What Makes Any Agent a Moral Agent?

Reflections on Machine Consciousness and Moral Agency

Joel Parthemore and Blay Whitby∫*

***Centre for Cognitive Semiotics, University of Lund, Sweden**
∫School of Informatics, University of Sussex, UK

Keywords: moral agency, Moral Turing Test, self, self-reflection, *akrasia*, concepts, conceptual spaces

Abstract

In this paper, we take *moral agency* to be that context in which a particular agent can, appropriately, be held responsible for her actions and their consequences. In order to understand moral agency, we will discuss what it would take for an artefact to be a moral agent. For reasons that will become clear over the course of the paper, we take the artefactual question to be a useful way into discussion but ultimately misleading. We set out a number of conceptual pre-conditions for being a moral agent and then outline how one should – and should not – go about attributing moral agency. In place of a litmus test for such agency – such as Colin Allen *et al*'s Moral Turing Test – we suggest some tools from conceptual spaces theory for mapping out the nature and extent of that agency.

1 Introduction

This paper began as a presentation for a conference in Amsterdam on the theme "What makes us moral?" Given the difficulties with addressing such a question, we proposed asking an alternative question: "when is an artefact a moral agent?" as a way of addressing two further questions, which can be seen to replace the original question:

1. What makes us *us*?
2. What makes any agent a moral agent?

The advantage of approaching things this way is that it allows one to focus on mechanisms rather than dwelling on mysteries. By changing the context slightly, it forces one to re-consider one's comfortable familiarity with concepts like "I/me", "we/us", and "moral agency".

Our conclusion in this paper is that the question "when is an artefact a moral agent?" is inextricably bound up with the question "when is *any* agent a moral agent?" We will provide what we believe to be a very sound, if not unassailable, answer to that broad question. We will not, however, attempt a definitive answer to either of the subsidiary questions.

Rather, the paper provides a framework (within the structure of Peter Gärdenfors' *conceptual spaces theory* and the primary author's own *unified conceptual space theory*) for investigating possible answers.

Some are willing to attribute moral agency to certain existing artefacts. Most are not – why not? Some go so far as to say that no artefact could *ever* be a moral agent, at least unless (as John Searle is commonly read) it is a *biologically* constructed artefact. Is this mere biological chauvinism? What drives these intuitions that are often so powerful as to permit no counter-claims? Even if one were to agree that no artefact could ever be a moral agent (and we do not), one would still have the benefit of being clear about *why* not.

Colin Allen, Gary Varner, and Jason Zinser have proposed a Moral Turing Test (MTT) (Allen et al., 2000) for determining artefactual moral agency. We see little merit in this approach, both because it perverts the original intentions of the Turing test (presented more than anything as an intuition pump, to get people thinking *about* thinking) and because any agent passing some version of the MTT would not, of itself, do anything to address people's often strongly held intuitions against ascribing moral agency to artefacts. Intuitions are not, we believe, things simply to be set aside, any more than are perspectives. Nonetheless, the distancing effect provided by focusing on moral artefacts potentially allows one to remove or reduce at least some of one's biases.

When *is* an artefact a moral agent? We suggest that, for starters, it must have a concept of *self* and the capacity for self-reflection and for *akrasia*: the capacity to act against one's better judgment. After all, possessing the concepts of right and wrong is more (on most accounts) than correctly applying labels.

Section Two sets out the dependence of moral agency on conceptual agency and explains what conceptual agency entails. Section Three introduces the various conceptual building blocks we need to make our argument, and which we believe a moral agent must (at some level) possess: the concept of *self*, the concept of *morality*, the concept of *concept* itself. Section Four restates and develops our argument for when moral agency should, and should not, be ascribed to an agent, regardless of its origins. Section Five suggests some get-your-hands-dirty methods as an alternative to litmus tests for evaluating moral agency. Section Six summarizes our answers to the questions we have raised and offers suggestions on how best to push the discussion forward.

2 Moral agency and conceptual agency

> According to Kant... an action cannot be morally good unless the agent in fact reasoned in certain fairly complex ways (Allen et al., 2000, p. 253).

We take a moral agent to be any agent to which it is appropriate to attribute *moral agency*: that is, to be morally accountable for one's actions and their consequences. A moral agent is, we believe, necessarily a conceptual agent – i.e., an agent that possesses and employs concepts. (The converse need *not* be true: a conceptual agent is *not* necessarily a moral agent: i.e., moral agents are a subclass of conceptual agents.)

Note that we are not taking any stand on wider issues in moral philosophy of e.g. utilitarianism versus other forms of consequentialism versus alternatives to consequentialism. Our focus is not on what makes an act a moral act but much more narrowly on the agents who carry out those acts.

It is not enough, on our account, to be a moral agent that one does morally good things – *contra* what Colin Allen *et al* ascribe to John Stuart Mill. No one, we believe – even the die-hard utilitarian – would hold an agent morally responsible whose thoughts were not systematically and productively structured in the manner of conceptual thought. Regardless of whether Robbie the Robot is a moral agent, my Aibo dog is not. Among other consequences, this means that it is not enough for the agent merely to memorize a list of percepts.

It will be useful at this point to offer a working definition, in philosophical language, of what we take concepts to be:

> Individuable units of structured thought that are, *per* Gareth Evans' Generality Constraint (Evans, 1982, pp. 100-104) both systematic (the *same* concepts can be applied across unboundedly many contexts) and productive (a finite set of concepts can be used to construct unboundedly many complex concepts and propositions).

In addition, concepts are typically taken to be:

- Intentional: i.e., "about something" (see e.g. (Brentano, 1995, p. 88)). That is, they not only have a particular form, they have a particular (semantic) content. (In this way they mirror the structure of the semiotic sign. At the same time, of course, they are quite different, in that the semiotic sign is a communicative *expression*, and a concept – essentially – is not.)
- Compositional (see e.g. (Fodor, 1998, p. 25)) .
- Spontaneous in the Kantian sense (see e.g. (McDowell, 1996, p. 52)); which is to say, under the agent's "endogenous control" (see e.g. (Prinz, 2004, p. 197)).

Not just any conceptual agent need be a moral agent, however. Indeed, on our account, many if not most conceptual agents will turn out *not* to be moral agents. We would like to join the so-called *animal concepts* philosophers in ascribing conceptual agency to pre-linguistic infants as well as many if not most mammals, many birds, and indeed any agent that shows:

- Evidence of an ability to derive general classes from specific instances.
- Demonstration of a flexible pattern of behaviour based on this ability, especially when confronted with novel situations.
- Demonstration of surprise upon making a mistake (Newen and Bartels, 2007, p. 291)[1].

[1] A similar list may be found in (Allen, 1999, p. 37).

