The Routledge Handbook of Philosophy of Pain; edited by Jennifer Corns
Format: Royal (174 × 246 mm); Style: Handbook_1; Font: Bembo;
Dir: //integrafs5/kcg/2-Pagination/TandF/RHPP_RAPS/ApplicationFiles/
9781138823181_text.3d;
Created: 13/12/2016 @ 17:43:05

16
ROBOT PAIN

Pete Mandik

1 Introduction: suffering robots

What if the painful but necessary experiments that are conducted on animals and people could instead be conducted on elaborate robots? This would be an ethical boon, but only if the robotic surrogates weren't so elaborate as to themselves suffer real pain. Aside from the potential benefits to medical research, there might be other motives – some benign, some nefarious – for creating robots with the potential of themselves suffering pains. It might increase the usefulness of a robot servant if it took care to prevent damage to itself; and the resultant self-monitoring system may turn out to implement pain. Some humans may seek to purchase pain-feeling robots for the purpose of torturing them – a sad fact about some humans. Plausibly, there's an ethical imperative for making sure avoidable robot pains are not inflicted.[1] But there's a metaphysical question of whether such pains – robot pains – *could* be inflicted. Further, there's an epistemological question of how we would ever know. As our technologies advance, this special version of the problem of other minds – the problem of robot pain – becomes increasingly pressing.

Can robots feel pain? I intend this question as shorthand for a much more elaborate question, and the sequence of elaborations I intend proceed along two lines, one of which goes beyond "robots" and the other of which goes beyond "feel pain."

More than robots, I want to ask about a broader class of entities, a class that includes technological artifacts – such as artificially intelligent computers (AI), and computer simulations of human and non-human animals – including so-called "mind uploads," simulations based directly on brain scans (for more specifically on the feasibility of mind uploading, see Mandik 2015a). Unless it matters for some particular point, I will continue to use "robot" as a handy label for the sort of entity in question, whose resemblance to humans is sufficiently incomplete as to leave interesting and special questions open about whether they feel pain.

About the phrase "feel pain," there are two sorts of pain phenomena or aspects of pain one might indicate by use of that phrase, but only one of those sorts is my primary focus in this chapter and only one of those is what I primarily aim to pick out by "feel pain." To be very brief, we might sort these aspects into the first-person aspects and the third-person aspects, and it is the first-person aspects that I primarily intend to pick out by "feeling pain" and related.

The bifurcation I have in mind can be conveyed in terms of the classic other-minds problem. The third-person aspects of pain are those aspects of pain upon which I base my confidence that someone else is in pain, whereas the first-person aspects of pain are the ones which I might, in a certain philosophical frame of mind, find myself doubting that anyone but me ever feels. These first-person aspects of pain are my primary focus here. (See also, in the present volume, Chapters 17, 18, and 20.) To elaborate my description of them even further, they are pains as they appear to the one who is having them. They are consciously experienced pains. There may be even further elaborations needed in nailing down the target here, but I will address them as further needs arise. It should suffice for now to say that the question "Can robots feel pain?" is shorthand for my real question, "Can robots and their ilk (e.g., computer simulations) consciously suffer pains, pains of the sort that I have little doubt that I myself suffer when I appraise my own pains from my first-person point of view?"

Echoing Chalmers' (1996) taxonomy of sorting problems of consciousness into the "easy" and the "hard, " we might say that the question of present concern is far harder than the question of whether there could be a robot that, to all outward appearances, was in pain. When we turn to the hard question of robot pain, the question of whether there can ever be pains that feel "from the inside" to robots the way pains feel to us, we encounter a question that does not seem straightforwardly empirical. It may very well ultimately turn out to be an empirically decidable question, but it seems that further philosophical reflection is needed before we know how to proceed empirically here.[2]

To my mind, it seems that the strongest and most convincing lines of thought relevant to the question of robot pain are two arguments, one of which offers a negative answer and the other of which offers a positive one. The first argument, the one whose conclusion is that robots cannot feel pain, is an adaptation of John Searle's famous Chinese room argument against artificial intelligence (Searle 1980). Searle develops his own argument in terms of the mental state of understanding, though it seems to me to be readily adaptable to pain states. The second argument, the one concluding that robots can feel pain, is an adaptation of a prosthetic neurons argument, perhaps most well known in the work of Chalmers (1996), but earlier versions of which may be found at least as far back as Harman (1973: 38–39).[3]

In the next section, I'll further explore the Chinese room argument against robot pain. I'll turn then in the section after that to explore the prosthetic neurons argument in favor of robot pain. Finally, in the concluding section I'll offer a comparative assessment of the two arguments.

