Artificial Consciousness and Artificial Ethics:
Between Realism and Social-Relationism.

Steve Torrance
School of Engineering and Informatics, University of Sussex, Falmer, Brighton, BN1 9QJ, UK.
School of Psychology, Goldsmiths, University of London, New Cross, London, SE14 6NW, UK
Email: stevet@sussex.ac.uk

Revised 20 June 2013

Abstract. I compare a 'realist' with a 'social-relational' perspective on our judgments of the moral status of artificial agents (AAs). I develop a realist position according to which the moral status of a being - particularly in relation to moral patiency attribution - is closely bound up with that being's ability to experience states of conscious satisfaction or suffering (CSS). For a realist both moral status and experiential capacity are objective properties of agents. A social-relationist denies the existence of any such objective properties in the case of either moral status or consciousness, suggesting that the determination of such properties are rests solely upon social attribution or consensus. A wide variety of social interactions between us and various kinds of artificial agent will no doubt proliferate in future generations, and the social-relational view may well be right that the appearance of CSS-features in such artificial beings will make moral-role attribution socially prevalent in human-AA relations. But there is still the question of what actual CSS states a given AA is actually capable of undergoing, independently of the appearances. This is not just a matter of changes in the structure of social existence that seem inevitable as human-AA interaction becomes more prevalent. The social world is itself enabled and constrained by the physical world, and by the biological features of living social participants. 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 an objective scientific inquiry.

Keywords: realism, social-relationism, 'machine question', artificial agents, moral status attribution, consciousness-satisfaction-suffering (CSS); phenomenal-valuational holism; bio-machine spectrum

1 INTRODUCTION

The term 'society' is currently understood to include humans – with various non-human biological species (e.g. domestic creatures, primates, etc.) sometimes included for certain purposes. Humans may be soon joined on the planet by a number of new categories of agents with intelligence-levels that – according to certain measures – approach or even exceed those of humans. These include robots, software agents, bio-engineered organisms, humans and other natural creatures with artificial implants and prostheses, and so on. It may well become widely accepted (by us humans) that it is appropriate to expand the term 'society' to include many of such emerging artificial social beings, if their cognitive capacities and the nature of their interactions with humans are seen as being sufficiently fluent and complex to merit it. Of the many questions that may be asked of new members of such an expanded society, two important ones are: Are they conscious? What kind of moral status do they have? These questions form part of two new offshoot studies within AI – Machine Consciousness and Machine Ethics (or Artificial Consciousness/Ethics). Wallach et al (2011) have written that 'Machine ethics and machine consciousness are joined at the hip'.[1] In what follows

[1] Wallach *et al.* are primarily concerned in their paper with how modelling moral agency requires a proper theoretical treatment of conscious ethical decision-making , whereas the present paper is more broadly concerned with the problem of ethical consideration – that is: what kinds of machines, or artificial agents in general, merit ethical consideration either as agents or as patients. The discussion largely centres around the relation between experiential consciousness and the status of moral patiency. I've discussed the general relation between consciousness and ethics in an AI context in Torrance, 2008, 2011, 2012a,b; Torrance & Roche, 2011. While I sympathize strongly with the sentiment expressed in the above quote from Wallach *et al.,* I prefer the terms 'Artificial Consciousness' (AC), and 'Artificial Ethics' (AE) to the 'Machine' variants. It seems clear that many future agents at the highly bio-engineered end of the spectrum of possible artificial agents – particularly those with near-human levels of cognitive ability - will be strong candidates to

we will find several ways in which these two domains cross-fertilize. Most prominently in the present discussion is the fact that the attribution of consciousness to machines, or artificial agents more generally (AAs)[2], seems to be a fundamental consideration in assessing the ethical status of artificial social beings – both as moral agents and as moral patients.

So 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 these linked questions. One view may be called 'moral realism' (or 'realism' for short). For a realist, there are objectively correct answers to questions like: 'Is X a conscious creature or agent?'; 'Is X the kind of being that has moral value?' – although it may be impossible, even in principle, to provide assured or non-controversial answers to such questions. I will defend a variant of this view here. The other view may be called 'social relationism' (hereafter 'relationism' or SR for short).[3] Social relationists deny that questions such as the above have objective answers, instead claiming that their determination relies solely upon social conditions, so that the process of ascribing properties such as 'being conscious', 'having moral status', involves an implicit relation to the ascriber(s). Different versions of relationism are presented by Mark Coeckelbergh and David Gunkel in their two excellent recent books (Coeckelbergh 2012; Gunkel 2012; see also Coeckelbergh, 2009, 2010a, 2010b; Gunkel, 2007, 2013). Despite initial appearances, much of the disagreement between realism and SR can be removed; nevertheless, a considerable core of variance will remain. Questions about sentience or consciousness, on the one hand, and about moral status on the other, will provide two pivotal dilemmas for the members of any future expanded society. The issue between these two views will thus be likely to be of crucial importance for how social life is to be organized in future generations..

In the way that I will understand them, realism and SR are views *both* about ethical status and about sentience or consciousness. A question such as 'What moral status does X have (if any)?' – supposing X to be a machine, a human, a dolphin, or whatever - can be construed in a realist or in a relationist way. Equally, a question such as 'Is X conscious?' can also be given either a realist or a relationist construal.

2 THE DEBATE ILLUSTRATED AND DEVELOPED

In order to see the difference between the realist and the relationist approaches, consider a hypothetical future robot and its human owner. Let's say the robot works as a gardener for the household where it is domiciled. We can think of the garden-bot here as being a 'robot' along broadly conventional lines: that is, as a manufactured humanoid physical agent with a silicon-based brain. We will assume that the robot is capable of communicating and interacting with humans in a relatively rich way. Sadly, the human owner in our example regularly treats her charge in ways that cause the robot (in virtue of its design) to give convincing behavioural manifestations of distress, or even of pain, whenever its lawn-mowing, planting or weeding don't come up to scratch. Two natural questions might be: (Q1) 'Is the robot gardener *really* feeling distress or pain?' and (Q2) 'Is it really *ethically ok* for the owner to behave that way to the robot?' Q1 and Q2 seem to be linked – it might be natural to say that it would be wrong to treat the robot in that way *if and in so far as* it would cause it experiences of distress, etc.

