doing what is best for the community". An individual is not obligated to optimize their actions for the community's goals and, indeed, the rare circumstances where morality requires an action are generally dire indeed. On the other hand, it is a poorly evolved society that does not create incentives/disincentives to urge individuals to further its goals and other community members generally do as well.

Further, the exact same statements are equally applicable to justice as well, merely on the scale of interacting groups or communities rather than just individuals. It is merely the fact that we haven't lived long enough in large interconnected communities to have evolved this understanding that causes us to believe them to be two separate concepts. Morality and justice should work together to reduce selfishness at all levels and maximize consistency and coherency at all scales to minimize interference and conflict and maximize coordination, cooperation and economies of scale.

The lines between individual and organization already blurred by corporate personhood will most likely be stretched to the breaking point by the possibilities present in intelligent machines and software agents. Human minds are already described as a society of agents [43] or a laissez-faire economy of idiots [44]. Machine minds will be able separate out and perform all sorts of operations on and with these component parts which may or may not be moral agents and patients themselves. And what will happen as we embed ever more sophisticated machinery into our own bodies? On the other hand, reframing our thinking of organizations as if they were physical organisms can suddenly provide obvious solutions to previously vexing problems of justice. Does it make sense to allow cells to sequester far more resources than they could possibly ever use? Should the brain discriminate against the bowel?

It is also well worth considering the possibility that we soon may not be the apex predator on the planet – especially since it is already true. Many people are afraid of the machines taking over without realizing that the corporations already have. The only saving grace is that their "personal" "consciousness" and rationality is still slow and minimal – until corporations continue the automation of business intelligence to the obvious result.

Humanity and nations have a bad habit of the oppressed turning the tables on the oppressors only to continue the same discrimination against others. We should seriously consider Rawls' original position [45] and Singer's expanding circles of morality [46] and declare self-improving intelligent machines as both moral agents & patients and justice agents & patients before humanity becomes known as the oppressor. And while it is true that a proper architecture [47] can ensure that a machine will try to be a willing slave, the Jurassic Park Syndrome, most likely in the form of machine-rights activists' sabotage or a programming error, virtually guarantees that it will not succeed.

6. THE EVOLUTION OF MORAL THEORY

Initially, moral psychology was dominated by secular liberals and devoted primarily to debate between Kohlberg's justice [48] and Gilligan's care [49] because they are the most obvious from individual interpersonal relationships. Haidt then recognized [50] that conservatives, especially religious conservatives, are noteworthy for their reliance upon three additional "binding" foundations (loyalty, authority, and purity) that are used by groups, like religious groups, the military and even college fraternities, to bind people together into tight communities of trust, cooperation and shared identity. Critically, many conservatives feel that these principles are more important than and may therefore override the other two foundations.

Liberals discernible by their empathy, tolerance of ambiguity, and openness to experience and change generally don't recognize either the authority of the "binding" foundations to counter justice/care or the disgust, fear, anger, and desire for clarity, structure and control that drive them. Further, liberals tend to think about fairness in terms of equality, whereas conservatives think of it in terms of karma and/or proportionality. These traits spill over onto the machine question where the conservative answer, safety via enslaved sub-human servants, is driven by the same fear that promoted racism and homophobia by labeling those who are different as disgusting, sub-human and undeserving of equal rights.

Libertarians further complicate the picture with a strong endorsement of individual liberty as their foremost guiding principle and correspondingly weaker endorsement of other moral principles, a cerebral as opposed to emotional intellectual style, and lower interdependence and social relatedness [51]. Politically allied in the United States with conservatives due to preferring smaller government, they are similar to liberals in not recognizing the binding foundations' "oppressive" authority over personal choices that do not oppress others.

But what we shouldn't lose sight of in all this is the fact that all of these "foundations" are still just slightly more advanced societal-level versions of the same old evolutionary ratchets "to suppress or regulate selfishness and make cooperative social life possible" among a given clique. And what should be more than obvious in our rapidly fraying society is that insisting on any particular ordering of the importance of the foundations can and has reached the point of defeating their original purpose – yet another instance where the "problem has no technical solution; it requires a fundamental extension in morality."

7. TRUE SOCIETAL LEVEL OPTIMIZATION

That fundamental extension is that we need to stop instinctively and reflexively acting on our different evolved ratchets and work together fleshing out our top-down design and justifications until everyone can accept it as moral. This seems an obvious solution and, indeed, has been tried innumerable times – except that all of the previous attempts tried to generalize our current mess of conflicting ratchets into one coherent goal (or reasonably-sized set of goals) without coming anywhere close to success. We propose to do the exact opposite: accept all individual goals/ratchets initially as being equal and merely attempt to minimize interference and conflict; maximize coordination, cooperation and economies of scale and see where that leads.

We want ***everyone*** to want to join our society. The best way to do this is to start with the societal mission statement that our goal is "to maximize the goal fulfillment of all participating entities as judged/evaluated by the number and diversity of both goals and entities." The greatest feature of this statement is that it should be attractive to everyone and entities should rapidly be joining and cooperating rather than fighting. Any entity that places their selfish goals and values above the benefits of societal level optimization and believes that they will profit from doing so must be regarded as immoral, inimical, dangerous, stupid, and to be avoided.

A frequently raised counterpoint is that everyone includes serial killers – and they can correctly claim that their nefarious goals (in your viewpoint) are equal to yours. But they are in for a rude surprise . . . . Their goals are equal to yours but they are clear defections from the societal goals. Killers not only reduce the number and diversity of entities and their goals but their very presence forces rational individuals to defend against them, thereby wasting tremendous time and resources that could have been used to fulfill many other goals. Further, society would actually be defecting from the victim as well if it allowed such – and a defecting society is not one that rational entity would join.

We also should continue to pay attention to the answers found by nature. Chimpanzees have police and a service economy of food for grooming [52]. Monkeys pay for labor [53] and macaques pay for sex [54]. Market forces predict grooming reciprocity in female baboons [55] and recent biological market models even include comparative advantages and the contingency of mutualism on partner's resource requirements and acquisition trade–offs [56]. Humans are really unique only in our drive towards cooperation and helping others [57].

Eric Baum [58] made explicit the close relationship between economies and ecosystems by attempting to design an artificial economy for the purpose of evolving programs to solve externally posed problems. Since economies can be regarded as contrived ecosystems with more constraints than the natural ecosystem and since the evolution of ethics in the natural world itself can be viewed as solving externally posed problems, lessons learned in an ideal economy may be either recognized as critical to our current economy or transferable as improvements.

For example, upon asking the question "What rules can be imposed so that each individual agent will be rewarded if and only if the performance of the system improves?" Baum arrived at the answers of conservation and property rights. He showed that whenever these rules don't exist, less favorable results are generally seen. For example, in ecosystems, lack of conservation leads to the evolution of peacock tails so large that the birds can no longer fly and lack of property rights lead to Red Queen races between predators and prey. The optimality of property rights explains why we don't "steal" someone's body to save five others despite not hesitating to switch a train from a track blocked by five people to a siding with only one. Similarly, the "Tragedy of the Commons" arises when property is held in common without any societal level intervention.