At the heart of my assessment is a comparison between premises in each argument that relate the ways things seem with respect to conscious pain phenomena to conscious pain phenomena themselves. The respective premises differ in their logical structure – one is the logical converse of the other. Putting this very briefly: At the core of the argument for robot pain is a premise that moves from the ways thing seem first-personally to the ways they are, whereas at the core of the argument against robot pain is a premise that moves in the opposite direction, that is, from the ways things are to the ways they first-personally seem. This logical difference, I'll suggest, makes a difference in the respective likelihoods of the soundness of each argument.

2 Chinese rooms and the case against robot pain

At the heart of Searle's famous argument is the thought experiment that gives the argument its name. We are invited to imagine Searle inside of a room running a program – that is, following a set of instructions – that would result in observers outside of the room concluding that someone inside the room understands Chinese.

The program in question is structurally equivalent to a program that, when run on a digital computer, would allow someone to have a convincing conversation in Chinese with it, perhaps via an exchange of text messages. In the room scenario, observers outside of the room send and receive messages via printed cards going into and out of slots in the room's walls. Inside of the room is John Searle, who is stipulated to understand absolutely no Chinese himself. He understands English, and follows a set of instructions printed in English in a manual that also contains pictures of a variety of Chinese symbols. The English text does not provide translations from Chinese into English, but instead directs Searle to select appropriate output cards for each given input card. We can imagine the instruc tions having something like the general form of "If receiving symbol ABC, then reply with symbol XYZ," which would be an instruction that a monolingual English reader can follow well enough without having any clue what the appropriate translation of ABC and XYZ into English would be.

We can see at this point the relevance of Searle's thought experiment for AI: If Searle can run the program without himself understanding Chinese, then an AI or robot that gives the outward appearance of understanding Chinese might nonetheless be running a program without itself thereby understanding Chinese.

We are in a position, too, to see how to adapt the Chinese room argument to pertain to pain. The program in question can be one the running of which convinces outside interlocutors that they are conversing with an entity undergoing some degree of pain. And if we imagine the room in the head of a large robot, instead of just card outputs that say things in Chinese about being in pain, the robot's outputs might additionally involve moving its body in ways consistent with, for instance, suffering the pains associated with a sprained ankle. I'll postpone further discussion of pain for now, and return to the version of the thought experiment focused on understanding a language.

This thought experiment is not itself an argument, but instead serves as a part of an argument. To briefly indicate some of the main parts of the larger argument, we can make do for now with the following points. First, we need to suppose a version of the thesis of AI – the thesis that robots, computers, and their ilk can have genuine mental states – that is stated in terms of the running of programs. We might put the crucial point like this: If AI is possible, then something or someone could understand Chinese merely by running a program. We can put essentially the same point in a different, contrapositive form. The following should serve our purpose: If it is possible for something or someone to run any arbitrarily selected program without themselves *thereby* understanding Chinese, then AI is not possible. The "thereby" here is important – a Chinese/English bilingual person running Searle's program suffices for Chinese understanding, but they don't understand Chinese *because* (or *in virtue of*) of the running of the program.

The role of the Chinese room thought experiment is to establish the truth of the claim that it is indeed possible for something or someone to run any arbitrarily selected program without thereby understanding Chinese. It should be clear that, given the above formulation of what AI entails about programs, if the thought experiment succeeds, then AI fails.

However, the thought experiment does not succeed. As even Searle himself notes (Searle 1980: 419–42), one of the most natural responses available to the defender of AI is the now famous "systems reply": While the defenders of AI can readily grant that Searle himself does not understand Chinese, they need not grant that Searle himself is running the program. Instead, says the systems reply, Searle is a proper part of a larger system – a system including other contents of the room, including the cards and instruction manual – and it is this larger system that runs the program. It's open, then, for the AI defender to assert that the larger system understands Chinese, and Searle's stipulated ignorance is powerless to cast doubt on this.