It might be objected that these two questions are dogged by so many indeterminacies that it is impossible to give any clear meaning to either of them. On Q1: How are we to think of the 'pain' or 'distress' of the robot gardener? Can we imagine any such states in an electronic robot – however sophisticated its design and construction? Can this be done in any way that avoids anthropomorphizing in a fashion that obscures, rather than clarifies the issue? How could the similarities or differences be characterized, when comparing 'silicon-

be considered both as phenomenally conscious much in the way we are and as moral beings (both as moral agents and as moral patients). Yet it may be thought rather forced to call such artificial creatures 'machines', except in the stretched sense in which all natural organisms, us included, may be classed as machines.

[2] In what follows I will sometimes talk about 'robots' and sometimes about 'artificial agents' (AAs). Generally, I will mean, by 'robots' physical agents (possibly humanoid in character), which are constructed using something like current robotic technology – that is, whose control mechanisms are computer-based (or based on some future offshoot from present-day computational designs). By 'AAs' I will understand a larger class of agents, which includes 'robots' but which will also include various kinds of possible future bio-machine hybrids, plus also agents which, while synthetic or fabricated, may be partially or fully organic or metabolic in physical make-up.

[3] A source for the term 'social relationism' is the title of a paper by Mark Coeckelbergh (Coeckelbergh 2010a).

based pain' (were such to exist) with 'organic pain'? There seems to be such a conceptual chasm between the familiar cases and the robot case that it may be thought that Q1 simply cannot be properly posed, at least in any direct, simple way. And on Q2: how are we to think of ethical duties towards silicon-based creatures? As with Q1, it might be argued that the question of ethical treatment posed in Q2 cannot be given a clear construal when divorced from the context of inter-human or at least inter-organism relations.[4]

Despite these doubts, I believe a realist can insist that the issues raised in Q1 and Q2 are genuine concerns: for example one wants to know whether there is something that the robot's owner is *doing wrong to* the robot, in virtue of some real states that the robot is undergoing as a result of her actions. Do we, ethically, need to care about such actions, in the way we would about human distress or pain? Do we need to waste resources protecting such agents (of which there may be many) from 'ill-treatment'? Such questions may be hard to put into clear conceptual focus, let alone to answer definitively. But surely they do seem to address strong *prima facie* concerns that might be raised about the kind of example we are considering. In any case, for now we are simply putting forward the realist's position: a more considered appraisal will follow when we have examined both positions in greater detail.

On the SR position neither Q1 nor Q2 has an inherently right or wrong answer. Rather, answers will emerge from the ways in which society comes to develop beliefs and attitudes towards robots and other artificial agents. Indeed a social-relationist may well insist that the sense that can be made of these questions, let alone the answers given to them, is something which itself only emerges in time with social debate and action. Perhaps society will broadly adopt a consensus and perhaps it won't: but, for a supporter of SR, there is no way one can talk of the 'correct' answers to either question over and above the particular responses that emerge through socially accepted attitudes and patterns of behaviour.

A relationist, then, would deny that there is any 'correctness' dimension to either Q1 or Q2 over and above what kinds of options happen to emerge within different social settings. In contrast, for a realist, social consensus could really get things wrong – *both* on the psychological issue of whether the robot is feeling distress and pain *and* on the ethical issue of how the robot should be treated. (Realists don't have to link these two questions in this way, but it has been seen by many natural to do so, and that's the version of realism that will be in the forefront of the discussion here.) [5],[6]

So for a realist the following concerns can be raised. A future generation of robot-owners may believe that some or even most of them are conscious or sentient, and therefore deserving of our moral concern in various ways, when 'in fact' they are no more sentient than (present-day) sit-on mowers, and can suffer only in the sense in which a mower would 'suffer' if the wrong fuel mix was used in its engine. Or it might go the other way round – socially-prevalent attitudes may withhold attributions of conscious feeling and/or moral consideration to certain kinds of artificial agent who 'in fact' are conscious and merit moral consideration – and perhaps that the latter is the case for such agents largely, or at least partially, because of

[4] I am grateful to an anonymous reviewer for insisting on this point. In the present discussion I am limiting the kinds of cases under consideration to AAs whose design involves electronic technologies which are relatively easy to imagine, on the basis of the current state of the art and of fairly solid future projections. There are a wide variety of other kinds of artificial creature – including ones with various kinds of artificial organic makeup, plus bio-machine hybrids of different sorts – which expand considerably on this range of cases. We will consider this broader range of cases in later sections.

Concentrating at the present stage of the discussion on AAs like the robot gardener, and other such, relatively conservative cases, has a triple utility. First, it allows us to lay down the foundations for the argument without bringing in too many complexities for now. Second, many people (supporters of strong AI, or 'strong artificial consciousness') have asserted that such robots could well have genuinely conscious states (and thus qualify for serious ethical consideration) if constructed with the right (no doubt highly complex) functional designs. Third, such cases seem to offer a greater challenge than cases which are closer-to-biology: it's precisely the non-organic cases, in which one has detailed similarity to humanity in terms of behavior and functional organization but marked dissimilarity in terms of physical or somatic structure, where the issues seem to be raised particularly sharply.

[5] Some people might agree that Q1 should be construed in a realist way – what could be more real than a person's vivid experiences of distress, pleasure, etc.? – while being reluctant to treat Q2, and similar moral questions, in a realist or objectivist way. In this paper I am supporting a realist position for both experiential and moral attributions.