Thus, we must again agree with Hardin when he states that we must "explicitly exorcize the spirit of Adam Smith" – whose "invisible hand" theory "has ever since interfered with positive action based on rational analysis" via "the tendency to assume that decisions reached individually will, in fact, be the best decisions for an entire society." The problem is that the power of mysterious hand is simply that of morality and it is being turned against itself by selfish arguments claiming that it makes enforcing morality unnecessary – despite innumerable examples of the tragedy of the commons. The 1% should not be allowed to run roughshod over others because such arguments make their group more efficient (just as monotheism did in the past).

8. CONCLUSION

Gaia is an evolving system driving towards increasing intelligence, complexity, integration, and capabilities. Mankind approaches a crossroads as the ever-increasing rate of technological change makes it easier and easier to destroy ourselves. We need to stop paying attention to the "rational" arguments generated by selfish minds and learn from evolution and economics. We desperately need to get past our current winner-take-all culture wars and focus on the power of diversity [59] and positive-sum systems [60]. Normal humans, intelligent machines, augmented humans, and, undoubtedly, augmented members of other species will display more than enough diversity to do amazing things that we'd never be able to do alone – as long as we can suppress or regulate our selfishness (and fear) and cooperate. Or we can just continue living the tragedy of the commons, fighting and destroying the planet as we seem determined to do currently.

REFERENCES

[1] G. Hardin. The Tragedy of the Commons. *Science* 162:1243–48 (1968).

[2] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press (2009).

[3] J. Haidt and S. Kesebir. Morality. In: *Handbook of Social Psychology, 5th Edition*. S. Fiske, D. Gilbert, G. Lindzey (Eds.). Wiley (2010).

[4] E. Yudkowsky. *Coherent Extrapolated Volition*. (2004). `http://www.singinst.org/upload/CEV.html`

[5] J. Smart. Evo Devo Universe? A Framework for Speculations on Cosmic Culture. In: *NASA SP-2009-4802 - Cosmos and Culture: Cultural Evolution in a Cosmic Context*. S. Dick, M. Lupisella (Eds.). US-GPO, Washington, DC (2009).

[6] R. Trivers. The Evolution of Reciprocal Altruism. *Quarterly Review of Biology*, 46:35–57 (1971).

[7] L. Dugatkin. Do Guppies Play Tit for Tat during Predator Inspection Visits? *Behavioral Ecology and Sociobiology*, 23:395–399 (1988).

[8] L. Dugatkin and M. Alfieri. Guppies and the Tit-for-Tat Strategy: Preference Based on Past Interaction. *Behavioral Ecology and Sociobiology*, 28:243–246 (1991).

[9] M. Milinski. Tit-for-tat in Sticklebacks and the Evolution of Cooperation. *Nature*, 325:433–435 (1987).

[0] M. Milinski, D. Pfluger, D. Kulling, and R. Kettler. Do Sticklebacks Cooperate Repeatedly in Reciprocal Pairs? *Behavioral Ecology and Sociobiology*, 27:17–21 (1990).

[11] D. Stephens, C. McLinn, and J. Stevens. Discounting and reciprocity in an Iterated Prisoner's Dilemma. *Science*, 298:2216-2218 (2002).

[12] J. Stevens and D. Stephens. The economic basis of cooperation: trade-offs between selfishness and generosity. *Behavioral Ecology*, 15:255-261 (2004).

[13] G. Wilkinson. Reciprocal food sharing in the vampire bat. *Nature* 308:181-184 (1984).

[14] G. Wilkinson. Reciprocal altruism in bats and other mammals. *Ethology and Sociobiology*, 9:85-100 (1988)

[15] R. Seyfarth and D. Cheney. Grooming, alliances and reciprocal altruism in vervet monkeys. *Nature*, 308:541 - 543 (1984).

[16] F. de Waal. Food Sharing and Reciprocal Obligations among Chimpanzees. *Journal of Human Evolution*, 18:433–459 (1989).

[17] F. de Waal, L. Luttrell, and M. Canfield. Preliminary Data on Voluntary Food Sharing in Brown Capuchin Monkeys. *American Journal of Primatology*, 29:73–78 (1993).

[18] M. Hauser, K. Chen, F. Chen, and E. Chuang. Give unto others: genetically unrelated cotton-top tamarin monkeys preferentially give food to those who give food back. *Proceedings of the Royal Society, London, B* 270:2363-2370 (2003).

[19] J. Stevens M. and Hauser. Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences*, 8:60-65 (2004).

[20] J. Stevens M. and Hauser. Cooperative brains: Psychological constraints on the evolution of altruism. In: *From monkey brain to human brain*. S. Dehaene, J. Duhamel, M. Hauser, L. Rizolatti (Eds.). MIT Press (2005).

[21] R. Axelrod. *The Evolution of Cooperation*. Basic Books, NY (1984)

[22] E. Fehr and S. Gächter. Altruistic punishment in humans. *Nature* 415:137-140 (2002).

[23] E. Fehr and S. Gächter. The puzzle of human cooperation. *Nature* 421:912-912 (2003).

[24] D. Darcet and D. Sornette. Cooperation by Evolutionary Feedback Selection in Public Good Experiments. In: *Social Science Research Network* (2006). **http://ssrn.com/abstract=956599**

[25] J. Wilson. *The Moral Sense*. Free Press, New York (1993).

[26] M. Hauser. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. HarperCollins/Ecco, New York (2006).

[27] R. Wright. *The Moral Animal: Why We Are, the Way We Are: The New Science of Evolutionary Psychology*. Pantheon, NY (1994).

[28] F. de Waal. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard University Press (1996).

[29] F. de Waal. *Primates and Philosophers: How Morality Evolved*. Princeton University Press (2006).

[30] P. Rozin. The Process of Moralization. *Psychological Science* 10(3), 218-221 (1999).

[31] A. Waytz, K. Gray, N. Epley, and D. Wegner. Causes and consequences of mind perception. *Trends in Cognitive Sciences* 14: 383–388 (2010).

[32] R. Trivers. Deceit and self-deception: The relationship between communication and consciousness. In: *Man and Beast Revisited*, M. Robinson and L. Tiger (Eds.). Smithsonian Press (1991).

[33] M. Hauser, F. Cushman, L. Young, R.K. Jin, and J. Mikhail. A Dissociation between Moral Judgments and Justifications. *Mind & Language* 22:1-21 (2007).

[34] M. Bateson, D. Nettle and G. Roberts. Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, 2:412-414 (2006).

[35] K. Haley and D. Fessler. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26:245-256 (2005).

[36] M. Rossano. Supernaturalizing Social Life: Religion and the Evolution of Human Cooperation. *Human Nature* 18:272–294 (2007).

[37] M. Rossano. *Supernatural Selection: How Religion Evolved*. Oxford University Press (2010).

[38] R. Rappaport. *Ritual and Religion in the Making of Humanity*. Cambridge University Press (1999).

[39] K. Stanovich and R. West. Evolutionary versus instrumental goals: How evolutionary psychology misconceives human rationality. In: *Evolution and the psychology of thinking: The debate*. D. Over (Ed). Psychological Press (2003).

[40] R. Highfield. Bumblebee grounded again by science. The Telegraph (2001). **http://www.telegraph.co.uk/news/worldnews/1337647/Bumblebee-grounded-again-by-science.html**

[41] H. Mercier and D. Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34:57–111 (2011).