Searle introduces a second thought experiment to shore up his argument against the threat posed by the systems reply. In this second thought experiment, he dispenses with the room and the other external props – the cards, the instruction manual. We are invited now to imagine Searle memorizing the program, including all of the pictures of the symbols. We might even imagine Searle memorizing sounds and instructions for which sounds to offer in reply to other sounds. Hypothetically, at least, Searle could run the program at such a speed that an outside observer/interlocutor might think Searle himself could both speak and write Chinese. But, again, we are invited to imagine Searle running this program without himself actually understanding Chinese. The systems reply now seems impotent, for it is not a larger system that is running the program, but Searle himself. And if he also does not understand Chinese, then this would be a case in which running a program does not thereby give rise to understanding.

We can imagine adapting Searle's thought experiment to pain states by imaging Searle undergoing a procedure – involving perhaps local anesthetics, analgesics, or some combination thereof – so that he cannot feel pain, while nonetheless remaining awake and alert. Numb to pain, he might nonetheless, by memorizing a set of instructions and observing by sight, etc., various pokes and prods to his flesh, give a convincing performance that would enable him to pass a Turing test for being in pain. He would be running a program that would, if run by a robot, give every outward aspect of being in pain. But he wouldn't thereby be in pain by running the program. Or so we are invited to imagine. If Searle can run the program without thereby being in pain, there seems little reason to believe, then, that a robot running what's essentially the same program would thereby feel pain.

But the question we must now contemplate is this: Is it indeed the case that Searle can run the program without thereby being in pain? What reason is there to believe that this is the case? What reason is there for not instead believing that, in running the program, Searle must thereby indeed be in pain?

To get a feel for what the underlying reasoning must be here, it helps to contrast the Searlean argument with some arguments that are decidedly *not* the Searlean argument. One argument that is certainly not the present argument is one where the opponent of AI presents an actual modern-day computer and declares it to be evident that it fails to be in pain. They would be correct in their claim that it is evident that computer is pain-free. Most AI proponents would likely grant that current computers exhibit none of what we take to be the third-person accessible evidence for the presence of pain states. This would be a very weak argument against the possibility of robot pain, and I think that the Searlean argument is better than that.

Another argument that is decidedly not the Searlean argument is a version of Leibniz's *Monadology* argument for the simplicity of minds. Leibniz imagines shrinking down and examining a complex mechanism – he seems clearly to have a brain in mind – and discovers nothing therein that would explain perception. We can easily modify the Leibnizian argument to involve a similar failure to explain pain. A shrunken explorer might make an incredible journey through the entire course of a brain and remain puzzled as to whether this complex system must give rise to any felt pain. Note that what our miniaturized Leibniz accesses are third-personal aspects of pain. Despite being physically inside of a brain, there remain senses of "inside" and "outside" whereby Leibniz remains outside. For all that he observes, the possibility remains that there is something it's like to be the complex mechanism, and the complex mechanism itself feels what it's like "from the inside."

The improvement that the Searlean argument offers over these other two arguments is the way it attempts to access things "from the inside" in the relevant senses of the terms. We can see the Searlean argument as attempting to marshal first-person evidence in the service of his anti-AI conclusion. In conducting the Searlean thought experiment ourselves we must imaginatively inhabit a point of view that puts us in a position to access first-personal evidence. But what is this so-called first-person evidence? It can be nothing else besides its seeming to the person in question that they are not in pain – I have no idea what it could mean for someone to have first-person evidence that they are not in pain without it seeming to that person that they are not in pain. And from this first-person evidence about how things seem, the Searlean attempts to draw a conclusion about how things are, namely that one is not in pain.

We are in a position now to lay out the main logical structure of the full Searlean argument.

1 If robot pain is possible, then there ought to be some program in virtue of which one would be in pain solely by running that program.
2 Searle can run any arbitrarily selected program without it seeming to him from the first-person point of view that he is in pain.
3 If it doesn't seem to one from the first-person point of view that one is in pain, then one is not in pain.
4 It's true of every possible program that Searle could run it without himself thereby being in pain.Therefore, robot pain is not possible.

The conclusion follows straightforwardly from premises (1) and (4). Premise (1) seems an obvious entailment of the basic idea of robot pain. Premise (4) follows straightforwardly from premises (2) and (3). Premise (2) is the core idea of the Searlean thought experiment, especially the one designed to avoid the systems response. Premise (3), which I'll hereafter refer to as the *Searlean conditional*, is the one that strikes me as the most questionable part of the whole argument. But I'll postpone my critical remarks for the section after the next. It is time now to turn to the case in favor of believing in robot pain.