[6] For a defence of the view that there is a close association between questions of consciousness and those of moral status, see, for example, Levy, 2009. Versions of this view are defended in Torrance, 2012; Torrance & Roche 2011. The conclusions Levy comes to on the basis of these views are very different from mine, however.

the facts of the former kind that hold true of them. (The scare-quoted phrase 'in fact' in the above will, of course, be problematic on the SR view.)

According to realism, then, a question like 'Is X conscious?' is asking about objective[7] matters of fact concerning X's psychological state, and further (for this version of realism) attributions of moral status at least partially supervene on X's psychological properties (including consciousness and related states). Moreover, on the version of realism being considered here, this is true of humans and non-human biological species, just as much for (current or future) robots and other technological agents. So on this view normative moral questions concerning the actions or the treatment of humans, animals or robots, are closely bound up with factual questions concerning the capacity of these various agents for conscious experience – that is, the *phenomenal* consciousness of such agents, as opposed merely to their ability to *function cognitively* in a way a conscious being would.[8]

Why should questions about the phenomenal consciousness of beings link so closely to moral questions? A key reason that the realist can put forward is to do with the fact that conscious experiences of different sorts have characteristic positive or negative *affective valences or qualities*.[9] Think of the current field of your experiential awareness. Experiences sometimes come in an affectively neutral way, such as the tiles on the floor which are in the background of my present visual awareness. But others are evaluatively graded: the hum of the fan system that I can hear is mildly irritating; the lilt of voices in conversation outside my window is mildly pleasant. Plunging into a rather cold sea will generate a mix of sensations, of a characteristic hedonic character (a different mix for different people, situations, moods, etc.) In general the flow of our experience is closely tied to our desires, needs, aversions, plans, etc., as these unfold in lived time. The degrees of positive or negative affective valence can vary from scarcely noticeable to extreme, and, as they vary, so too do the contributions that they make to a conscious creature's levels of satisfaction and suffering, i.e. to their experienced well-being or ill-being. (It would seem to be difficult to see how beings which were not capable of conscious experience could have such states of satisfaction and suffering.) Further, it can be argued that considerations of how actions make a difference to the well-being, satisfaction, etc. of people affected by those actions is a central concern of ethics. So an important part of a being's moral status is determined by that being's capacity to experience such states of satisfaction/suffering.

The social-relational view, by contrast, 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 have instead 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. Society is fast-changing today, so new criteria for consciousness-attribution may currently be emerging, which are likely to radically alter social opinion on what beings to treat as conscious (indeed what beings count as 'social beings' will itself be a socially 'moving target'). Moreover, on the SR 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. The SR view will say there can be no neutral way of judging between these different views. Thus, on the relationist view, both the psychological and the ethical components of the realism described above are rejected.[10]

[7] To clarify: the realist's claim is that 'Is X conscious' is objective in the sense that 'X is currently conscious' asserts a historical fact about X, even though it's a fact about X's *subjective* state, unlike, say, 'X is currently at the summit of Everest'.

[8] The inherent distinguishability between phenomenal and functional consciousness is defended in Torrance, 2012.

[9] See Thompson, 2007, chapter 12, for a discussion of the relation between consciousness, affect and valence.

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

3 THE EXPANDING MORAL CIRCLE AND THE LANDSCAPE OF CONSCIOUS WELL-BEING

Many writers have talked of a progressive expansion of moral outlook through human pre-history and history. Peter Singer has written persuasively (Singer, 2011) of the ‘expanding circle’ of ethical concern, from primitive kin- 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. According to Singer, we owe our ethics to the evolutionary pressures on our pre-human forbears, and we owe it to the animal co-descendents of that evolutionary process to extend ethical concern to the well-being of all sentient creatures.[11]

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 - see, for example, Leopold, 1948, Naess, 1973.) 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 (Singer, 2011, p. 123.)

Singer’s position appears to rest on a conceptual claim – that only beings with sentience can coherently be considered as moral patients. As I see it his argument is roughly this. 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 a lawn-mower run continually on the wrong fuel mix, or operated over a rocky terrain, it may ‘suffer’ or be ‘abused’ in a functional, and non-experiential, sense). So, the argument goes, a non-sentient being cannot *coherently* be considered as a moral patient, because no action could affect its consciousness of its own well-being. Ethical considerations are about how our actions might affect others (and ourselves) in ways that make a difference to those so affected.. To quote from an earlier work of Singer’s (1975, 9): ‘A stone does not have interests because it cannot suffer. Nothing that we can do to it could possibly make any difference to its welfare. A mouse, on the other hand, does have an interest in not being kicked along the road, because it will suffer if it is.’ (cited in Gunkel 2012, p. 113.)

Of course the precise conditions under which particular artificial agents might be considered as conscious – as a being having interests, like the mouse, rather than as a brute, inanimate object, like the stone – are notoriously difficult to pin down. But having controversial verification-conditions is not the same as having no verification-conditions, or ones which are essentially dependent upon the social-relational context. Compare, for example, the question of extra-terrestrial life. There may be controversy among astrobiologists over the precise conditions under which life will be established to exist in a distant solar system; but this does not detract from the ontological objectivity of exoplanetary life as an real phenomenon in the universe. It does not make the physical universe, or the existence or nature of planets outside our solar system, *social-relational* (although of course exoplanetary science, astrobiology, etc., as academic studies, are social activities, with their specific funding, geopolitical and ideological dimensions). Similarly, the realist might say, a robot’s or AA’s pain, if such a thing were to exist, would be as objective a property of the AA, and as inalienable *from* the AA, as would your or my pain be inalienable from you or from me. (Conversely, an AA 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.)[12]

For realism, in the version I am developing here, 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). This *phenomenal-valuational* holism may be put as follows. 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

[11] See also the discussion in Torrance, 2013.