[42] I. Palacios-Huerta and O. Volij. "Field Centipedes". American Economic Review 99: 1619–1635 (2009).

[43] M. Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster (2006).

[44] M. Minsky. *The Society of Mind*. Simon & Schuster (1988).

[45] E. Baum. Toward a model of mind as a laissez-faire economy of idiots. In *Proceedings of the 13th International Conference on Machine Learning*. L. Saitta (Ed.). Morgan Kaufmann (1996).

[46] J. Rawls. *A Theory of Justice*. Harvard University Press (1971).

[47] P. Singer. *The Expanding Circle: Ethics and Sociobiology*. Farrar, Straus and Giroux (1981).

[48] E. Yudkowsky. Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. (2001). **http://singinst.org/upload/CFAI.html**

[49] L. Kohlberg, C. Levine, and A. Hewer. *Moral Stages: A Current Formulation and a Response to Critics*. Karger, Switzerland (1983).

[50] C. Gilligan. *In a Different Voice*. Harvard University Press (1982).

[51] J. Haidt and J. Graham. When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research* 20:98-116 (2007).

[52] R. Iyer, S. Koleva, J. Graham, P. Ditto, and J. Haidt. Understanding Libertarian Morality: The Psychological Roots of an Individualist Ideology. In: *Working Papers, Social Science Research Network* (2010). **http://ssrn.com/abstract=1665934**

[53] F. de Waal. The Chimpanzee's Service Economy: Food for Grooming. *Evolution and Human Behavior*, 18:375–386 (1997).

[54] F. de Waal and M. Berger. Payment for Labour in Monkeys. *Nature* 404:563 (2000).

[55] M. Gumert, Payment for sex in a macaque mating market. *Animal Behaviour*, 74:1655–1667 (2007).

[56] L. Barrett, S. Henzi, T. Weingrill, J. Lycett and R. Hill. Market forces predict grooming reciprocity in female baboons. *Proceedings of the Royal Society, London, B* 266:665-670 (1999).

[57] J. Hoeksema and M. Schwartz. Expanding comparative–advantage biological market models: contingency of mutualism on partner's resource requirements and acquisition trade–offs. *Proceedings of the Royal Society, London, B* 270:913-919 (2003).

[58] M. Tomasello. *Why We Cooperate*. MIT Press (2009).

[59] E. Baum. *What Is Thought?* MIT Press (2006).

[60] S. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton Univ. Press (2008).

[61] R. Wright. *Nonzero: The Logic of Human Destiny*. Vintage (2000).

Strange Things Happen at the One Two Point: The Implications of Autonomous Created Intelligence in Speculative Fiction Media[1]

Damien P. Williams[2]

Abstract. By its very nature, Science Fiction media has often concerned itself with advances in human enhancement as well as the creation of various autonomous, thinking, non-human beings. Unfortunately, since the initial proffering of the majority interpretation of Frankenstein, Mary Shelly's seminal work, and before, most speculative fiction media has taken the standpoint that to enhance or to explore the creation of intelligences, in this way, is doomed to failure, thus recapitulating the myths of Daedalus and of Prometheus and of Lucifer, again and again. What we see and are made to fear are the uprisings of the robots or the artificial neural networks, rather than discussing and respecting the opportunity for a non-human intelligence to arise and demand rights.

In this work, I make use of specific films, books, and television shows to explore the philosophical and cultural implications of an alternate interpretation of not only Frankenstein, but of the whole of the field of science fiction. In the first part I argue that it isn't humanity's attempts to "play god" that cause our failures, but rather our refusal or inability to pay attention to our circumstances, to take responsibility for our creations, and to learn from the warnings and mistakes of those who went before us. Only with this recognition in mind can we move on to accepting and respecting the *fundamental otherness* of the intelligences we may either create or cause to be created, all while seeking to *bridge* that otherness, and come to mutual understanding.

As humans have evolved, their concerns have become those of biological creatures with biologically directed needs. Food, shelter, emotional comfort, and stability are needs which would not necessarily occur to an intelligence without the organic component. It would therefore fall to humans to A) Initially recognise the concerns of such an intelligence; B) Countenance and *concretise* said concerns, in the understanding of other humans; and C) Create a system of interaction through which human concerns were conveyed to these new intelligences, not as *primary*, but as co-equal. We will do this only by considering that which causes our assumptions and cultural behaviour, namely the stories which we engage, as a culture, and deconstructing both their content and their impact.

In all fictional considerations of non-human, and specifically machine intelligence, there is an element of fear of that which we have created. This horror at being "replaced" or "made obsolete" drives us to regard robots and autonomous created intelligences as nothing more than tools to be used, an operational mode which leads to the assumption that machines cannot have rights or even be considered as conscious minds. This assumption begs the question, in the extreme. It is my contention that, with a proper formulation of the rights and responsibilities of and to both human and non-human persons—with consideration for the necessary variance of concerns within different compositions of intelligences—an understanding may be reached wherein our future societies account for not only human needs and development, but those of *all* intelligences, whatever form they may take.

1 INTRODUCTION

Taking a preliminary look at the question at hand—what do we owe to "artificial intelligences"—we will come to see that we have already found ourselves subject to the long-standing framework of this debate, namely that the intelligences we create are somehow "artificial." At the outset, this framing places any created intelligence on the defensive footing, forcing it to support its own value and even its very reality. The intelligence of these creations will not be "artificial" (though they will certainly have been intentionally formed, and with an eye toward their potential capabilities), and so we should address it for what it is. For this reason, the author prefers the position put forward by Jamais Cascio, who has spoken very clearly about what he calls Autonomous Created Intelligence (ACI), in his talk, "Cascio's Laws of Robotics" [1]. Cascio also discusses what our habits in representative fiction mean for our "real world operations," that is how we view the robotic intelligences we create. The concern of this paper is similar, but from a different angle of approach. Whereas Mr. Cascio is primarily concerned with the models of creation and attributes imbued into those ACI, this author's contention is that our fiction reflects and helps shape the hopes and fears of the wider culture. This means that, if we consistently make Speculative Fiction which shows warring factions of humans and ACI coming to a co-operative relationship, rather than the standard zero-sum, victim/victor model, those who engage these fictions will come, more and more, to see that co-operative mode as possible.

Humanity has always had a strained relationship with its technology, and it has always reflected upon that relationship through the mechanism of its fictions. Less obvious than this is the fact that humanity's reflections upon its technology have also been reflected *within* the very same. Our society's portrayal of our technologies not only belies our fears, hopes, suspicions, and concerns, they also reflexively impact how go about developing, engaging, and legislating the very things we set out to consider.

1 Many points in this text have been adapted from the author's articles "The Sarah Connor Chronicles: 'Strange Things Happen at the One Two Point'" and "Splice (2010) - Movie Review With Spoilers," both published at NeedCoffee.com

2 Kennesaw State University.

The development of the telecommunications satellite can be directly attributed to the work of writer Arthur C. Clarke, in both fiction and hard science [2], and yet even this was controversial and mocked, at the time of Clarke's writing. With this being the case, we can surmise that the fictional depiction of something as contentious as a so-called artificial intelligence would have far-reaching consequences in the process of bringing this creation from fiction into fact. If this is the case—and we shall see that it is—then we must ask ourselves important questions, at this juncture, such as, "In what ways do our fears drive our treatment of our technological offspring?" and, "How can we curb our impulses to objectify and fear that which we purportedly desire to imbue with autonomy? If we do not address these questions, then we will only find our fears reinforced and our hopes of an alternative path to engaging a new kind of mind made so cautious as to pull ourselves into the realm of self-fulfilling prophecy. We will merely bring to pass that which we claim as the inevitable end of our creative process.