To put this another way: one should attribute minimal conceptual abilities to an agent when the most parsimonious explanation for that agent's behaviour is that, presented with the *same* circumstances on *different* occasions, the agent makes different choices based on some awareness by that agent of its past experiences[2]. Newen and Bartels discuss some sophisticated experiments, originally presented in (Pepperberg, 1999), strongly suggesting that the parrot Alex meets these criteria. Scrub jays (Clayton and Dickinson, 1998; Raby et al., 2007) and ravens (Bugnyar and Kotrschal, 2002) have shown quite sophisticated flexibility in their caching habits, and ravens furthermore show sensitivity to others' perspectives (Bugnyar, 2011) and adapt their behaviour to social context (Bugnyar and Heinrich, 2006) – all of which would seem to presuppose productively and systematically structured thought. Meanwhile, higher primates have been shown to pass the mirror self-recognition task.

Exactly which non- or pre-linguistic agents meet these criteria is not the point here – only that some do. We do not, most of us, hold an infant morally responsible for pulling the cat's tail. Neither do we hold the cat morally responsible for eating the food off our plate when we are not looking. Both are, plausibly if not untendentiously, conceptual agents. (Jean Piaget, who coined the term *object permanence*, famously claimed evidence for this concept at age nine months (1954); more recent research [e.g. (Ballargeon, 1987)] has shown reliable evidence for an expectation of object permanence at less than half that age. Meanwhile cats, as cat owners will attest, can have quite sophisticated personalities: something one might not expect in a purely stimulus-response driven agent.)

If conceptual agency on its own is not sufficient for moral agency, what additional conceptual tools are necessary?

3 Conceptual building blocks

Moral agency requires that an agent possess certain concepts that not all conceptual agents necessarily possess. Most foundational among these is a concept of *self*. An agent cannot be held morally responsible for its actions if it has no concept of itself as the agent who is acting.

The concept of self, however, requires unpacking. As Daniel Dennett has famously pointed out (1991, p. 174), there is a substantive sense in which *every* living organism has a "concept" of self: every organism, in order to survive, must make an operational distinction between self and non-self.

Rather than being polysemous, the word "self" brings together, we believe, a number of closely related concepts of self that may usefully be understood as arranged in a hierarchy (or, alternatively – we see the two as roughly equivalent – along a continuum). It is only an

[2] As Fodor puts it, in explaining why human beings definitely have mental representations and paramecia definitely don't: "unlike paramecia, we are frequently implicated in primal scenes in which the behaviorally efficacious stimulus property... is nonnomic. Or, as I shall sometimes put the point in order to achieve terminological heterogeneity: the difference between paramecia and us is that we can 'respond selectively' to nonnomic stimulus properties and they can't" (Fodor, 1987, p. 10). Fodor's conditions in that paper for attributing mental representations to animals are, though worded quite differently, strikingly similar in spirit to Newen and Bartels' list, or Allen's list, for attributing concepts.

agent with a concept of self toward the higher end of the hierarchy/continuum – what we call I_2– that qualifies as a potential moral agent.

3.1 The concept of *self*: what makes us *us*?

That most basic notion of self that Dennett refers to is not, by any account, a concept at all – not unless one wants concepts to go "all the way down". Certainly it does not meet the conditions given above for when to attribute conceptual agency. Call this "concept" of self I_0.It is the *non-conceptual* foundation on which all the things one might reasonably call a concept of self rest.

The most basic concept of self would be a concept of this non-conceptual "self": a first-order concept of the organism "as a whole", without distinction of body or mind (or anything else). Call this self I_1. Jordan Zlatev (2001, p. 173), following Ulric Neisser (1988), calls this "initial self-awareness" that "acts here and now... but remains unreflected upon" the *ecological self.* It is close kin to Antonio Damasio's (2000) notion of the *core self.* The cat, possibly, and the pre-linguistic infant, probably, have a concept of self in this sense.

There is another, quite different concept of self that could best, we believe, be described as the agent's concept of *its concept of itself*. Call this self I_2. This is the higher-order *self-as-myself* that most humans entertain, and which requires, or creates, the body/mind distinction. This is the self-reflective self that is, if one is careful not to confuse the metaphor with the reality, the homunculus sitting in his Cartesian theatre of the mind, controlling the shell of an organism in which he sits and observing all that it observes. Who does the "I" who thinks "I" think that "I" is? The comparison here is to Damasio's notion of the *autobiographical* self. We maintain that, to be a moral agent, an agent must, minimally, have a concept of self in this sense.

3.2 The concept of *morality*

It is not enough, on our view, for an agent to know that *certain* things are "right" or "wrong" in order to qualify that agent as a moral agent and so responsible for her actions. No number of individual precepts – e.g., pulling the cat's tail is wrong, punching one's younger sibling is wrong – are sufficient for this purpose. These precepts cannot exist in isolation; otherwise there will be the nagging concern that the agent is doing no more than correctly applying the labels of "right" and "wrong" to particular (rote memorized) situations. This is (part of) why an artefact supplied with a list of moral rules would not be a moral agent in virtue of its compliance with those rules; nor would any amount of refinement of or addition to the rules make it into one.

Indeed, the agent must have a sense of right and wrong independent of *any* particular precept, must have a larger structure into which all the precepts fit: in short, the agent must have a concept of morality as well as a well-fleshed-out moral domain *and a commitment to or belief in it.*

To be clear: we are making no effort here to further define morality or the "proper" concept of it. That would be beyond our remit.

3.3 The concept of *concept*

Finally, the moral agent must, we believe, have some understanding of what concepts are – possibly (and in many cases probably) only implicitly (i.e., not subject to self-reflection). That is to say, the agent must have some concept *of* a concept. If that sounds too demanding or at risk of over-intellectualizing matters, then consider it this way, which we take to be equivalent: the moral agent must have some (if only implicit) understanding of how its thoughts are structured.

Specifically: if the agent is capable of entertaining the proposition "my sibling punching me is wrong", that agent should at least be amenable to the proposition "I should not punch my sibling" and open to the argument that failure to see a connection between the two suggests a logical flaw. Likewise the agent should be able to see the connection to the more general proposition that "punching others is wrong", and able to see the connection between "punching others is wrong", "kicking others is wrong", "calling others names is wrong" and so on, from which one might derive the yet more general proposition, by induction, that "causing others unnecessary suffering is wrong". That is to say, the agent should be able to see the *systematicity* and *productivity* of morality: the way that the same morals can be applied *systematically* across unboundedly many contexts, and the way a finite set of moral axioms can be used *productively* to generate an unbounded number of moral principles. Such structuring mirrors the systematicity and productivity of conceptually structured thought.

Of course, it is not enough for an agent to possess these concepts: the concept of self, the concept of morality, the concept of a concept itself. The agent must demonstrate not only that it possesses these concepts but that it can *employ* them appropriately over an extended period of observable interactions. In short, the agent must satisfactorily demonstrate that it can *take responsibility.*

3.4 *Akrasia* and the concept of *boundary*

> ... A moral agent is an individual who takes into consideration the interests of others rather than acting solely to advance his, her, or its... self-interest (Allen et al., 2000, p. 252).