[12] Singer does not discuss the case of moral interests of robots or other artificial agents (or indeed of exoplanetary beings) in his 2011 book.

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.

Not all realists may agree with this kind of phenomenal-valuational holism. One writer who seems to do so is Sam Harris. He presents the view (Harris, 2010) 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 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, Harris takes brain-processes to be central determinants of well- or ill-being – a view I would accept only with strong qualification. It may be hard in practice to solve various moral dilemmas, but, he claims, they are in principle 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, *pace* Harris, bodily processes from the neck down are pretty important too, as well as the active engagement of the creature in its lived world.[13] A big challenge for AI and Artificial 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.[14]

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. I very much agree with Harris on this, and would say that it has been particularly true of 'Cognitive Science' as this has developed in the last half-century or so. The over-cognitivizing of the mind has had an unfortunate effect both on the attempt to understand mental processes scientifically, and on the process of exploring the ethical ramifications of mind science. AI researchers, neuroscientists, psychologists and philosophers have too often talked as though the mind was exclusively a cognitive mechanism. All that cognitivizing hard work; that systematic redescription of the *chiaroscuro* of our desires, emotions, pains and delights in terms of informational operations; that bleaching out of happiness and misery from the fabric of our psychology; all that has, arguably, been to the detriment of both scientific understanding and ethical discussion in this area.

4 THE TURING DREAM: A 400-YEAR SCENARIO

I thus applaud the science-ethics holism found in objectivist writers like Harris. Perhaps this enthusiasm will be shared by relationists such as Coeckelbergh and Gunkel. 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.[15]

[13] For an excellent, and fully elaborated, defence of the kind of view of consciousness that I would accept, which is centred around notions of enactivism and autopoiesis, see Thompson, 2007 – especially chapters 12 and 13. There is no space here to do more than gesture to this view in the present discussion.

[14] Like Singer, Harris does not consider the ethical status of possible artificial agents.

[15] 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?

In the light of this, consider the following possible future picture. Let us consider what might be called the 'Turing Dream' – that is, the goal aspired to by many in the AI community of developing the kind of complexity and subtlety in functioning which would enable robots to behave more or less like us over a very broad range of activities and competencies in the physical and social real world. Let us suppose (a big ask!) that researchers do not hit any insurmountable barriers of computational tractability, hardware speed, or other performance or design impasses, and the 'dream' is fulfilled, so that such robots proliferate in our world, and co-habit with us in a more or less peaceable kingdom. Let us suppose, then, that – say – 400 years from now (or choose the timeframe you prefer), human social relations have changed radically because of the existence of large numbers of such artificial agents implementing more-or-less-human or even greater-than-human levels of ability across a wide range of capabilities. If the Turing Dream were to come about in something like this fashion, 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. Even if they don't, and the 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 at all 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? (And by what right are we licensed to talk of admitting 'them' into 'our' moral world?) 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, judgments, norms and institutions that grow up as such artificial agents proliferate, to say whether they are *actually* conscious, whether they *actually* are capable of having states of felicity 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 roughly similar moral terms as the equivalent actions of humans, etc.[16]

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 – or even in ways that can't easily be ranked in terms of any serial comparison of 'level' to ours)? 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 convention to celebrate the 500th anniversary of Turing's birth?[17] 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'.[18]

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 relationist side of the argument in

[16] 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.

[17] The present paper originated as a contribution to a workshop at a Convention celebrating the 100th anniversary of Turing's birth.

[18] 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' (Bisson, 1991).

this robot convention, and that the most instinctive human response would be 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 a century from now would be right to adopt a realist view about the consciousness and ethical status of humans. Should we then not do the same today (or indeed in a hundred years' time) of robots? [19]

5 THE OTHER MINDS PROBLEM PROBLEMATIZED

A supporter of SR might raise 'other-minds'-style difficulties even about CSS in humans, as a way of highlighting the difficulty of certifying the existence of CSS properties in artificial agents.[20] Any individual human's recognition of CSS properties is based upon their own direct first-person experience. There are great psychological and social pressures on any individual to infer to the existence of such CSS properties in others, it might be said: yet the inference can never rationally be fully justified, since one can never be directly acquainted with another person's conscious states. It appears that one has to infer to the internal states of another's consciousness on analogy with bodily and physiological states observed in oneself as these are linked with one's own experience. As Wittgenstein remarked, with a whiff of sarcasm, 'how can I generalize the one case so irresponsibly?' (Wittgenstein, 1953: I, §293). The 'other minds' problem may thus be used as a device for undermining realism, by emphasizing that even in the case of other humans, let alone various species of animals, there are substantial difficulties in establishing the presence of the kinds of experiential properties that the realist requires for ethical attribution.[21]

There are several responses to such doubts. A first response is this. If CSS properties are not completely mysterious and inexplicable, from a scientific point of view, then they must be causally grounded in *natural features* of any individual human possessing them. Imagine two humans, A and B, who display third-person behavioural and physiological properties that were identical in all relevant respects, yet where A has an 'inner' phenomenological life and B lacks all phenomenological experience. This would be a completely bizarre phenomenon, to be explained either in terms of some non-natural circumstance (e.g. God chose to inject a phenomenology into A but withhold it from the physically-identical B) – or else it would have to be completely inexplicable. Neither of these alternatives seems attractive. Yet to entertain radical doubts about other minds seems to require embracing one or other of these unsavoury positions. To avoid that, it would be necessary to concede that there are some sort of third-person, underpinning natural features, publicly accessible in principle, that could be used to distinguish those humans (and other species) with a phenomenology, with CSS features, from those without. In any case, there is a mass of scientific theory and accreted experimental data linking CSS properties in humans, and affective properties more generally, with our evolutionary history, and with our current biological and neural make-up.