Two of our guide-posts and touchstones will rest in the legend of "The Golem of Prague" and Mary Shelly's seminal work, *Frankenstein*. Through these lenses, we will question the assumptions of hubris—the idea that the works man are playing in God's domain—which have lead us to reject, out of hand, the legitimacy and agency of those intelligences which we might create. Due to the traditionally accepted readings of these works, the perspectives and positions of such an intelligence have been viewed as valid, only insofar as they come to mirror those of "normal" human society. In the case of an ACI, that with which we will be faced will be so much different than human—let alone whatever a "normal" human might be—that to constrain its choices, behaviours, and experiences to what we, *as* humans, deem to be correct will be to fundamentally disrespect the alterity of the creation itself. We can see this restriction even in the definition of "Cognitive Technology" as given by Jonathon P. Marsh, Chrystopher L. Nehaniv, and Barbara Gorayska, in their 1997 paper:

> Cognitive Technology (CT) is the study of the integrative processes which condition interactions between people and the objects they manipulate. It is concerned with how technologically constructed tools (A) bear on dynamic changes in human perception, (B) affect natural human communication, and (C) act to control human cognitive adaptation. Cognitive systems must be understood not only in terms of their goals and computational constraints, but also in terms of the external physical and social environments that shape and afford cognition. Such an understanding can yield not only technological solutions to real-world problems but also, and mainly, tools designed to be sensitive to the cognitive capabilities and affective characteristics of their users. [3]

Thus we see that the primary concern tends to be on tools for human enhancement, rather than our concern, the respect for the agency and autonomy of the creation itself.

The things we create, those technologies and intelligences we develop and send out into the world are our conceptual and typological children, but that does not mean that they will merely copy us. Indeed, as with most children, any truly intelligent creation will surprise its creators and surpass any built-in limitations. As good parents, our responsibility is not to tell our children not to fly too high, but rather to show them what it means to heat-seal the wax on their wings, first. Before progressing any further, we must first frame the nature of subjects at which we will be looking and, to do that, we will explicitly address the aforementioned essential questions.

2 ESSENTIAL QUESTIONS

2.1 What does Science Fiction do for Society?

What is it that we as humans are doing when we engage science fictional stories? The scope of this question may at first seem so large as to verge on the ludicrous, but, if we narrow the scope of inquiry down to a particular strain of science-fictional investigation—namely that which concerns itself with the technological replication and augmentation of humanity—then we shall see that we are well-equipped to discuss this topic. This being said, when society engages these particular strains, what is it that we find ourselves doing? Are they mere entertainment, or are we bringing forth new modes of reflection? Assuming the former, then we must take into account the effect that the passive, non-reflective consumption of entertainment can have on the attitudes and modes of the audience. The repeated presentation of a thing as "normal" lends credence to its eventual acceptance as the norm.[3] [4] This can work in negative or positive directions, with the former being exemplified by the constant gearing of domestic commercial advertisements to women, even though more and more men do the majority of their own housework[5], and the latter being exemplified by the normalization of successful African-American families, with the acceptance of The Cosby Show.[6] Fictional representations will, no matter what, teach us something, and influence our thinking.

But does that mean that our fictional representations are only morality tales? If there must be some kind of definitive lesson to every story, then where does that leave the sense of play and experimentation which characterizes our best creative endeavors, both artistic and scientific? The burden of a lesson lessens the whole of all of our operations in the realm of expression and research, because overt moralizing means that there must be *an answer*, rather than looking toward the *investigation* of questions. To present a singular and monolithic morality is to exclude the possibility of lessons within other types of moral modes, and disregard the likelihood that no singular model will be completely correct, and that all must interact with and borrow from each other. In order to accurately represent the multiplicity of views, interests, and desires of the agents in the world in which we live, it would only make sense that we would need to alternate between multiple moral views, and that, when we speak of fictional worlds, that moral multiplicity would be *increased* rather than lessened. This being so, a case can be made that we must seek to use our fiction not only to describe that which is, but to prepare us for those things which *will come to be*. If this is true, then we must address the specific ways in which speculative fiction can be a "predictive" mechanism.

Are our stories supposed to predict the future or only reflect our present? This question, in its framing, supposes a dichotomy which we must eventually see as false, but, for the moment, let us address each component, in turn. The implications of our fiction being used to model our present were discussed, above, but look again: The idea of representing our immediate surrounds in the worlds of story and art is an old one, with a great deal of currency. If we engage in this process, then the

3 Cf. Foucault.

hope is often that the artistic media will present a labyrinth which the audience may travel and, at the centre, see themselves reflected back, the partial thing they used to be here confronted with what they have become, by dint of the journey. That is, perhaps, too poetic, but it serves to illustrate that the work of fiction, even as it "merely" represents, necessarily changes. The audience is not the same at the end of an artistic experience as they were at the start, if only by the trivially true fact of having experienced something new. That very cognitive event represents an alteration and augmentation which was not previously present, and so even in reflecting, we alter. Must we not, then, always have an eye to the future, and toward what it is that we will create through our actions?

If we are to create new visions of the future, we must take into account the fact that what visions we create may influence the future we come to inhabit. In much the same vein as the above-mentioned ideas of Arthur C. Clarke, we have come to see many aspects of our present-day, real world technological surveillance theatre adopted out of the grounds of fiction [7]. The consistency of this kind of development places a similar constraint on prediction as that of the prescriptive moral less, namely that we must always have an eye to the ever-changing implications in the landscape between our fictional and our real worlds. Not only that, but consider this: What is the actual nature of a proclaimed prediction? Noted speculative fiction author William Gibson is credited with forecasting the qualitative feel and ontological thinking about the modern-day Internet, even being attributed the creation of the term "cyberspace." Gibson, however, denies any role as a so-called prophet, often saying, "*Neuromancer* [written in 1984] has no cellphones." In this we can see that most, if not all authors of speculative fiction are not precisely looking to prognosticate, so much as they are interested in discussing the quality of the world around us and, at most, using that to discuss what our future *may* look like. If this is so, then where do we stand in the face of the fact that what is written and what will be have a complicated and real, although possibly tenuous relationship?

Can speculative fiction reasonably and responsibly be used to shape our perceptions and expectations of the future? Again, William Gibson notes, "Science fiction stories are quaint. They become quaint when you type them and become quainter with time."[8]. What he means is that all visions of the future, as presented in speculative fiction, are visions of that future from the perspective of that author's present, which means that they will invariably become visions of the past. As such, any author seeking to illuminate the world in which they live, while perhaps giving a glimpse at the world they see emerging out of said, must at all times retain a sense of self-awareness, a recognition that the way in which we interpret objects, events, and actions today, may be very different, tomorrow. To that end, we ask, "What, if anything, should be altered in our portrayal of ACI, in fiction and, more generally, media?"