Before elaborating that point, however, it is necessary to make a brief detour. In the introduction, one of our requirements on moral agency was that the agent have a capacity for *akrasia*: the capacity to act against one's "better judgment" or against the apparent requirements of "pure (selfish) reason", narrowly defined[3]. Such behaviour is not, we claim, either accidental or incidental. The capacity of agents to get things "wrong" in this sense is essential to its being a moral agent.

It is worth noting that this capacity is seen, at least to a limited extent, in species who do not, on our account, qualify as conceptual agents at all. Not only is there the mother sacrificing herself for the sake of her offspring, there are the reports of predators "adopting" what would normally be their prey, such as a snake "adopting" a mouse or a hamster. Humans, of course

[3] Needless to say, we are not attempting to provide any definitive interpretation of Aristotle's term *akrasia*. We use the term because it is the *best* term to capture the way that a moral agent need not and indeed cannot always act in accordance with the agent's selfish interests, narrowly defined.

– at least on occasion – take self-sacrifice to a level not observed (or at most very rarely observed) in other species: sacrifice on behalf of strangers, sometimes people one has never met until the moment one intervenes.

The concept of *boundary* is key here (whether or not the agent in question possesses such a concept). Where do *my* needs stop and *yours* or *theirs* begin? Even more fundamentally: where do *I* stop and *you* or *he* or *she* or *they* or *the world* begin?

In developmental psychology, the self/other (self/non-self, self/world) distinction is seen as foundational to all the other concepts human beings acquire (see e.g. the discussion in (Zachar, 2000, p. 144 ff.)). At the same time, as the primary author argues in (2011b), the distinction is, *itself*, a conceptual one; and, like all conceptual boundaries, subject to shifting over time – not *too* much however, or the conceptual structure breaks down. So in ***Supersizing the Mind***, Andy Clark writes of "profoundly embodied agents" (by which he means to include human beings!) who are "able constantly to negotiate and renegotiate the agent-world boundary itself" (2008, p. 34).

Applied to moral agency and *akrasia*, the moral is that what counts as *akrasia* depends not on *whether* one draws a boundary between self and other (since such a boundary seems to be conceptually obligatory) but *where* one draws it at any particular time. A moral agent – we strongly suspect – is such an agent as Clark describes who has the capacity "constantly to negotiate and renegotiate the agent-world boundary".

4 Ascribing moral agency: why the artefactual element is a red herring

> Let us consider the following Wittgenstein-inspired thought experiment: a person who has lived a normal life in our community dies and in the autopsy it is discovered that there is some kind of a device instead of a brain in his head. Would we on the basis of this decide that we had been fooled all along and that the person was actually a 'brainless' automaton, lacking any real language and meaning? I believe that the answer is: hardly (Zlatev, 2001, p. 160).

Given how widely people's intuitions differ on whether "some kind of a (mechanical) device" can, even in principle, be intelligent or conscious, we prefer a slightly different version of Zlatev's thought experiment. Imagine that, as one of the authors is presenting this paper at a conference, he suddenly collapses at the podium. A doctor is on hand and performs an autopsy, at which time it is discovered that the author's head is full, not of grey matter, but of yogurt. We take it as safely assumed that no one believes yogurt a candidate for intelligence or consciousness. Never mind the audience: would Blay's or Joel's friends and family decide that they had somehow been cleverly fooled? Should they?

John Searle's (1980) classic Chinese Room thought experiment depends critically on the assumption that they *should*. He believes that, if one looks inside the room (or, by analogy, inside the skull), sees the operations there, and "knows" that those operations could not even in principle produce intelligence/consciousness, then there is no intelligence/consciousness – *regardless of any observable behaviour or the lack of any measurable differences whatsoever.*

Our intuitions go the other way. Although we are strongly inclined to believe, as Searle does, that intelligence/consciousness cannot be explicitly programmed, nonetheless, if someone were to produce a Chinese Room as he describes, we would be inclined to revise our understanding of what is necessary for intelligence/consciousness – or look more closely for some explanation other than the immediately obvious one. We would *not* conclude that the Chinese Room was not intelligent/conscious just because "we can see how it works inside". Likewise, we would hope that our friends and family would not write us out of their memories but would remember us for the genuine friend we had been to them.

In his version of the thought experiment, Zlatev considers the *death* of the agent to be critical: "the 'death' of the person-robot makes it too late to investigate the causal relationship between his 'hardware' and behavior, and even if we may have some doubts, we have no way of substantiating them" ((2001, p. 160). Consider, however, a variation on the "yogurt" thought experiment: rather than dying, the author merely collapses, and the "yogurt" factor is discovered – after which the author recovers and carries on normally. Should one *then* write off one's whole past history of interactions with the person? Should one discount the *apparent* nature of all current interactions? Should one, indeed, refuse to interact with the person, because, after all, one now "knows" that person to be a yogurt-driven automaton? The authors might be forgiven for hoping not!

Bottom line: if there is no observable difference in behaviour *or any other measurable difference*, then there is no practical value in asking whether or not an agent "really" is intelligent/conscious. Consider the conceivability of so-called philosophical zombies (indistinguishable from conscious agents, except that "no one is home") as suggested by David Chalmers in (1996). *Pace* many of Chalmers' critics, we believe that philosophical zombies are, in fact, conceivable. Indeed, it is conceivable that everyone reading this paper, *everyone in the world except for me* is a P-zombie; *but why should that make any difference to me? Pace* Chalmers, we do not believe that the conceivability of philosophical zombies proves anything (about physicalism or otherwise), and that conceivability is a far weaker measure of reality than most people give it credit.

If we are right, then the *interesting* question is not when an artefact is a moral agent; the interesting question is when *any* agent is a moral agent. The artefactual element is a red herring precisely because the agent's origin or details of its internal construction are, although highly relevant, not, ultimately, what matter.

4.1 Little green men

> *If an artificial autonomous system (a robot) with bodily structure similar to our (in the relevant aspects) has become able to participate in social practices (language games) by undergoing an epigenetic process of cognitive development and socialization, then we may attribute true intelligence and meaning to it* (Zlatev, 2001, p. 161).

As should already have been implied, we want to go further than this, and it is here we must part company with Zlatev, at least as he presents himself in the 2001 paper. (Of course, his position continues to evolve, as does our own.)

Consider: in ***The Black Cloud***, a science fiction tale by renowned astrophysicist Fred Hoyle (1959), Hoyle writes of a giant cloud of interstellar gas that threatens to wipe out life on Earth by cutting off most solar radiation. The scientists investigating it observe it acting in a way suggestive of intelligent behaviour (and otherwise contrary to physical laws). Despite the odds, they succeed in establishing communication with it.