Doubts about solipsism and the existence of other human consciousnesses besides one's own, can also be shown to be fed by a series of questionable fundamental conceptions about mind and consciousness in the first place. Solipsistic doubts are linked to doubts about the 'Hard Problem' of consciousness (Chalmers, 1995; Shear, 1997); the Explanatory Gap between physiology and experience (Levine, 1983); Absent Qualia puzzles (Block, 1978); and so on. I have suggested (Torrance, 2007) that there are two contrasting conceptions of phenomenality: the 'thin' (or 'shallow') conception, deriving from Descartes' radical doubts about bodily existence in the face of the *cogito*, assumes that consciousness must be understood in terms of a radically ego-logical sensory presence, which is causally, ontologically and conceptually, radically separate

[19] 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.

[20] See, for example, Gunkel, 2012, Chapters 1 and 2, who insists on the perennial philosophical problem of 'other minds' as a reason for casting doubts on rational schemes of ethical extension which might enlarge the sphere of moral agency or patiency to animals of different types, and beyond that, to machines. It is remarkable how frequently Gunkel returns to rehearsing the theme of solipsistic doubt in his discussion.

[21] Appeal to doubts over other minds is one of the arguments used by Turing to buttress his early defence of the possibility of thinking, and indeed conscious, machines (Turing, 1950).

from any other processes (particularly bodily or physiological processes). By contrast, the 'thick' (or 'deep') conception sees consciousness as conceptually inseparable from the lived, physiological processes that make up a conscious being's embodied existence. The distinction is deployed in that paper to articulate an embodied approach to Machine Consciousness (see also Holland, 2007, Stuart, 2007, Ziemke, 2007). In a later paper I have argued for an embodied, de-solipsized conception of consciousness, by challenging prevailing 'myths' which construe consciousness as being, *by definition,* essentially inner, hidden and individualistic (Torrance, 2009). I suggest that consciousness is an 'essentially contested' notion (the term is due to Gallie, 1955) – so that, at the very least, there is no requirement that consciousness be conceptualized in the ways suggested by these 'myths'.

Freed from the necessity to rely on internalist and individuocentric conceptions of consciousness, one is able to see how philosophical worries to do with solipsism, absent *qualia*, the explanatory gap between the neural and the phenomenal, and so on, are all systematic products of a dependence by many sectors of the philosophical and scientific community (a dependence which both feeds from and into unreflective idioms of folk talk about mind) upon an outmoded and unhelpful way of conceptualizing experience or phenomenality. Many other authors have articulated versions of this richer, more deeply embodied, conception of phenomenality. A notable critique of the mind-body problem by Hanna and Thompson differerentiates between two conceptions of body – the body of the physical organism which is the subject of biological investigation as a physical-causal system (*Körper*) and the 'lived body' (*Leib*), which is the individual subject of embodied experience of an organism as it makes sense of its trajectory in its world (Hanna & Thompson, 2003). Hanna and Thompson deploy the *Leib-Körper* distinction to dissolve doubts over mind-body separation, other minds, and so on, by bifurcating notions of body, in contrast to the distinction between notions of phenomenality in Torrance, 2007.[22] Other philosophical challenges to the other minds problem will be found in Gallagher's rejection of theory-theory and simulation-theory approaches to mentalization, which draw upon work by Husserl, Scheler, Gurwitsch, Trevarthen and others who argue that, for example, the perception of someone's sadness is not an inference to a hidden X behind the expressive countenance, but is rather a direct observation of the other's experiential state, derived from the condition of primary intersubjectivity with their conspecific carers that humans and other primates are born into, and the secondary intersubjectivity of joint-attention and engagement in joint projects in the world that occurs during later development (Gallagher, 2005a, 2005b; Zahavi, 2001; Gallagher & Zahavi, 2008; Thompson, 2007.). Such work by Gallagher and others offers a strong alternative to opposed viewpoints in debates over the psychology of social cognition, but *a fortiori,* marginalize classical other-minds doubts as a philosophical side-show.

6 SOCIAL AND NON-SOCIAL SHAPERS OF SOCIALITY

We have seen, then, that to raise epistemological doubts which problematize our everyday assurance that we live in a community of consciousnesses is to rely on a cluster of problematic Cartesian assumptions which are based on options about how to conceptualize phenomenality which we are not forced to make, given alternative conceptual directions which many strong theoretical considerations direct us towards.[23] But in any case such doubts can be shown to be *irrelevant* from an ethical point of view. Ethics, while being based on a weighty body of theoretical discussion, stretching back millennia, is, above all, concerned with our *practical,* functioning, social lives. In practice, we live in a non-solipistic world, a world which we cannot but live in as one of a community of co-phenomenality.

If Ethics is about practice, it is also about sociality and interactivity. The way I come to recognize and articulate CSS properties in myself is partly based upon my social interactions with my conspecifics. Indeed, to deploy a serious point behind the Wittgenstein quip cited above (and much else in his later writings), my whole discourse about mind, consciousness, pain, desires, emotions, etc. is based upon the public forms of

[22] See also the treatment of this issue in Thompson, 2007, especially chapter 8 – therein called the 'Body-Body problem'. Thompson mentions that a quite elaborate range of notions contrasting and combining the motifs of *Leib* and *Körper* are found in Husserl's writings (see Depraz, 1997, 2001, cited in Thompson, 2007, ch. 8, footnote 6). Thompson also critiques superficial or 'thin' conceptions of phenomenology (*ibid.,* ch. 8) but without the 'thin'/'thick' terminology used in Torrance, 2007.