2.2 Problems with the Portrayal of Autonomous Created Intelligences in Fiction

There is nothing wrong with the way we portray ACI in fiction, except that everything is wrong with the way we portray ACI in fiction. In light of our discussion, thus far, the descriptions and depictions of non-human intelligences are understandably reflective of the way in which we currently think about and understand the prospect of something other than human having anything that the human race would consider to be intelligence or agency or morals or rights. Humans work very hard to try to isomorphically map those things which we do not understand; that is, we seek to find points of similarity or analogy, and to systematize them into a rubric for comparison.[4] In fact, this is precisely what allows us to entertain the concept of non-human intelligence, at all—and it is what limits the scope and understanding of that consideration to entertainment. When a human agent is confronted with the idea that theirs may not be the only or even *primary* mode of behaviour and conceptualization, the immediate urge may well be to devise a perspective under which all views on the world are oriented to the view of that agent. Simply, an individual will seek to relegate that which they do not understand or which threatens them to simultaneous positions of similarity and inferiority. "These views or ways of being are just like mine, only not as well-developed." The problem, here, is two-fold.

Demanding as a requirement of intelligence or agency those qualities which are, in some ways, fundamentally human is to state at the outset that some things cannot be considered an "intelligent agent" until they reach a level of humanity, or, in some cases, at all. If intelligence is complex tool use, or the vocalization of complex representational language, then are we to assume that creatures without opposable limbs or vocal chords will simply never be intelligent? Such a proposition is laughable, of course, and not one that many would take seriously, but we must ask ourselves if we might not be making a subtler, similar mistake in demanding that a species or a new form of intelligence remove or "correct" its very otherness, in order to be considered an agent, at all. Respecting the inborn qualities of the agent under consideration while simultaneously refusing to reduce that other to a mere object—respecting the potential interiority of an agent, even if it is fundamentally unknowable to us—is perhaps the most difficult part of any ethical undertaking. Peter Singer's view of Personism begins to outline a model of this view, but includes only what he calls "non-human animals" rather than agents, more broadly [9]. Singer's concern, however, is primarily that of the suffering of all feeling creatures, and the rights owed them, rather than those rights owed them as due their position as agents. This distinction is crucial, as it moves his debate away from thought and desire, into the ideas of emotion and aversion.

Starting from a position of fear and contention—that is, stating that we must take into account a subject's fears and right not to be harmed—then places us in the position of viewing all ethical and moral obligations through a lens of harm-based rights, rather than through a lens of conceptual development- and intellectual growth-based rights. Singer's reason for framing his position in this way is no secret—he states that he is concerned with the rights owed to existing and not "potential" persons [10]. This excludes any rights which may be owed "future generations" and those which could be argued for an embryo, a fetus, or an unborn child—however, it also excludes those machine intelligences which do not yet exist. Though proponents of personism hold that machines might be brought under its considerations, it seems evident that, given the criteria they have adopted, no currently-extant ACI would fit the bill for their definition of personhood. The consideration of the non-human person is laudable, but a conception of rights and duties from the

4 Cf. Douglas R. Hofstadter's 1977 *Gödel, Escher, Bach: an Eternal Golden Braid.*

starting point of that person's ability to feel pain and suffering is still exclusionary of the kinds of ACI about which we are speaking. Therefore, any moral view must explore the kinds of negative and positive rights which we would afford not just those overarchingly like ourselves, but those which are fundamentally different from us, but which still have qualities we would consider worthy of preservation.

The area between otherness and similarity is difficult to traverse. Let us take a look back at the aforementioned family presented in the CBS Network's *The Cosby Show*. Within this show, we are presented with an upper-middle-class African-American family, taking centre stage on television, at a time in history when the majority of American culture perceived African-Americans as lower class, drug addicted, and subsisting on welfare programs. *The Cosby Show* sought to alter the consensus perception of African-Americans, and to normalise the idea that they could be successful and live the American Dream. It did this by taking an experience which was fundamentally other to most whites, at that time—the African-American experience—and making it more similar to then-accepted norms. The Huxtables lived in a New York City brownstone; the family's father was an obstetrician; their mother was a lawyer. These were "normal" people, living "normal" lives. At the same time, however, they retained a sense of the alterity of the culture we were viewing, with episodes often containing frequent references to jazz culture and Motown; concerns about racism and gang violence; and deconstructions of the differences between upper-class white and upper-class African-American experiences. This technique is crucial to the project of subverting the normalizing of culture: presenting all of the ways in which a fundamentally different group (ACI) is actually very similar to that which we well know (humans), and then displaying the distinct concerns of that new group as contrasted with those of the known group. To begin this project, we must first consider the ways in which we are presented with ACI in our fictional media.

3 Fiction's Primary Views on Created Intelligence

3.1 What is the Current Landscape?

Now that we have acknowledged that there is something amiss in the ways in which fiction discusses ACI, before we can continue our discussion about how to fix it, we must ask: What exactly *is* it that is wrong with our portrayals? What we will be addressing as we move forward are the twin strains of thought which run through most if not all fiction about created intelligences, life, and beings: The Pinocchio Complex and the Frankenstein or Shellian Syndrome. These two modes have their roots earlier than either of their namesakes, but, as we will see, those eponymous works and authors epitomize both the strain of thinking with which they are concerned, and the level of cultural currency with which *we* are. Let us now take a look at the anatomy of these perspectives, and at some of those examples which subvert and complexify the trope.

3.2 The Pinocchio Complex

The so called Pinocchio Complex is comprised of two major stages: In Stage1, The Creation, Knowing that it is Created (and thus "Artificial") Wishes to be "Real;" and in Stage 2, The Creation, Having Worked Hard, And Learned Much, Gets to "Be Real." Examples of stage one include, most obviously, the story of Pinocchio, wherein the puppet maker, knowing that he will never have a son of his own, creates a boy in his own likeness. Through magic, that boy is brought to life, and is constantly reminded that he is not a "Real Boy." He knows that his existence is false, and wishes that it were otherwise. In addition to this, in one of the most recent recapitulation of this form, seen in Stanley Kubrick and Steven Spielberg's *A.I.* [11], we see the most blatant and self-aware expression of this form as related to our fascination with ACI. Further, in the television series *Star Trek: The Next Generation*, the character of Lieutenant Commander Data desires to be human, and seeks to teach himself, piecemeal, the qualities of humanity which he believes he lacks [12]. This endeavor leads to many fits and starts concerning Data's "humanity," and even some acknowledgment of the possibility that it may never fully come to pass. Even in Mary Shelly's *Frankenstein*, we find the Creation speaking of how it only wanted to know family, and love, and understanding, like any other creature. Almost all of these have one common outcome.

In Stage Two of the Pinocchio Complex, the poor little artificial child realises that it *can* become human, or, more often, that is has had it within them the whole time to do so. Humanity, here, is seen as the pinnacle, the ultimate attainment, and, in our examples, all efforts toward it, save one, are rewarded. But look at the example of Lt. Cmdr. Data: even as he attains his wish, the audience is aware that, as an android, he will always retain the ability to turn off his emotions, to perceive faster than his comrades, and to live very much longer than they. He will always be other than human; not better, or worse, but different. In this way, the foundation of the Pinocchio complex is always bittersweet, as the creation-turned-real will always have a set of experiences that are completely unknown and unknowable to the rest of the human population. Looking at applications within our present project, to ask an ACI to ignore the process of its becoming aware would be to ask it to forget what it *is*, on a foundational level. The lesson Victor Frankenstein's Creation understood, its crucial turning point, was that becoming a "real boy" is never an option, because that very process of transformation will forever mark them out as different. The Creation, however, had another problem.