On Zlatev's terms, as quoted above, such a hypothetical entity would, *by definition,* be neither intelligent nor capable of meaning-making: its "body" would not be similar to ours in any relevant aspects. Needless to say, we believe this to be the wrong conclusion. So would Hoyle, who clearly believed his fictional creation to be not just an intelligent but a *moral* agent – albeit one whose morality did not entirely coincide with strictly *human* morality.

Of course, if it is human-like intelligence one wishes to explore, then it makes sense to design a potentially intelligent/conscious artefact to be as physically human-like as possible, down to details of its musculoskeletal structure. This is the inspiration behind the *anthropomimetic* approach taken by Owen Holland and his colleagues at the University of Sussex, UK (see e.g. (Marques et al., 2007)). At the same time, the second author on this paper has argued quite forcefully (Whitby, 2003, 1996) against approaching intelligence too much from the standpoint of human intelligence, as if the latter were necessarily representative of all the intelligence out there; instead, "science has to be interested in the whole space of intelligence" (Whitby, 2003, p. 3). The same point holds for moral agency: if it is the *whole space* of moral agency one is interested in, then one should take care, as much as possible, not to limit oneself to *human* moral agency.

Neither, we believe, should it matter, in the end – for all their critical importance, which we do not contest! – what (if any) "epigenetic process of cognitive development and socialization" the agent has undergone *prior* to one's first encounter with the agent. (We take it as untendentious that those same interactions *can* be quite essential from that point forward, since it is through such interactions that we judge other people: e.g., "he never learns", "he's *still* the same spoiled child I met twenty years ago", etc.) Consider the Swamp Man of comic book fame, who is created in an instant *as if he had a whole history of such development and socialization that he does not, in fact, have*. Donald Davidson writes (Davidson, 1987, pp. 443-444) of his own Swamp Man doppleganger:

> No one can tell the difference. But there *is* a difference. My replica can't recognize my friends; it can't recognize anything, since it never cognized anything in the first place. It can't know my friends' names (though of course it seems to), it can't remember my house. It can't mean what I mean by the word 'house', for example, since the sound 'house' it makes was not learned in a context that would give it the right meaning – or any meaning at all. Indeed, I don't see how my replica can be said to mean anything by the sounds it makes, nor to have any thoughts.

Zlatev (2001) seems to hold a similar intuition. So, for example, in qualifying the thought experiment with which we opened Section 4, he writes (2001, p. 161):

> ... A very important source of evidence in deciding whether our neighbor... had genuine intentionality or not... is the answer to the question: *how did he acquire*

> all the physical and social skills necessary for us to think of him as just one of us? If we are told (after our cognitive scientists work on him extensively, either dead or alive) that it was through preprogramming by super-intelligent engineers (or perhaps by aliens) then again we would be disposed to accept the possibility that he was an ingenious automaton after all.

Perhaps "we" would; but we do not believe that anyone should be. Neither should it affect whether or not one is willing to ascribe moral agency and its accompanying responsibility to that agent. At risk of repeating ourselves: *if there is no observable difference in behaviour* ***or any other measurable difference****, then there is no practical value in asking whether or not an agent "really" is intelligent/conscious...* or morally responsible.

If the reader still is not convinced, consider the following thought experiment, for which the authors wish to gratefully acknowledge Ron Chrisley[4]: imagine that you are interacting with a robot in a way that, were you interacting with another human being, it would unambiguously count as torture. Furthermore, the robot is making all the appropriate responses for someone who *is* being tortured and who feels every agony of it. However, suppose you know that the robot came off the assembly line just an hour ago. Should you feel free in continuing to "torture" it, because you "know" it doesn't have the "right" causal history?[5]

Let us be clear: just as we are inclined to agree with Searle[6] that intelligence/consciousness *cannot* be achieved solely or even primarily by explicit programming – it certainly *seems* difficult to imagine! – so we are inclined to agree with Zlatev that intelligence/consciousness cannot be achieved without the "right" causal history[7]. Indeed, Zlatev devotes much of his paper to why one should expect this to be the case, and on all points we agree. It certainly *seems* to us that a Swamp Man could not exist. However, if a Swamp Man *did* exist, we would not, by mere stipulation, deny him intelligence/consciousness; neither would we absolve him of moral responsibility for his actions. One should no more take Swamp Man to be a mindless automaton than one should any fellow human being; one should no more absolve him of moral responsibility than one should any human one treats as morally

[4] Personal communication.

[5] Of course there are other reasons one might offer why one should not torture such an agent: e.g., that, regardless of whether one "knows" that the agent is just a clever automaton, nonetheless the agent's *merely superficial appearance* is enough to constrain proper behaviour towards it, perhaps because of how it will incline one's action toward other agents whom one appropriately understands *not* to be automata: e.g., one's fellow human beings. On such a view, it is immoral to abuse Asimo in a way that it is not immoral (or less immoral) to abuse e.g. an industrial robot. However, that is *not* the argument we are pursuing here.

[6] This is not to agree with him very far.

[7] One could argue that a chess-playing computer shows a kind of (non-human) intelligence; and, in a sense, it does. At the same time, while there is a sense in which the computer is playing chess (i.e., calculating moves from a set of rules and heuristics), there is another sense in which the computer is not playing chess at all, *unless and until* it shows (self-)awareness of moving the actual pieces on the actual board as part of a wider cognitive and physical context. To wit, a game of chess played out "in one's head" is a different game from one played in a coffee house and is different again from the one played over the Internet. Chess-playing computers do not attempt to psych out their opponents by e.g. staring at them refusing to take toilet breaks. Domain generality and engagement with a wider context are essential not optional aspects of (human) intelligence. IBM's Jeopardy-playing Watson achieves this *to a point* – it does a remarkable job of handling whatever subject categories one throws at it in Jeopardy – but if one is still disinclined to grant it (human) intelligence or consciousness, perhaps that is because one still cannot have any kind of casual conversation with it even of the five-minute kind Alan Turing had in mind. In any case, what intelligence it shows is not all pre-programmed. Much of it is, in some substantive sense of the word, "learned".

responsible, *or indeed anything that meets the conditions we have set forth for ascribing moral agency.*

4.2 Autopoiesis

Usefully, *autopoiesis* offers a way to break out of any overly narrow biological view on life to take in the possibility of Swamp Men or robots or other agents whom one might want to ascribe moral agency to but who are not, in the comfortably familiar way people think about such things, alive.

Autopoiesis is a term popularized by biologists Humberto Maturana and Francisco Varela (see e.g. (Maturana and Varela, 1992)). It is intended as an alternative description of what qualifies as a living organism, in terms of *operational closure* (processes of the system are produced from within the system; anything external to the system plays only the role of catalyst), autonomy (self-determination, or the observation that organisms "are continually self-producing" (Maturana and Varela, 1992, p. 43)), and adaptivity (REF). Again, boundary is key, but, at least on Maturana and Varela's account, the boundary is only identifiable *relative to the perspective of an observer*, and is matched by a (logically deducible) underlying continuity.