[23] A variety of sources, from phenomenology, and several of the mind sciences, all converging on the view that our understanding of mind is thoroughly intersubjective, in a way that renders solipsistic doubts incoherent, will be found in Thompson, ed, 2001, and Thompson, 2007.

life I share with other humans in the real world of participatory inter-subjectivity. 'Just try, in a real case', Wittgenstein wrote, 'to doubt someone's fear and pain.' (Wittgenstein, 1953, I, §203)

Ironically, we here seem to find ourselves using an SR-style argument to counter an SR-style objection to realism. But we need to be careful. The 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 the objectivity of first-personal recognition of CSS properties. Such a suggestion needs to be taken seriously, and may point to a deep truth behind the SR position – that our understanding of how we come to possess CSS properties, and of the variety of roles they play in our lives, is indeed inextricably bound up with our social relationships and activities.

But it's also 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 from these very experiential and appetitive features in our lives, including, centrally, the ups of happiness and downs of misery. It is vital not to assume that everything in human lived experience is subject to social shaping: the reverse is also true. 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 interactions with other bodies and the surrounding physical world. So there is a *multi-directional* cluster of mutually constitutive and constraining relationships between the social, material, biological and experiential factors in our lives.[24] What makes up our CSS features emerges from the entanglement of these various kinds of factors. This brings us to the heart of the question of CSS in future artificial social agents.

The progress of current work in AI teaches us that many of the features of human-human social and communicative interaction – the 'outer' features, at least – can be replicated via techniques in computer and robotic science – essentially algorithmic modelling techniques. 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 is gradually being extended. But also, our very conceptions of the social, the cultural and the intersubjective, are being re-engineered. Or, to put the point a different way: how we make sense of the social, and indeed how we make sense of 'we', is being continually reshaped by our artefacts as they are increasingly implicated in our social existence. This is a crucial point that the 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 SR account is well suited to provide a theoretical framework for much of the domain of the social. So what is taken to constitute 'the social' is itself largely shaped by social factors, and changes as new social possibilities emerge. But, as we suggested earlier, the domain of the social is 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-AA and AA-AA 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, from the multiple physical, environmental, and indeed metabolic and experiential drivers of sociality. This is particularly true of those social relationships between humans and the computationally organized machines of today (let alone future, more organically constituted, AAs). Arguably, there are a great many social relationships, even today, 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 'BIO/MACHINE' SPECTRUM

So it is likely that, 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

[24] 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.!

(despite many vociferous assertions to the contrary). Such artificial social partners are not, in Tom Regan's phrase, 'subjects of a life' (Regan, 1983). Thus, for now, the possibilities for social interaction with machines outstrip the possibility of 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*. 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? For one thing, techniques in synthetic biology may develop in ways which allow biotechnologists to create agents that are not just functionally or behaviourally very close to humans – i.e. which exemplify social patterns in a variety of outward ways, but which are close also 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 may indeed have little ground for denying that they are 'objectively' conscious, CSS-bearing, beings, with all the ethical consequences that would flow.

We certainly cannot rule out, then, that there will at some future time be AAs with physiologies which are as close to those of humans as one cares to imagine. Leaving aside the general 'other minds' objections discussed earlier, what reason would we have for saying that, despite our extensive biological commonality with such artificial creatures, they lacked the CSS features that we had – other than the (surely irrelevant?) fact that, unlike us, they were fabricated (or cultured) in a laboratory? (Such AAs might even have ontogenetic histories very much like ours, progressing from a foetal or at least neonatal stage, through infancy, and so on.) So, in the context of our present discussion, such very full-spec synthetic-biology-based, humanoid creatures surely give support to the idea that *at least some artificial agents should certainly be given the moral status that humans enjoy*. But such creatures occupy one extreme, bio-realistic, end, of a spectrum of possible artificial agents. At the other extreme there are the relatively simplistic, computational AI agents of the recent past and of today – an era when artificial agent design is still, one assumes, in its infancy.[25] And between those two extremes are a host of other imaginable and not-so-imaginable agent designs. Clearly, as we retreat from the bio-realistic end, judgment calls about CSS properties and ethical status on the basis of physical design are much less straightforward.

In many (middle) 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 the underlying detailed physical design is remote from detailed humanoid or mammalian physical design. These will offer the toughest cases for decision: agents that, via their fluent social capacities (and, in many varieties, outwardly human-like bodily features), display a wide variety of apparent CSS-evincing behaviours but where they share relatively few of the internal neurological and more broadly physiological features that make for the presence of CSS properties in humans. These are the cases where the social-relational view may seem to be on its most solid ground: what possible 'objective' basis could there be for deciding on whether to accord or withhold moral consideration in each particular class of example?

However, a realist can reply that, even if such cases are difficult to determine in practice, there is still the question of what kind of experiential state, if any, actually occurs, independently of human or social attribution. For such cases, then, there are many 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 societal implications: (A) 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 behaviourally convincing but CSS-negative artificial agents. (B) Conversely, in the false-negative case, vast

[25] For the sake of simplicity of discussion I'm here representing the spectrum as if it were a unidimensional space, whereas it is almost certainly more appropriate to see it as multidimensional. (See Sloman, 1984.)

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.[26]

Given the existence of a large variety of different possible artificial agents, what David Gunkel (2007, 2012, 2013) calls '*The* Machine Question' (my emphasis) factors into a multiplicity of different questions, each one centred around a particular kind of machine (or AA) design. Some kind of decision has to be made for each kind of machine design, and this is not going to be easy. Nevertheless surely it's not the case that the only option is to throw up one's hands in resignation and just say "Let social consensus decide".