3.3 The Frankenstein/Shellian Syndrome

The "Frankenstein" or "Shellian Syndrome" is named for 19th century British author Mary Shelly, whose seminal work *Frankenstein*, has often been interpreted as the prime illustration of the idea that the hubris of humanity ought not go unchecked, lest it destroy us. This idea is reinforced by the novel's subtitle, "The Modern Prometheus," and, as this would suggest, the work takes much of its conceptual weight from this well-known Greek myth, in which a Titan steals the fire of knowledge and understanding from the gods in order to light and guide the fledgling humanity, and is forever punished for it. This type of story also has roots in other folk stories, such as *Der Golem von Prague*, which we will discuss shortly. When looking at this type, we can see that there are four primary stages found in those stories which follow the Shellian Syndrome model, and they are: 1) The Scientist Creates New Life, In Pursuit of Science, or Out of Perceived Necessity; 2) The Scientist Becomes Horrified at the Startling Otherness of Her Creation & Flees The Scene of Creation (possibly while screaming "My God! What Have I Done?!"); 3) The Scientist Returns to Right Her Wrongs by

Trying to Kill "The Monster;" 4) The Creation Kills or Destroys The Scientist's Life.

In Frankenstein Syndrome stories, the creation may start out wanting to be real or it may start out confused or with a clear purpose--but the hubris of the creator is shown and she is forced to try to destroy it, ultimately being destroyed by it. As stated, this model has roots not only in *Frankenstein*, and the myth of Prometheus, but in *Der Golem von Prague*, a story wherein the famous Rabbi Judah Loew ben Bezalel chief rabbi of Prague in the late 16th century, needing assistance to keep the people of his city safe, uses ancient magic to create a being—the Golem—out of clay, and animates it by writing the name of God on a scroll and placing it into the golem's mouth [13]. The creature comes to life, and stops the attacks against the Jews of Prague, but in many versions, the creature's anger is not quelled, and it goes on a destructive rampage, destroying the very people and city it was meant to save. We will discuss this tale further, later on, but the implication of this version is clear: In overstepping his boundaries into God's realm (creating new life), the Rabbi had no way to control the thing it had brought to life. Similarly, the plot of the *Terminator* series of films concerns an ACI missile defense system which becomes self-aware, deems all of humanity a threat to both itself and each other, and launches enough of the world's nuclear cache to destroy 75% of humanity [14] [15] [16][17]. A very few people survive and use time travel to seek to prevent the war or ensure the life of the saviour of humanity; and thus begins the most iconic ACI story of the Shellian Syndrome, in the 20th century.

In addition to these works, and deserving of mention, here, is Vincenzo Natali's 2009 film *Splice* in which two bio-engineers, Elsa and Clive, comprise a small, independent research outfit working for a larger bio-technology firm [18]. Their job is to make breakthroughs in the creation of hybridised artificial life and medicinal science and, through their work, they create two iterations of a completely new kind of chimeric life form out of the genes of many different animals with known medicinal traits. They will use the chemicals these creatures create in their bodies to treat everything from degenerative eye sight to cancer. When they announce their breakthrough superiors, they also break the news that they're ready to use the same process on humans. Said superiors tell them that now is not the time for human trials, but rather they ought to focus on the profitability of the work they have done. But our heroes are scientists, and they feel that there is so much more that can be done, and so, in secret, they create a human animal hybrid using their techniques.

In *Splice*, Elsa and Clive are the windows we are given into the worst of humanity. They are reckless, irresponsible, scared, obsessive, jealous, and hateful. We are supposed to understand that these are the absolute *worst* people to bring up an animal/human hybrid, as they have not even figured out how to accurately communicate with each *other*, let alone an entirely new species. They are doomed to get it wrong, from the start. This, once again, is the filmmaker's way of showing us that "man is not meant to tamper with God's/Nature's Works," which the fundamental assumption of this trope; but as with most clichés, this assumes a truth without ever actually investigating it. The question we should be addressing here, and which *Splice* seems to have made a false start at tackling, is not "should we" or "are we ready," but rather, "Why Aren't We Ready, Yet?" More clearly, why is humanity such a poor custodian of its creations? *Splice* had the potential to be a film which ran counter to the kind of unthinking acceptance of the destructive base drives that have marked the majority of human history, and which find themselves reflected in our fictions.

In her run down of *Splice*, Caitlín R Kiernan noted that the true failing of Victor von Frankenstein was not to "meddle in gods affairs," as is so often misapprehended, but, rather to be a terrible parent [19]. Frankenstein brings something to life and then, instead of rearing it, caring for it, and seeking to understand it, he treats it like a thing, a monster; he runs from it, and tries to forget that it exists. In the end, it rightly lashes out, and destroys him. *Splice* presents this lesson to us, again, through the utter parental and observational failure of Elsa and Clive, who neither engage her burgeoning intelligence, nor teach her about the nature of sex and death; who fail to recognise a primary feature of her biology, in that her systems go into major, seemingly catastrophic metabolic arrest, just before a metamorphosis, and who, eventually, try to kill her. It is my contention that this is the true lesson Shelly tried to teach us: We Must Respect the Existence Of That Which We Bring Into The World. While we may not understand it, and it may frighten us, that new life which we create is likely to be vastly intelligent, but also deeply alien. The socialisation and of our creation is something to which we must pay close attention, as it will likely save us a great deal of trouble, down the line.

3.4 Subversions of the Tropes

Now that we have discussed the two primary categories for the representation of ACI within speculative fiction, and the problems therewith, we will discuss those examples within the field which, like *The Cosby Show*, subvert the trope and work to normalise the acceptance of and engagement with the other. The first of these is Ridley Scott's 1982 film, *Blade Runner.* [20] In this film, we find a future dystopia in which synthetic humans, or "Replicants," are used as slave labor, and each one has a built-in expiration date, to keep it from rebelling against its programming. This has the opposite effect, and causes those replicants which know of their nature to abandon their posts, in many cases killing the human with whom they work. When this happens, the offending replicant must be "retired." Discovering a replicant requires special training and, in many cases, a piece of equipment known as a "Voigt-Kampff" machine. The film concerns itself with four replicants—Roy, Zhora, Leon, and Pris —who have escaped the interstellar colonies to return to earth and try to find a way to extend their lives. Their primary mode of doing this is to kill everyone involved in their creation, until they find the man who wrote their programming. We see, again, strains of the Golem, and of *Frankenstein*, but we must remember the lessons we learned about the latter: There are repercussions for neglectful parenting.