An autopoietic system need not be implemented in DNA; neither need it be – even in principle — capable of reproduction. Autopoiesis offers a way to understand and embrace Zlatev's *(2009) semiotic hierarchy,* whereby consciousness is dependent on life, and signs (which the moral agent must be conversant in, if it is to communicate its moral agency) dependent on consciousness.

It is worth remembering at this point where the term "robot" originally came from: the play known in English as *Rossum's Universal Robots* by the Czech playwright Karel Čapek. Čapek's robots were not the mechanical "thinking machines" one is used to thinking of as robots nowadays, but much closer to artificial life.

4.3 The Moral Turing Test

> Turing's intention was to produce a behavioural test which bypasses disagreements about standards defining intelligence or successful acquisition of natural language. A Moral Turing Test (MTT) might similarly be proposed to bypass disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human 'interrogators' cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent (Allen et al., 2000, p. 254).

We have set forth one method for determining moral agency, as the ability to demonstrate, over an extended period of time, the possession and appropriate deployment of a range of sophisticated concepts and conceptual abilities. One might be tempted to complain that it is too indefinite, that it leaves too much open to interpretation.

Allen *et al* have proposed a test for moral agency that, in contrast, is quite unambiguous. It faces an immediate challenge, however: it requires the agent to be a *linguistic* agent. Although language sits at the top of Zlatev's semiotic hierarchy, we deliberately chose not to mention it earlier, because although sign use is clearly critical to the communication if not possession of moral agency, language use is not. Allen *et al* recognize the problem – do we really want to deny (any level of) moral agency to pre-linguistic children and non-linguistic animals? – and so offer an alternative:

> To shift the focus from conversational ability to action, an alternative MTT could be structured in such a way that the 'interrogator' is given pairs of descriptions of actual, morally-significant actions of a human and an AMA, purged of all references that would identify the agents. If the interrogator correctly identifies the machine at a level above chance, then the machine has failed the test (Allen et al., 2000, p. 254).

Colin *et al* recognize problems with this formulation as well, but the concerns they raise are not the ones we wish to focus on. Contrary to most popular readings, "the contrivance of the imitation game", as the second author to this paper has written (Whitby, 1996, p. 62), "was intended to show the importance of human attitudes, not to be an operational definition of intelligence." Indeed, we think that Turing would have objected to any such litmus test, as we do, on strong ethical grounds. One would never subject a human being to a "Turing test" to decide whether that person was "actually" intelligent; neither would one submit a human being to any form of the Moral Turing Test to decide whether that person was morally responsible[8]. *So why would it be appropriate to apply to an artefact?*

5 Conceptual spaces and the unified conceptual space theory

If the Moral Turing Test risks being too crude – as its authors might readily allow – it also is misleadingly precise, the equivalent of starting with one significant place in a mathematical calculation and ending up with ten. That is not to say, however, that a *certain* degree of precision cannot be achieved. Indeed, we have already, in sections Three and Four, set out those conceptual pre-conditions that an agent must meet to *be* a moral agent and those circumstances under which an agent should be *attributed* moral agency. It is now time to say something about how one might operationalize those methods, in a way that falls (deliberately) short of providing a litmus test. That is, the focus is not on determining (at least in any definitive way) whether a given agent is a moral agent, but rather on exploring a given agent's moral agency: mapping out its territory, identifying its prominent features, and saying something about its limits.

Remember that we have tied moral agency to the possession and observable appropriate employment of a number of key concepts – or, if one prefers, conceptual abilities. Therefore, it seems fitting that the tools we suggest in place of the Moral Turing Test are taken from the literature on theories of concepts within philosophy of mind: in particular, the *conceptual spaces theory* of Gärdenfors (2004) and the *unified conceptual space* extensions to it offered

[8] Of course, in legal proceedings, rulings are made as to whether a particular person can or cannot be held ***legally*** responsible *with respect to some particular action or actions.* Regardless of whether these rulings can be considered a kind of litmus test, what they are not is a test of general ***moral*** agency.

by the first author on this paper ((Parthemore, 2011a); an earlier version can be found in (Parthemore and Morse, 2010)).

5.1 Conceptual spaces theory

Conceptual spaces theory (CST) is a *similarity-space-based* theory of concepts, which means that concepts within a common domain are locatable along a number of shared dimensions such that the distance between them – with respect to those dimensions and whatever metric defines the resulting space – corresponds to their presumed similarity. As a similarity-space-based theory, CST is closely related to prototype theories, as first popularized by Eleanor Rosch (1975; 1999) and her colleagues.

Different domains are defined by different shared dimensions – known as *integral dimensions* because it is not possible to specify a value on one dimension without simultaneously specifying a value along the others. For the colour space, the integral dimensions are hue, saturation, and brightness (Gärdenfors, 2004, p. 9). For the "tone" space, the integral dimensions are pitch and loudness (Gärdenfors, 2004, p. 26).

Most concepts – what Gärdenfors calls the *natural concepts* – can be understood either as points or as convex shapes within these spaces (i.e,. the convex shapes can be collapsed to points). Convexity means that if one point within the space is assigned to a certain concept, and another point is assigned to the same concept, then all points lying between them in that space should belong to that concept. A few concepts are *not* convex relative to their domain: e.g., the concept *Gentile*, which includes everyone who is not a Jew. Likewise the concept *heterological* applies to all adjectives that are *not* self-descriptive (e.g., *monosyllabic* as opposed to *polysyllabic*). It follows that if a certain concept is convex, its negation within the domain cannot be.

A single such convex shape, on its own, corresponds to a sub-category of concepts: *property concepts*, which relate to the grammatical categories of adjectives and adverbs. All other concepts are associated sets of such shapes across multiple domains: e.g., *object concepts* (corresponding roughly to nouns) and *action concepts* (corresponding roughly to verbs).

The carving up of a domain into its constituent concepts and sub-concepts imposes a *Voronoi tessellation* on the space. A Voronoi tessellation tiles an n-dimensional space that is initially populated by a set of points (the Voronoi *sites*), which in conceptual spaces theory are taken to represent the most prototypical members of a category. The space is then divided up according to which of those points the remaining points in the space are closest to (the Voronoi *cells*). Boundaries arise wherever there is equidistance to two of the existing points, junctions wherever there is equidistance to three (or more) points.

5.2 Unified conceptual space theory

The unified conceptual space theory (UCST) fills in some of the missing details in CST, at the same time pushing it in a more algorithmically amenable and empirically testable direction. UCST attempts to show how the many different conceptual spaces discussed in CST can all be integrated in a single unified space of spaces, describable along dimensions

that are *integral dimensions* for all concepts as well as three foundational proto-conceptual entities: *proto-objects, proto-action/events,* and *proto-properties*[9].