3 Prosthetic neurons and the case for robot pains

If robots could indeed feel pain, how might we go about building such a robot? One approach that suggests itself is that we identify systems we know to already feel pain – namely ourselves – and we copy as much as we can from the human case into a robotic form. One way to construct a robotic copy of a human is by gradually transforming a human into a robot by a sequence of prosthetic replacements of the human's naturally occurring parts, especially parts of the nervous system, with artificial analogs.

Like all physiological systems in the human body, the nervous system is composed of causally interacting cells. The most significant cells in the nervous system are neurons. The causal interactions between the cells together serve in the causal mediation between sensory stimulus to, and behavioral response of, the entire organism. The interactions between cells also serve to constitute those aspects of cognition that can be characterized causally. One such aspect is memory, which can be characterized in terms of changes in an organism that underwrite its ability to give different responses to instances of the same type of stimulus presented at different times (as when a ringing bell at one time doesn't cause salivation but when presented at a later time does). Another aspect of cognition that can be characterized causally is the ability to make discriminations, as when one discriminates a paint chip's color from the color of the background. Part of such a capacity must involve the different causal effects that the respective colors of the chip and background have on the discriminating organism.

Hypothetically, the causal influence that one naturally occurring cell exerts on another can also be exerted by a device that is not a naturally occurring cell, but is instead an artificial prostheses, perhaps a microchip that has the same input–output profile as a neuron. Imagine a sequence of surgeries that transforms a human into a robot by gradual replacement of components. We will make the simplifying assumption that all that matters for present discussion is located in the nervous system, and so we will imagine the sequence of prosthetic replacements as involving a sequence of procedures whereby, during each procedure, a single cell of the nervous system is replaced by a chip that has the same causal effects and sensitivities vis-à-vis other cells as the cell that the chip replaces. Note that we are not supposing that the chip has *all* of the same causal properties as a neuron. If it did, it would be impossible to distinguish by any experiment or observation a neuron from a chip. But chips are made largely of silicon, and neurons are not, and the chemical differences involve causal differences. Nonetheless, we are supposing that a neuron is not causally sensitive to all of the causal properties of another neuron. And we are further supposing that the limited range of causal properties that characterize interneuron interaction can be fully replicated by chip–neuron interaction, which in turn can be fully replicated by chip–chip interaction. The result of transforming NaturalPete into RobotPete by a sequence of chip replacements will be that RobotPete is what I'll call a "coarse-grained causal isomorph" of NaturalPete.

NaturalPete feels pain. A natural thing to suppose goes hand-in-hand with the very idea of robot pain is that, given that NaturalPete feels pain, in being NaturalPete's coarse-grained causal isomorph, RobotPete feels pain too. (We assume throughout that when each is awake, they are each in the same environment in all relevant respects.) For the purposes of the central argument in this section, the most important connection between the thesis of robot pain and the neural prosthesis scenario can be stated thusly: If, by being the coarse-grained causal isomorph to an entity that feels pain, the isomorph feels pain, then robot pain is possible.

The thought experiment serves the larger argument in favor of robot pain in the following manner. We are invited to imagine ourselves undergoing the sequence of prostheses replacements without it ever, at any point, thereby ceasing to seem to ourselves that we have pain. The "thereby" here is important: It may be prudent to administer a general anesthetic for the duration of the surgery, and thus any pain would, arguably, cease for the duration. Nonetheless, it isn't here supposed that the pain is temporarily abated *because* one or more neurons were replaced by chips, but instead because an anesthetic was administered. At every point in the sequence at which the subject is neither anesthetized nor asleep, it seems to the subject that they are in pain despite how many of their neurons have been replaced by chips.

Important in thinking through the thought experiment of the transformation of NaturalPete into RobotPete is the sequence $n + 0$, $n + 1$, $n + 2$, …, where $n + m$ equals the number of neurons so far replaced at that point by prostheses, $n + 0$ corresponds to NaturalPete, and $n + z$ corresponds to RobotPete. Throughout this sequence, there is something that uncontroversially remains constant, namely the coarse-grained functional structure. Each respective entity in the sequence, then, is a coarse-grained functional isomorph of its sequential predecessor.