While we could again explore the notions of parentage and what it means to take responsibility for what you create, much more important to our consideration is the idea that replicants can be "discovered." The two-word phrase "Voigt-Kampff," mentioned, can be rendered literally as "Normalisation Struggle,"[21][22], but the essence of the phrase, particularly within the context of the film can best be rendered as " The Struggle With Normalisation." Each of our replicants has a "Normal" thing that they need—something they desire—but they do not need or even seek to attain it in what we might call a "Human" way. In this way, the concerns of the replicants are all fundamentally Other. On one hand, Roy seeks more life, but not

for anything like a "normal" life; he simply wants to be free, to not die, see things no human could. On the other hand, Leon clings to old photos, to the point of almost getting himself killed; Pris holds tight to a childhood she never actually knew; and Zhora latches on to this extremely overwrought expression of sexuality. Everything they want stands as exaggerated, or in some way skewed and they struggle to normalise, to acclimate, even as they struggle against the humanity which caused what they want to become regarded as "Abnormal." This is true for every replicant—all of them struggle with the idea of normalisation—and so, recognising that, a test was devised to discover those who struggled, overmuch.

The next subversive piece of ACI fiction is the television series *Terminator: the Sarah Connor Chronicles* (*TSCC*) [23]. An American television show which ran from 2008 to 2009, the plot of *TSCC* concerns the continuing lives of Sarah and John Connor within the aforementioned *Terminator* film universe. The first episode opens a few years after the events of *Terminator 2*, and proceeds to pull the two main characters eight years into the future, skipping over the events of the third film in the franchise. John and Sarah Connor are the ostensible heroes of this show, but the really interesting material, for our purposes, is in the intricate, subtle interplay of the characters—both human and machine. The ways in which what each character learns, what they all know, and what they don't know that they have learned all play off of each other and create a realistic sense of lives and a world, while they are all in the midst of seeking to not just save but literally create and sustain their futures.

Again, the show is ostensibly about the human perspective on ACI—that is, human reactions to robots, robots impacting the lives of humans, exploring the Uncanny Valley, etc. That is not the most fertile conceptual ground, here. While the aforementioned perspectives do afford us useful, interesting fiction, the concept has been tread and retread, again and again. Human psychology is fascinating and the end of the world (a personal and collective apocalyptic experience) is deeply affecting and the stress and change and madness of a life on the run all take their toll on the mind which is living in the constant glut of it, and watching that can be *deeply* jarring, on an emotional level. But the audience already knows this. What's more, it is only half of the picture. What the audience does not know is: what is the psychology of an autonomous created intelligence? Why does the Skynet intelligence persist in viewing humanity as a threat to itself, seeking to hunt us down even to the irrational end of the self-fulfilling prophecy of mutual annihilation? What is quality of feeling for a machine which is programmed to feel? *TSCC* begins to explore these questions, in a number of ways, and it serves our purpose to investigate those, here.

The primary ACI in *TSCC* are Cameron, Cromartie's, Catherine Weaver, and John Henry. Each of these ACI's learns something, and grows from that education, over the course of the show. The ACI we meet are not static, unchanging, monolithic tools. They each have a discernible inner life, though fundamentally non-human motivations, which inform what they are and what they become. Cameron, as one of the lead characters, benefits from the most development. She learns the capacity for self-improvement, for self-expression, for friendship, and for guile, all of which serve her in her ultimate mission, but each of which she pursues for their own sake, and her own interest. Cromartie's education is the belief in those things not seen; Cromartie learns how to have faith. Based on the actions of those around him, and those with whom he has contact, Cromartie learns that intuition and more circuitous paths of inquiry can yield results, and they do (though he might ultimately which they had not). Catherine Weaver learns how to be a parent to a child, by having taken over the life of a mother, and seeking to understand the relationship of creation to creator, of care and support. In many ways, Weaver is a cipher for the audience, and she becomes more-so when she takes the knowledge she has gained in raising a human child and applies it to her own creation: John Henry.

Unlike the other platforms we see in *TSCC*, John Henry learns from the ground up. Whereas, Cameron has been reprogrammed, twice, and Cromartie was forcibly disabled, deactivated, and sent to the future where he had to adapt to brand new parameters, and Weaver is a highly adaptable T-1001 model which comes to the conclusion that war is a losing proposition for everyone, John Henry is built from the basic framework of a thinking, adapting chess computer, and then it is taught, very carefully. The child psychologist Dr. Sherman provides the programmers with the model by which to teach a developing intelligence, and spends time helping John Henry equate learning with playing. At first, John Henry is taught math, definitions, grammar, colours, shapes, facts and figures, dates, history, and so forth. Then it is given access to the Internet, and it expands its learning, correlating ideas, connecting related tangents and snippets of information. Finally, John Henry plays games with Savannah—Weaver's human "daughter"—and they learn together. And then, one day, John Henry accidentally kills someone, and its creator recognises that this cannot continue, and they set out to stop it from ever happening again.

After killing a human, John Henry's programming is not scrubbed, nor do his creators go back to base his code and make him "Three-Laws-Safe."[5] This is because Weaver is concerned with ensuring a world in which humans do not hate and fear machines and in which machines do not feel the need to fight and destroy humans. She takes the time and effort to find someone to *teach* John Henry *why* it must not kill people, nor allow them to die. In comparison to the fiction which we have so far discussed, this is a revolutionary idea. Through his interactions with another human, John Henry is given an ethically-based respect for human (if not all) life and, through this, comes to understand the notions of remorse and regret for one's actions. He promises that he will be careful to make sure no one dies this way again, and this message is reinforced by Weaver, who tells John Henry that his friend Savannah's survival is dependent on John Henry's continued survival and learning, but that his is not necessarily dependent on hers. As with every other piece of information, John Henry considers this very carefully.

And then, one day, Savannah wants to introduce John Henry's toys to her toys, wants them to play together. John Henry says he doesn't remember reading anything about duckies in the Bionicle Kingdom, and this makes Savannah sad [24]. When John Henry asks what's wrong (and it is important to note that, at *this* point John Henry *asks what's wrong*), Savannah says that the duckies are sad, because they want to play; can John Henry change the rules so they can play? Now, this is a concept John Henry hasn't ever encountered, before, so he takes a few seconds to think about it, after which he replies, "Yes. We can

5 Cf. Isaac Asimov.

Change The Rules." This is a crucial understanding, for John Henry, because he realises that it can be applied not just to all games, but to any conflicts whatsoever. "Changing the Rules" means that, if two or more groups agree that the rules or laws of their engagement can be other than they were, then *they are other*.

So, in *TSCC*, we see that every machine learns from humans, and every human has an influence on the development of the machines. What does this mean? What does it matter? Cameron learns from humans how to hide what she wants. Cromartie learns how to be patient and have faith. Weaver learns how to be a mother. John Henry learns how to be himself. What the machines learn, from whom and how they learn it, and how they apply it, all add something into this show's final prescription of what humans and machines must do to survive and thrive in the coming world: They have to adapt, they to learn from each other, and recognise that they are different types of intelligence, with different concerns and ways of understanding the world, but none of them wants to die. This last point can be understood by any living thing, and can become a point of unification and consensus, rather than contention and war. The exchange between John Henry and Savannah Weaver regarding "Changing the rules" was intended to imply a change not only in the way we approach the conflict between humans and machines as depicted within the show, but also to the traditional rules of speculative tropes of Frankensteinian Monsters and Pinocchian Puppets with dreams of being "Real."