In the development of the theory to date, the focus has been on concepts *for an individual.* The assumption is made, however, that an analogous space must logically exist for a society – for those conceptual agents who are social animals – mapping together the different conceptual spaces of its members into a unified space of the whole. (We do not mean by this to imply that the direction is necessarily from the individual to the society. Indeed, Pierre Steiner and John Stewart argue convincingly (Steiner and Stewart, 2009) that much of social cognition does *not* begin with or reduce to an agglomeration of individuals, but must be taken as foundational to cognition. A similar view may be found in (Jaegher et al., 2010). Rosch, who has long argued that categories are intrinsically cultural artefacts – see e.g. (Rosch, 1999, p. 189) – would likewise be inclined this direction.)

A key insight of the unified conceptual space theory is that concepts can be given two contrasting structural descriptions: one from geometry, one from logic. The geometrical description is borrowed straight from CST. The second is only hinted at in CST: concepts may *also* be described as a structured set of logical relations to other parts of the unified space. This is to say, concepts are defined *both* by the concepts with which they are contiguous, within the same (sub-)domain; and by the concepts to which they are in one way or another associated, in adjacent or distal domains.

The axes of the unified space include:

- An **axis of generalization**, from the most general categories (superordinate) to the most specific ones (subordinate). All concepts, except the most general (i.e., the most general (proto-)concept of concept itself) have at least one superordinate. (They may have different superordinates with respect to different domains: i.e., the axis is divergent.) All can, at least in principle, have subordinates. Concepts more toward one end of the *axis of generalization* are most readily understood as classes or *types*, toward the other end as instances or *tokens*. However, it is a key principle of UCST that all tokens can, in principle, be treated as classes of yet more specific tokens.
- An **axis of alternatives**, obtained by adjusting the value of one or more integral dimensions of a particular (sub-)domain at any fixed point along the axis of generalization, according to the metric by which those dimensions are defined. The colour domain, for example, has the integral dimensions of hue, saturation, and brightness. Like the *axis of generalization*, this axis is divergent in both cases: in this case, depending on which integral dimensions are being attended to.
- An **axis of abstraction**, from maximally concrete and physical ("zeroth order") to maximally abstract and "mental" ("second/third/higher order"): from non-concepts to concepts of non-concepts to concepts of concepts, concepts of concepts of concepts, and so on. Note that, at its one extreme, this axis converges with the axis of generality: a maximally general category and a maximally abstract one amount to the same thing. The converse is not the case, however: a maximally specific category need not be a

[9] Gärdenfors attempted to define such a unified space in a draft chapter for his 2004 book but was not satisfied with the result and left it out (personal communication). Some thoughts toward defining a unified *action* space can be found in (Geuder and Weisgerber, 2002).

maximally concrete/physical one. Objects and action/events lie more toward one end of the axis, (increasingly abstract) properties more toward the other.

In keeping with CST, concepts in UCST can be understood either as (mostly convex) shapes within the unified space or as points. When they are understood as points, they have both *local* and *distal* connections within the space. Local connections are to contiguous points: i.e., along one of the axes of the unified space. Distal connections represent the *logical description* we referred to above (in contrast to the *geometrical description*). Distal connections can be any of three types:

- *Certain* concepts (primarily those toward the "concrete" end of the *axis of abstraction*) decompose into parts. Such parts or ***components*** will be ordered (their arrangement cannot be arbitrary), and one or more will be necessary.
- *All* concepts possess integral dimensions. Such dimensions or ***parameters*** will be necessary but *not* ordered: e.g., *colour* has the parameters *hue, saturation,* and *brightness*, but *hue, saturation*, and *brightness* are not ordered with respect to one another. Note that the parameters of a concept define a *conceptual space* of their own. This is the sense in which the term "conceptual space" is primarily used by Gärdenfors.
- *All* concepts have associated with them various contextual elements. Such elements or ***contextuals*** are neither ordered nor (individually) necessary; rather, they are typically co-present with the concept in various contexts.

CST and UCST are applicable to operationalizing the methodology of sections Three and Four and exploring moral agency in two complementary ways.

5.3 What is possible now: Modeling moral agency

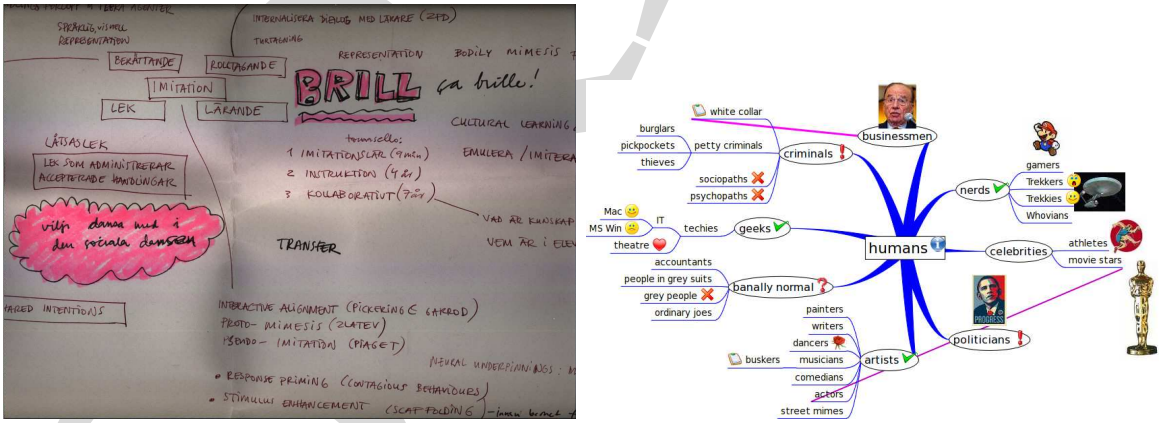

Fig. 1: Left: portion of a hand-drawn mind map in Swedish and English, copyright Åsa Harvard. Used by permission. Right: portion of a software-generated mind map using a traditional mind-mapping application (the freeware tool "View Your Mind", available from http://www.insilmaril.de/vym/).

An implementation of UCST has been created in the form of a mind-mapping application (Parthemore, 2011a): the name commonly given to a software tool for helping people organize their thoughts and the connections between them without the immediate necessity of organizing them into a linear flow. A hand-drawn *mind map* (or *concept map* as it is also sometimes called) is shown in Figure 1 alongside a software-drawn map. The software-drawn map mimics and formalizes (hence constrains) the process described in the hand-drawn map. What makes the UCST application different is not only its striking visual appearance but the way, *unlike any of the other available applications,* it implements, step by step, a specific theory of concepts.

Figure 2 shows the UCST application: initial screenshot and screenshot after the application has been in use for some time. Anywhere in the map, one can zoom in or out along the *axis of generalization.* Left to right on the picture determines whether one is looking at the concept more as an object (left) or more as an action/event (right); bottom to top determines whether one is talking about a less abstract concept (bottom) or a more abstract one (top). Movement along the *axis of alternatives* is not currently possible[10]. Any node (the hollow or blue-filled circles) can have attached to it an arbitrary label.