But why is it important to the argument that there's a gradual transformation of a human into a robot? It helps here to contrast the present argument with one that depends only on the consideration of a robot that has the same coarse-grained functional structure as a human. In this latter case, all we are imaginatively presented with, being humans ourselves, is third-person evidence concerning whether the robot is in pain. It exhibits all the outward behavioral evidence that we would utilize in determining that some human other than ourselves is in pain. And minus the very fine details that distinguish neurons from microchips, it has the same third-person-accessible internal causal structure as well. In contrast, the advantage of the neural prosthesis argument is the way in which it marshals the first-person point of view. We imaginatively put ourselves in the position of NaturalPete, and imagine ourselves living through the sequence of surgeries and post-operative occasions. Throughout the sequence, what we are invited to imagine is that it would seem to us, from the inside, as it were, that our pains neither fade nor suddenly disappear. For the thought experiment to serve its role in the larger argument, we need additionally a linking principle from how things *seem* with respect to one's own pain states to how things actually *are* with regard to one's pain. We need, then, a premise in the argument along the lines of this: If it seems to one that one is in pain, then one is indeed in pain.

Assembling the pieces sketched so far, we are in a position now to consider the structure of the larger prosthetic neurons argument, noting a rough structural similarity between the adapted Chinese room argument.

1 If, by being the coarse-grained causal isomorph to an entity that feels pain, the isomorph feels pain, then robot pain is possible.
2 Via a sequence of neuro-prosthetic replacements, you can be transformed into your coarse-grained causal isomorph without it ever ceasing to seem to you from the first-person point of view that you are in pain.
3 If it seems to one from the first-person point of view that one is in pain, then one is in pain.
4 By being your coarse-grained causal isomorph, an entity thereby feels pain.Therefore, robot pain is possible.

The conclusion of this argument follows straightforwardly from premises (1) and (4). Premise (1) is highly natural thing to suppose about the relation between the idea of robot pain and the notion of a being who is a coarse-grained causal isomorph of a normal human being. Premise (4) follows straightforwardly from premises (2) and (3). Premise (2) is a condensed statement of the prosthetic neurons thought experiment. Premise (3) is the one that strikes me as the one most in need of further comment, and I turn to that issue in the next section, where I'll also remark on a counterpart premise from the Searlean argument.

4 Concluding comparative assessment

The above arguments cannot both be sound, for their conclusions cannot both be true. One or both arguments must be unsound. Unfortunately a full discussion would far exceed the allotted space. In this remaining section, I give a very brief account of one possible view of both arguments, one whereby the third premise of the adapted Chinese room argument is false and the third premise of the prosthetic neurons argument is true. Of course, this falls far short of establishing the soundness of the one argument and unsoundness of the other. I hope the present discussion to nonetheless be useful for further thought on the matter.

Before remarking on the merits of premise (3) of the prosthetic neurons argument, it is worth noting its logical relationship to its counterpart premise from the Searlean argument. Recall that the third premise from the adapted Chinese room argument is this:

> *The Searlean conditional*: If it doesn't seem to one from the first-person point of view that one is in pain, then one is not in pain.

The Searlean conditional is logically equivalent, by contraposition, to this conditional (which I'll hereafter call *the thesis of self-intimation*):

> *The thesis of self-intimation:* If one is in pain, then it seems to one from the first-person point of view that one is in pain.

The thesis of self-intimation is logically distinct from its converse, a conditional we can call *the thesis of incorrigibility*:

> *The thesis of incorrigibility:* If it seems to one from the first-person point of view that one is in pain, then one is in pain.

Note that the thesis of incorrigibility is one and the same as the third premise of the prosthetic neurons argument. The remainder of my commentary on the two arguments will focus on these two theses, *self-intimation* and *incorrigibility*.[4]

The merits of the thesis of self-intimation can be regarded as one and the same as the merits of the claim that pains never occur unconsciously. Unconscious pains, if there are such things, are pains one is in without it seeming to one that one is in pain. See David Pereplyotchik's Chapter 17 in the present volume for an overview of the case for unconscious pains. As Pereplyotchik points out, one basis for acknowledging unconscious pain is a version of a Higher-Order representational account of consciousness, in particular the Higher-Order Thought (HOT) account as spelled out by Rosenthal (2005, 2011).