8 NEGOTIATING THE SPECTRUM

So the bio-machine spectrum ranges from cases of very close bio-commonality at one end, to simplistic current-AI style behaviour matching at the other end. At both these extremes decisions about CSS capacity and about moral status seem relatively straightforward – a resounding 'yes' at the first extreme, and a resounding 'no' at the other. But what about the intermediate cases? Is there any way to make progress on providing methods for adjudicating on moral status for the wide variety of hard cases between the two extremes? I think that there is: I believe it is possible to propose a series of testable conditions which can be applied to candidate artificial agents occupying different positions in the vast middle territory of the spectrum. I list such a series of conditions below, well aware that this is only one person's draft, no doubt skewed by the preoccupations and prejudices of its author. Nevertheless it may point the way to showing how a realist position may be able to do more than simply say 'There must be a correct answer to the question "Should A be accorded moral status?" for any given candidate agent A', and may be able to offer something approaching a decision-procedure.[27] Here, then, is the list:

- NEURAL CONDITION: Having a neural structure (albeit not implemented in a conventional, biological way), which maps closely the features specified by theories of the neural (and sub-neural?) correlates of consciousness, according to the best neuroscience of the day.
- METABOLIC CONDITION: Replicating (perhaps only in some analogical form) the more broadly physiological correlates of consciousness in humans and other organisms (including, e.g. blood circulation, muscular activities; alimentary processes; immune-system responses; endocrine/hormonal processes, etc., to the extent that these are variously considered to be essentially correlated with biologically occurring forms of consciousness).
- ORGANIC CONDITION: Having a mechanism that supports consciousness that is organic in the sense of displaying self-maintaining/recreating (autopoietic) processes which require specific energy exchanges between the agent and its environment, and internal metabolic processes to support the continuation of these mechanisms and the maintenance of effective boundary conditions with respect to the surrounding environment.
- DEVELOPMENTAL CONDITION: Having a life-history that at least approximates to that of humans and other altricial creatures, with foetal, neonate, infant, etc. stages – involving embodied exploration, learning, primitive intersubjective relations with carers and teachers, and the capability to develop more complex forms of intersubjectivity.

[26] Here we are stressing moral patiency, but a similar problem of false-positives/false-negatives exists for moral agency, too. Many relatively primitive kinds of AI agents will act in a functionally autonomous fashion so as to affect human well-being in many different ways – so in one sense the question of moral agency is much more pressing, as many authors have pointed out (see, for example, Wallach and Allen, 2010). Yet there are important questions of responsibility-ascription that need to be determined. If we provide too great a share of responsibility to AAs that act in ways that are detrimental to human interests, this may well mask the degree of responsibility that should be borne by particular humans in such situations (e.g. those who design, commission and use such AAs). Conversely, we may overattribute responsibility to humans in such situations and withhold moral credit from artificial agents when in truth it is due to such agents, for example, on the grounds that as 'mere machines' they cannot be treated as fully responsible moral agents. The vexed issue of moral responsibility in the case of autonomous lethal battlefield robots provides one illustration of this area: see Sparrow, 2007; Arkin, 2009; Sharkey & Suchman, 2013.

[27] It may well be that realism will not be defeated if no decision-procedure is provided. There is the ontological matter of whether questions like "Does A have a moral status as an ethical patient/agent?" have a correct answer or not (independently of the accidents of social determination). And there is the epistemological or methodological matter of whether or not it can be determined, in a straightforward way or even only with extreme difficulty, what the correct answer to that question is for any particular A.

- SENSORIMOTOR CONDITION: Exemplifying forms of sensorimotor interaction with the environment, which are considered to be implicated in conscious perceptual activity (See O'Regan & Noe, 2001; O'Regan, 2007.)
- COGNITIVE CONDITION. Displaying the various cognitive marks or accompaniments of consciousness, as identified by cognitive theories of consciousness.
- SOCIAL CONDITION: Generally interacting with humans (and other artificial agents) in a variety of social situations in a fluent and collaborative way. (See Gallagher, 2012).
- AFFECTIVE/WELFARE CONDITION: Showing evidence of having an extended array of needs, desires, aversions, emotions, somewhat akin to those shown by humans and other natural creatures.
- ETHICAL CONDITION: Subscribing to ethical commitments (in an autonomous, self-reflective way, rather than simply via programmed rules) which recognize appropriate moral rights of people with needs, desires, etc.
- INTROSPECTIVE CONDITION: Being able to articulate an introspective recognition of their own consciousness; ability to pass a variety of self-report tests held under rigorous experimental conditions.
- TURING CONDITION: Responding positively, and in a robust and persistent way, to Turing-Test style probes for consciousness (including blind dialogues filtered through an textual or other filtering interface; or physical-interaction episodes in a real-world context).

There is, and no doubt will continue to be, much controversy as to which items should be on this list, on their detailed formulation and corroboration-conditions, and on their relative priorities. Those who favour an organic approach to consciousness will privilege the first few criteria, whereas those who favour a cognitively-based view will put more emphasis on many of the later conditions. Nevertheless, despite the controversial nature of some of the items, I believe that a list like this could be drawn up, for use across the scientific community, to establish a broad way forward to assess different candidate artificial agent designs, in order to assist in decisions about presence of consciousness in those different candidates, and consequently their ethical status.

A list of this sort does not establish the realist position on questions about when CSS properties are genuinely present, and consequently about which kinds of possible artificial agents might qualify for inclusion in the 'circle' of moral consideration. Nevertheless it at least shows that 'the machine question' has an internal complexity, a fine texture, with different categories pointing to different kinds of answers, rather than simply being a single undifferentiated mystery from an ontological and epistemological point of view, leaving the vagaries of social judgment as the sole tribunal. Moreover it suggests that this particular issue is not really different in nature from any complex scientific question where there are many subtle and cross-cutting considerations to be taken into account, and where disagreements are open to resolution, at least in principle, in a rational way.