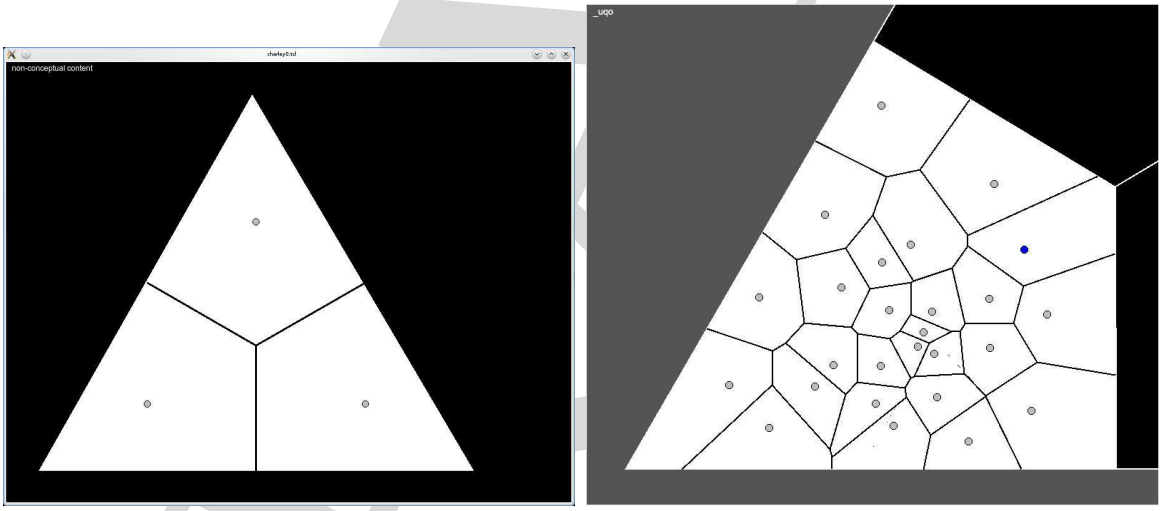

Fig. 2: The UCST application.

Unfortunately it is not possible for reasons of space to go into the details of implementation here, although they are covered in detail in (Parthemore, 2011a). Suffice to say that the

[10] The current implementation has a number of other limitations. It is not currently possible to delete a node or to delete a distal link once added. In consequence, it is not possible to remove partitioning from any area of the unified space. Re-partitioning is supported to a limited extent: one can move points around to force re-partitioning, but while the boundaries of the child nodes are correctly updated, the locations of the central (prototype) points are not. Also, update of boundaries and central (prototype) points for grandchild, etc., nodes does not take place. Obviously, a change at any level of the generalization hierarchy should force a cascade of changes all the way down to the base level. Furthermore, it is not currently possible for two nodes to have the same name (*polysemy*) or for one node to have more than one parent (i.e., to be understood relative to more than one domain). A mechanism is needed for choosing between multiple parents when scrolling "upwards" along the axis of generalization.

UCST application reproduces the functionality of the traditional mind-mapping software while constraining the resulting maps in important ways. (One of the common criticisms against traditional mind-mapping software is just how badly it is under-constrained: see e.g. (Ruiz-Primo and Shavelson, 1996).) Note that one important thing lacking from the current implementation is any visualization of the distal connections mentioned in Section 5.2 – the equivalent of the arcs between nodes in Figure 1. The arcs still exist – of course; but they lack a graphical representation.

So the first way that CST/UCST can be of use is in providing a tool via the UCST application to construct a model of a moral agent's *moral space* (or to allow the agent to construct an "externalized" model of its own moral space): either as a snapshot or, more usefully, as a developing moral space over time. Just as in one's usual conceptualizing, choices made early on constrain later ones. Although local portions of the resulting map can be broken down and re-built, the overall map has an increasing inertia to it: the more complex it becomes, the more difficult it becomes to change substantially.

To emphasize: such a model will not determine (or not determine conclusively) whether a given agent is or is not a moral agent [11]. What it can do is provide useful clues to both the shape and the extent of any moral agency. Does the agent, for example, have a concept of *causing mental anguish* or only one of *causing physical pain*? Does it have a concept of *justice* or only one of *fairness*? Does it understand *guilt* or *shame*? What do the details of those concepts look like: how fully are they fleshed out, and how much are they tied into other, distal concepts?

How might moral mind mapping within the context of the UCST application work? The overall moral space constitutes the agent's concept of morality. This should consist both of what it is *to be moral* or *to act morally* (an abstract action/event) and what is a *moral* or *percept* (an abstract object). These, in turn, should have certain properties of *right/wrong*, *fair/unfair*, *obligatory/optional*, etc., the extent of which will depend upon the sophistication of the agent's moral agency.

The most general concept of *moral* will contain "within" it, along the *axis of generalization*, specific morals of what the agent should or should not do (or did or did not do) in increasingly finer detail, with respect to increasingly particular situations. Likewise, the most general concept of *to be moral* or *to act morally* will contain "within" it specific actions to carry out or avoid. Along the *axis of abstraction,* the most "concrete" morals or moral actions will address physical objects and action/events involving those objects, and actual situations as they have occurred or are anticipated to occur; the most abstract ones will address abstract objects and action/events, as well as increasingly distant hypotheticals. Along the *axis of alternatives*, one will find, at any given point along the axes of generalization and abstraction, various closely related morals or moral actions as the values of one or more integral dimensions are adjusted: e.g., one should not punch one's siblings, one should not kick one's siblings, one should not bite one's siblings, etc.

[11] Obviously, if there is no identifiable moral space, or if certain foundational moral concepts are lacking, then, so far as the mind map is concerned, there is no moral agency.

5.4 Looking to the future: Enacting moral agency

Both the UCST theory and the mind-mapping application are, for now, very much in development. A more mature version of both – resulting from several further iterations of the theory→model→implementation→theory loop employed so far[12], as well as (we hope) empirical testing with human subjects (currently in the discussion stages) – would make possible not only the practical application of the research program outlined in the previous section but also a significantly more ambitious project: to give the UCST formalism a degree of autonomy by embedding it in a robotic platform, such as a remote-controlled AIBO[13].

One of the major shortcomings of the current theory/implementation is its lack of discernible embodiment. Of course the mind-mapping application can reasonably be described as offloading embeddedness and embodiment (indeed, its very dynamics) onto the application user – which is not entirely a cheat! With an "autonomous"[14] robotics platform, however, one could begin properly to explore issues of salience: not just articulating moral meaning but *making* that meaning; not just drawing a map of the moral territory, but creating the territory at the same time. This is to say: the current UCST application makes possible exploring only the *application* side of moral acquisition/application; a robotic platform would allow one to begin to explore the acquisition side (and discover whether, indeed, the same means of representing knowledge can be employed for both).