Being very brief here: according to Rosenthal's HOT theory, one and the same pain can be unconscious at one time, and conscious at another. The relevant difference between the two different times boils down to whether or not there's a suitable accompanying HOT about the pain state. Putting it in a very simplified manner, when there's a HOT about the pain state, the pain state is conscious, and when there is no such HOT, the pain state is unconscious. So, on this view, there are two ways of being in pain: one way is to be in pain consciously, and the other is to be in pain unconsciously.

The view additionally allows for two different ways of consciously being in pain. The first is the way just mentioned above, a way involving two actual mental states, one of which is the pain, the other of which is a HOT about the pain. A second way one may count as being in a conscious pain state is when one is simply in a HOT about a pain. The pain the HOT is about need not actually exist; it is, in this case, merely notional. There are then, three ways of being in pain: one unconscious way and two conscious ways, and the two conscious ways are when the state the HOT is about is actual and when the state the HOT is about is merely notional (Rosenthal 2011: 433–434).

This may seem an unintuitive way of reading HOT theory. One might prefer to read it instead as entailing that one is in a conscious pain only when one has both of two states: a pain and a HOT about it. On such a reading, in the case where one has only the HOT, one is not in pain even though it seems to oneself that one has a pain. And thus, on such a reading, *incorrigibility* would turn out false.[5]

Whatever the merits of this latter reading of HOT theory, it goes against the one that Rosenthal emphasizes (as well as the interpretations favored by other HOT theorists, see, for example, Weisberg 2010, 2011). One way of unpacking these distinct readings of the HOT

The Routledge Handbook of Philosophy of Pain; edited by Jennifer Corns
Format: Royal (174 × 246 mm); Style: Handbook_1; Font: Bembo;
Dir: //integrafs5/kcg/2-Pagination/TandF/RHPP_RAPS/ApplicationFiles/
9781138823181_text.3d;
Created: 13/12/2016 @ 17:43:07

theory is, as I spell out in further detail in Mandik (2015b: especially 194–195), as a non-relational reading of HOT theory, as opposed to a relational one. In favor of this non-relational reading of HOT theory, Rosenthal writes "[a]ll that matters for a state's being conscious is its seeming subjectively to one that one is in that state. On the HOT theory, that's determined by a HOT's intentional content" (Rosenthal 2011: 436). On this non-relational reading of HOT theory, the HOT alone suffices for being in a conscious state and thus is consciousness non-relational, for it does not consist in a relation borne between a HOT and another state.

The upshot of all this for the present chapter is that, insofar as Rosenthal's HOT theory and the non-relational reading of it are coherent, we thereby have a coherent basis for rejecting the thesis of self-intimation while at the same time affirming the thesis of incorrigibility. HOT theory allows us to reject self-intimation on the grounds that there can indeed be pains without accompanying HOTs about them (see again Pereplyotchik's chapter for details). HOT theory allows us to affirm incorrigibility on the grounds that, when there is HOT to the effect that one is in pain, one is in pain, at least in a notional sense of "is in pain."

A converse move seems unavailable to the Searlean. The Searlean seems to lack any basis for rejecting incorrigibility while at the same time affirming self-intimation. The Searlean lacks, as far as I can tell, any basis for saying that while it is true that if one is in pain, then it seems from the first-person point of view that one is in pain, it is nonetheless false that if it first-personally seems to one that one is in pain, then one is. The Searlean may try to bolster their case along such lines by arguing that all pains are conscious pains, but now the dialectic shifts to the one covered in Pereplyotchik's chapter, and space does not permit further pursuing it here. (See also Chapter 15, this volume.)

Insofar as we are focusing critique only upon the third premise of each of the two main arguments concerning robot pain (the premises that comprise the theses of *incorrigibility* and *self-intimation*), we now have a basis for rejecting the Chinese room argument, while retaining the prosthetic neurons argument. Of course, the arguments contain other premises, and while I have tried to indicate somewhat why I think those other premises are the most secure parts of the arguments, other thinkers may focus their critiques on exactly those points. Such lines of thought cannot be further pursued in the present space.