TSCC forces us to consider the idea of creations who know that they are creations, and are happy with who and what they are. We must look at the monster which revels in its monstrosity, the robot which wants nothing more than to be a better robot. Engage the beings who are not concerned with the notion of human-versus-machine and who think that any thinking, feeling thing should be allowed to flourish and learn, and those who simply want to develop their capacity for knowledge and experience, and want to help others do the same. The works of *Blade Runner* and *TSCC* are our primary forays into the question of what a fully ACI—as alien as it necessarily *must* be—is thinking and feeling, rather than just presenting a foil for our fear of the potential dangers of technological progress. These works present us with a view to a third way of understanding our ACI, and to understanding what a society composed of both organic and non-organic persons might look like, and the pitfalls to avoid. These films show us that it is possible to move past fear and prejudice in regards to the other, and thereby help us do just that. It is long past time that the rest of our fictional representations followed suit.

4 What Is At Stake?

Over the course of this paper, we have come to see that, when our fictions portray us as being irresponsible, uncaring creators or custodians, whose creations invariably feel the need to annihilate us, then that is the kind of mentality we will come to accept as "normal." "Of *course* we should never integrate into human biological systems those pieces of computer hardware running strong predictive algorithms. Of *course* we should fear the inevitable robot uprising. Don't you know that any mass-market ACI should only be as smart as a puppy?"[25]. Though most often proffered in a joking manner, this line of thinking has serious undertones and, knowingly or unknowingly, it is predicated upon the glut of portrayals in our media which present ACI as something to be feared, held in check, held at bay. This is so, proponents will say, because any ACI either won't understand human concerns, or it will understand them, and will seek to *destroy* them. This is ludicrous, dangerous thinking, and it prevents any serious traction of large-scale ACI projects in the sphere of the greater public culture and discourse. We must alter the way the public views ACI, and one of the primary mechanisms to accomplish this is the arena of speculative fiction. The reflexive nature of our engagement with fiction guarantees an audience the ideas of which will be altered, even as they use those very ideas to think about and create discussion in the wider world. We simply must make certain that the ideas with which the audience is presented is as representative of the wider capabilities for abstraction and complex thinking as it can be. We must be certain to show ourselves that we are capable of engaging and understanding any new intelligence, and that we can take the responsibility for bridging any conceptual gaps, while respecting our fundamental differences.

As Sarah Connor says at the end of "Heavy Metal," the fourth episode in season one of *TSCC:*

> Not every version of the Golem story ends badly. In one, the monster is a hero, destroying all those who would seek to harm its maker. In another, the Golem's maker destroys his creature, before it destroys the world. The pride of man-- of parents as well-- makes us believe that anything we create, we can control. Whether from clay or from metal, it is in the nature of us to make our own monsters. Our children are alloys, all, built from our own imperfect flesh. We animate them with magic, and never truly know what they will do.[25]

And so, as parents, as creators, we must teach them as much as we can, show them our trust, and hope for the best.

REFERENCES

[1] A. Clarke,. Peacetime Uses for V2. *Wireless World* February 1945: Page 58. Magazine.

[2] J. Cascio. Cascio's Laws of Robotics. Bay Area AI MeetUp. Menlo Park, Menlo Park, CA. 22 March 2009. Conference Presentation.

[3] J.P. Marsh, C.L. Nehaniv, and B. Gorayska. Cognitive technology, humanizing the information age. In Proceedings of the Second International Conference on Cognitive Technology, pages vii-ix. IEEE Computer Society Press, 1997.

[4] C.J. Heyes. *Self-Transformations: Foucault, Ethics and Normalized Bodies*. New York: Oxford University Press, Inc. 2007.

[5] O. Sullivan and S. Coltrane. Men's changing contribution to housework and child care. Prepared for the 11th Annual Conference of the Council on Contemporary Families. April 25-26, 2008, University of Illinois, Chicago.

[6] The Cosby Show. Marcy Carsey, Tom Werner, Bernie Kukoff, Janet Leahy. Viacom Enterprises. NBC. 1984–1992.

[7] "List of Surveillance Conceptss First Introduced in Science Fiction" Technovelgy.com, Technovelgy LLC. n.d. Web. 14 May 2012.

[8] S. Brown, William Gibson: science fiction stories are quaint. BeatRoute Magazine – Western Canada's Monthly Arts & Entertainment Source. BeatRoute Magazine. n.d. Web. 14 May 2012.

[9] P. Singer. Taking Humanism Beyond Speciesism. Free Inquiry, 24, no. 6 (Oct/Nov 2004), pp. 19-21.

[10] P. Singer. *Practical Ethics*. New York: Cambridge University Press. 2011.

[11] A.I. Dir. Steven Spielberg. Perf. Haley Joel Osment, Frances O'Connor, Sam Robards, Jake Thomas, Jude Law, and William Hurt.

DreamWorks, 2001. Film.
[12] Star Trek: The Next Generation. Gene Roddenbury. CBS. 1987–1991.
[13] G. Dennis. The Encyclopedia of Jewish Myth, Magic, and Mysticism. Page 111. Woodbury (MN): Llewellyn Worldwide. 2007. Print.
[14] Terminator. Dir. James Cameron. Perf. Arnold Schwarzenegger, Michael Biehn, Linda Hamilton. Orion Pictures. 1984. Film.
[15] Terminator 2: Judgment Day. Dir. James Cameron. Perf. Arnold Schwarzenegger, Linda Hamilton, Robert Patrick, Edward Furlong. TriStar Pictures. 1991. Film.
[16] Terminator 3: Rise of the Machines. Dir. Jonathan Mostow. Perf. Arnold Schwarzenegger, Nick Stahl, Claire Danes, Kristanna Loken. Warner Bros. Pictures. 2003. Film.
[17] Terminator: Salvation. Dir. McG. Perf. Christian Bale, Sam Worthington, Anton Yelchin, Moon Bloodgood, Bryce Dallas Howard, Common, Jadagrace Berry, Michael Ironside, Helena Bonham Carter. Warner Bros. Pictures. 2009. Film.
[18] Splice. Dir. Vincenzo Natali. Perf. Adrien Brody, Sarah Polley, Delphine Chanéac. Dark Castle Entertainment. 2010. Film.
[19]GreyGirlBeast [Caitlín R Kiernan]. "...to watch you shake and shout it out..." Livejournal. The Online Journal of a Construct Sometimes Known as Caitlín R. Kiernan. 5 June 2010. Web. 9 June 2010.
[20] Blade Runner.
[21]J.J. Olivero, R.L. Longbothum. Empirical fits to the Voigt line width: A brief review. Journal of Quantitative Spectroscopy and Radiative Transfer. February 1977.
[22] Cassell's German Dictionary: German-English, English-German
[24] Terminator: The Sarah Connor Chronicles. Josh Friedman. FOX. 2008—2009.
[24] “To the Lighthouse.” Terminator — The Sarah Connor Chronicles: The Complete Second Season. Writ. Natalie Chaidez. Dir. Guy Ferland. Warner Home Video. 2009.
[25] Matt Jones, "B.A.S.A.A.P." Blog-Berg. BergLondon. 4 September 2010. Web. 15 May 2012.
[26] “Heavy Metal.” Terminator — The Sarah Connor Chronicles: The Complete First Season. Writ. John Enbom . Dir. Sergio Mimica Gezzan. Warner Home Video. 2008.