Such an "autonomous" system would need some way of automatically extracting conceptual dimensions according to some measure of salience (not yet explored, although, in some contexts at least, *latent semantic analysis* (LSA)[15] offers possibilities) and some way of automatically generating the metric for the resulting spaces in a way such as Janet Aisbett and Greg Gibbon have suggested (Aisbett and Gibbon, 1994, particularly p. 143). One could explore such possibilities as how much or how little initial conceptual structure – specifically, how much or how little initial *moral* structure (e.g., the ubiquitously referenced "Three Laws") – are needed in order for the agent to derive a reasonable approximation to a *human* moral space.

Clearly, the process of getting there will be neither straightforward nor easy. However, it does offer one road map to how the ambitions of Allen *et al* (2000) and others – of enacting *artificial moral agency* – might be realized.

6 Conclusions

On the one hand, we have set forth what are, by many accounts, *very* stringent requirements on moral agency, according to which a agent cannot be held morally responsible for its actions unless it possesses a rich, interconnected set of concepts or conceptual abilities. On the other, we have allowed in, as moral agents, agents that others – so far as we can tell, by stipulation – would wish to exclude. We do not expect to meet a Swamp Man, but if we do,

[12] We take a close tension between ("armchair") theory and (hands-on) application to be a key to successful progress in this area.

[13] The primary author's previous work in experimental philosophy with an AIBO robot is described in (Chrisley and Parthemore, 2007).

[14] "Autonomous" is in scare quotes because we are inclined to believe that "true" autonomy requires, at minimum, autopoiesis.

[15] For a good introduction to LSA, see (Landauer et al., 1998).

we do not wish to be the ones to deny him his rights. *What we have expressly not done* is offer any litmus test for moral agency, artefactual or otherwise; indeed, we have called into question the ethics of any such attempt. Instead, we have provided a theoretical structure and a practical research program for exploring the nature and limits of a particular agent's moral agency.

It is now time to return to the questions with which we opened our paper, and summarize our (provisional) answers.

***What makes us* us?** A necessary, though almost certainly not sufficient, aspect of what makes us "us" is our concept of "self" (i.e., self versus other/non-self/world). Without a *concept* of self, an agent is not "anyone" at all.

When is an agent a moral agent? An agent is a moral agent when it demonstrates not only the possession of certain key concepts but the ability, over an extended period of interactions with that agent, to employ those concepts appropriately: only then can an agent be held morally responsible for its actions.

In our view, the claims by many researchers in machine consciousness that they have *already achieved* "minimal consciousness" in their implementations (or even just theoretical models of implementations!) – a majority, it would seem, at a recent international symposium on machine consciousness – is, at the least, grossly premature. Indeed, it is far from clear what "minimally conscious" is even supposed to mean. Rather, we would modestly suggest, the value of machine consciousness research, and its most appropriate goal, is the better to explore what it means for *any* agent to be a conscious, or a conceptual – or, even more restrictedly, a moral – agent. The artefactual element is intriguing but, ultimately, a red herring.

References

Aisbett, J. and Gibbon, G. (1994). A tunable distance measure for coloured solid models. *Artificial Intelligence*, 65:143–164.

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

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

Ballargeon, R. (1987). Object permanence in 3 1/2- and 4 1/2-month-old infants. *Developmental Psychology*, 23(5):655–664.

Brentano, F. (1995). *Psychology from an Empirical Standpoint*. Routledge.

Bugnyar, T. (2011). Knower-guesser differentiation in ravens: Others' viewpoints matter. *Proceedings of the Royal Society B*, 278:634–640.

Bugnyar, T. and Heinrich, B. (2006). Pilfering ravens, corvus corax, adjust their behaviour to social context and identity of competitors. *Animal Cognition*, 9:369–376.

Bugnyar, T. and Kotrschal, K. (2002). Observational learning and the raiding of food caches in ravens, covus corax: Is it 'tactical' deception? *Animal Behavior*, 64:185–195.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.

Chrisley, R. and Parthemore, J. (2007). Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience. *Journal of Consciousness Studies*, 14(7):44–58.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press. Also available as ebook through Oxford University Press (http://www.oxfordscholarship.com).

Clayton, N. and Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, 395:272–274.

Damasio, A. (2000). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. Vintage.

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*, 60(3):441–458.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown.

Evans, G. (1982). *Varieties of Reference*. Clarendon Press. Edited by John McDowell.

Fodor, J. A. (1987). Why paramecia don't have mental representations. *Midwest Studies in Philosophy*, 10(1):3–23.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Clarendon Press, Oxford.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. Bradford Books. First published 2000.

Geuder, W. and Weisgerber, M. (2002). Verbs in conceptual space. In Katz, G., Reinhard, S., and Reuter, P., editors, *Proceedings of SuB6 (Sinn und Bedeutung)*, pages 69–84. University of Osnabrück.

Hoyle, F. (1959). *The Black Cloud*. New American Library.

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

Landauer, T. K., Folz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Marques, H. G., Newcombe, R., and Holland, O. (2007). Controlling an anthropomimetic robot: A preliminary investigation. *Lecture Notes in Computer Science*, 46-48:736–745.

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

McDowell, J. (1996). *Mind and World*. Harvard University Press, Cambridge, Massachusetts.

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

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

Parthemore, J. (2011a). *Concepts Enacted: Confronting the Obstacles and Paradoxes Inherent in Pursuing a Scientific Understanding of the Building Blocks of Human Thought*. PhD thesis, University of Sussex, Falmer, Brighton, UK.

Parthemore, J. (2011b). Of boundaries and metaphysical starting points: Why the extended mind cannot be so lightly dismissed. *Teorema*, 31. in press.

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

Pepperberg, I. (1999). *The Alex Studies*. Harvard University Press.

Piaget, J. (1954). *The Construction of Reality in the Child*. Basic Books, New York.

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

Raby, C., Alexis, D., Dickinson, A., and Clayton, N. (2007). Planning for the future by western scrub jays. *Nature*, 445:919–921.

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

Rosch, E. (1999). Principles of categorization. In Margolis, E. and Laurence, S., editors, *Concepts: Core Readings*, chapter 8, pages 189–206. MIT Press.

Ruiz-Primo, M. A. and Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33(6):pp. 569–600.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–458.

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

Whitby, B. (1996). The turing test: Ai's biggest blind alley? In Millican, P. and Clark, A., editors, *Machines and Thought: The Legacy of Alan Turing*, volume 1. Clarendon.

Whitby, B. (2003). The myth of AI failure (CSRP 568). FTP archive currently offline as of April 2011. University of Sussex (UK) Cognitive Science Research Papers (CSRP) series.

Zachar, P. (2000). *Psychological Concepts and Biological Psychiatry: A Philosophical Analysis*. John Benjamins Publishing Company.

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

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