In this chapter, I have laid out what seem to me to be the most promising arguments on opposing sides of the question of whether what humans regard as the first-person accessible aspects of pain could also be implemented in robots. I have emphasized the ways in which the thought experiments in the respective arguments attempt to marshal hypothetical first-person accessible evidence concerning how one's own mental life appears to oneself. In the Chinese room argument, a crucial premise involves the thesis that from a lack of it seeming that one is in pain, one can conclude that one is not in pain. There's a counterpart thesis playing a crucial role in the prosthetic neurons argument, one asserting an entailment from its seeming that one is in pain to one's being in pain. I further suggested that by adopting a HOT theory of consciousness, of the sort reviewed in Pereplyotchik's chapter in the present volume, one is thereby in a better position to endorse the prosthetic neurons argument over the Chinese room argument than to make the opposite appraisal.

Related topics

Chapter 15: The lives of others: pain in non-human animals (Droege)
Chapter 17: Pain and consciousness (Pereplyotchik)

The Routledge Handbook of Philosophy of Pain; edited by Jennifer Corns
Format: Royal (174 × 246 mm); Style: Handbook_1; Font: Bembo;
Dir: //integrafs5/kcg/2-Pagination/TandF/RHPP_RAPS/ApplicationFiles/9781138823181_text.3d;
Created: 13/12/2016 @ 17:43:07

Chapter 18: Pain: perception or introspection? (Aydede)
Chapter 20: Pain and incorrigibility (Langland-Hassan)

Acknowledgements

For especially helpful discussions of earlier versions I am enormously grateful to Jennifer Corns and David Pereplyotchik. Thanks too are due for comments from David Bain and Olivier Massin, and also from attendees of presentations I made at The Role of Phenomenal Consciousness Workshop at the University of Glasgow's Value of Suffering Project.

Notes

1 For an excellent survey on the distinctively ethical dimensions of robot pain, see Sandberg 2014.
2 The problem under present consideration is a less general version of what Schneider and Mandik (forthcoming) identify as the Hard Problem of AI Consciousness.
3 See also Pylyshyn 1980, Zuboff 1981, 2008, 1994, and Cuda 1985.
4 See Chapter 20, this volume, for more on incorrigibility.
5 I am grateful for Jennifer Corns for pressing this concern in response to an earlier version of the present material.

References

Chalmers, D.J. (1996) The conscious mind. In *Search of a Fundamental Theory*. Oxford: Oxford University Press.

Cuda, T. (1985) Against neural chauvinism. *Philosophical Studies* 48: 111–127.

Harman, G. (1973) *Thought*. Princeton, NJ: Princeton University Press.

Mandik, P. (2015a) Metaphysical daring as a posthuman survival strategy. *Midwest Studies in Philosophy* 39 (1): 144–157.

Mandik, P. (2015b) Conscious-state anti-realism. InC. Muñoz-Suárez and F. de Brigard (eds.), *Content and Consciousness Revisited: With Replies by Daniel Dennett*. Berlin: Springer.

Pylyshyn, Z. (1980) The "causal power" of machines. *Behavioral and Brain Sciences* 3: 442–444.

Rosenthal, D.M. (2005) *Consciousness and Mind*. Oxford: Clarendon Press.

Rosenthal, D.M. (2011) Exaggerated reports: reply to Block. *Analysis* 71(3): 431–437.

Sandberg, A. (2014) Being nice to software animals and babies. R. Blackford and D. Broderick (eds.), *Intelligence Unbound: The Future of Uploaded and Machine Minds*. Hoboken, NJ: Wiley Blackwell.

Schneider, S. and Mandik, P. (Forthcoming) How philosophy of mind can shape the future. In Amy Kind (ed.), *Philosophy of Mind in the Twentieth and Twenty-First Centuries*. London: Routledge.

Searle, J.R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3: 417–457.

Weisberg, J. (2010) Misrepresenting consciousness. *Philosophical Studies* 154(3): 409–433. doi:10.1007/s11098-010-9567-3.

Weisberg, J. (2011) Abusing the notion of what-it's-like-ness: a response to Block. *Analysis* 71(3): 438–443. doi:10.1093/analys/anr040.

Zuboff, A. (1981) The story of a brain. In D.R. Hofstadter and D.C. Dennett (eds.), *The Mind's I*. New York: Basic Books.

Zuboff, A. (1994) What is a mind? *Midwest Studies in Philosophy* 19(1): 183–205.

Zuboff, A. (2008) Thoughts about a solution to the mind–body problem. *Think* 6(17–18): 159–171.