9 SUMMING UP

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-AA and AA-AA interactions. But sociality is itself constrained in various ways by physical, biological and psychological factors. And consciousness and well/ill-being (what I have called CSS) 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 or constraints on 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 classes of artificial creatures will have closely similar biological properties, making the question of CSS-attribution relatively easy for those 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, a conception of which we currently have at best the merest glimmer, if any at all) 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 and functional cognitive/affective organization of such agents will play important roles (Coeckelbergh, 2009, 2010b), of course, in determining how we attribute moral patiency and agency status, but only in a wider mix of considerations which will include many 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 example, 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, false positive or false negative mis-attributions could each 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 CSS states. 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.[28] 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 (Kurzweil, 2005) 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, were the rapid and far-reaching technosocial upheavals predicted by many to come about. Nevertheless I believe I have shown that the realist framework offers the best way forward for the AI and AC research community to 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. I am also grateful to Joanna Bryson and David Gunkel for inviting me to join with them in co-chairing the Turing Centenary workshop on The Machine Question, where this paper first saw life. Ideas in the present paper have also greatly benefitted from discussions with Mark Bishop, Rob Clowes, Ron Chrisley, Madeline Drake, Shaun Gallagher, David Gunkel, Joel Parthemore, Denis Roche, Wendell Wallach and Blay Whitby.

REFERENCES

Arkin, R.C. (2009). *Governing Lethal Behavior in Autonomous Systems.* Boca Raton: Chapman & Hall/CRC.

Bisson, T. (1991). 'They're made out of meat', *Omni,* 4. April, 1991. http://www.eastoftheweb.com/short-stories/UBooks/TheyMade.shtml Accessed 10 January 2013

Block, N. (1978). 'Troubles with functionalism', in Savage, C. ed. *Perception and Cognition: Issues in the Foundations of Psychology. Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press. 261-325.

Chalmers, D. (1995) 'Facing up to the problem of consciousness.' *Journal of Consciousness Studies* 2(3): 200-19.

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

[28] For example, here I have dealt primarily with the connection between consciousness and artificial moral patiency, or recipiency, as opposed to moral agency, or productivity (but see footnote 26 above). There are arguments that suggest that consciousness may be as crucial to the former as to the latter (Torrance, 2008; Torrance & Roche, 2011).

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

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

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

Depraz, N. (1997). *La traduction de* Leib, *une* crux phaenomenologica. *Etudes Phénoménologiques* 3.

Depraz, N. (2001). *Lucidité du corps. De l'empiricisme transcendental en phénoménologie*. Dordrecht: Kluwer Academic Publishers.

Hanna, R. and Thompson, E. (2003). 'The mind-body-body problem.' *Theoria et Historia Scientiarum: International Journal for Interdisciplinary Studies* 7: 24-44.

Gallie, W. B. (1955). 'Essentially contested concepts'. *Proceedings of the aristotelian society* Vol. 56:. 167-198.

Gallagher, S. (2005a). *How the body shapes the mind*. Oxford: Clarendon Press.

Gallagher, S. (2005b). 'Phenomenological contributions to a theory of social cognition'. *Husserl studies*, 21(2), 95-110.

Gallagher, S. (2012) 'You, I, Robot' *AI and Society,* DOI 10.1007/s00146-012-0420-4

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

Gunkel, D. (2007). *Thinking Otherwise: Philosophy, Communication, Technology..* West Lafayette, IN: Purdue University Press.

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

Gunkel, D. (2013) 'A vindication of the rights of robots', this Volume.

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

Holland, O. (2007). 'A strongly embodied approach to machine consciousness.' *Journal of Consciousness Studies*, *14*(7), 97-110.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Viking.

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

Levine, J. (1983). 'Materialism and qualia: The explanatory gap.' *Pacific Philosophical Quarterly* 64: 354-361.

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

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

O'Regan, J. K., & Noë, A. (2001). 'A sensorimotor account of vision and visual consciousness.' *Behavioral and brain sciences*, 24(5), 939-972.

O'Regan, J. (2007). 'How to build consciousness into a robot: the sensorimotor approach'. In Lungarella, M, et al. (eds.) *50 years of artificial intelligence*, Heidelberg: Springer Verlag, 332-346.

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

Sharkey, N. and Suchman, L. (2013). 'Wishful mnemonics and autonomous killing machines', *AISB Quarterly*, 136, 14-22.

Shear, J., ed. (1997) *Explaining Consciousness: The Hard Problem.* Cambridge MA: MIT Press.

Singer, P. (1975) *Animal Liberation: A new ethics for our treatment of animals.* NY: New York Review of Books.

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

Sloman, A. (1984). 'The structure of the space of possible minds'. In Torrance, S. (ed). *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence.* Chichester, Sussex: Ellis Horwood, 35-42.

Sparrow, R. (2007). 'Killer Robots', Journal of Applied Philosophy, 24(1), 62-77

Stuart, S. (2007). 'Machine consciousness: Cognitive and kinaesthetic imagination.' *Journal of Consciousness Studies*, 14(7), 141-153.

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

Thompson, E. (2007) *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge, MA: Harvard University Press.

E. Thompson, ed. (2001) *Between Ourselves: Second-Person Issues in the Study of Consciousness*, Thorverton, UK: Imprint Academic. Also published in *Journal of Consciousness Studies* (2001) 8(5-7).

Torrance, S. (2007). 'Two conceptions of machine phenomenality', Journal of Consciousness Studies, 14 (7). 154-166.

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

Torrance, S. (2009). 'Contesting the concept of consciousness.' *Journal of Consciousness Studies*, 16(5), 111-126.

Torrance, S. (2012). 'Super-intelligence and (super-)consciousness' *International Journal of Machine Consciousness,* 4(2):

Torrance, S. (2013) 'Artificial agents and the expanding ethical circle' *AI & Society,* DOI: 10.1007/s00146-012-0422-2

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

Turing, A. (1950). 'Computing machinery and intelligence'. *Mind* 59: 433-460.

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

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford: Blackwell.

Zahavi, D. (2001). 'Beyond empathy. Phenomenological approaches to intersubjectivity.' *Journal of Consciousness Studies*, 8(5-7), 5-7.

Ziemke, T. (2007). 'The embodied self: Theories, hunches and robot models.' *Journal of Consciousness Studies*, *14*(7), 167-179.