Cruelty to Robots?
The Hard Problem of Robot Suffering

Bruce MacLennan[1]
[1] University of Tennessee, Knoxville
maclennan@utk.edu

Abstract

Ethical treatment of future robots will be contingent on their capacity to suffer, for example, to feel pain in the same sense that we feel pain. Neurophenomenology, which illuminates the relation between human experience of emotion and its neural correlates, posits empirical questions whose solutions define the preconditions for robots' experience of their emotions.

1 Emotional Robots

In this paper I focus on just one issue in the ethical treatment of robots, namely the question of cruelty to robots in the context of *the Hard Problem of robot emotions*: the possibility and preconditions for a robot experiencing its emotions (MacLennan, 2009). Everyday notions of ethical treatment depend in part on the recipient's capacity to suffer. Suffering includes pain, but goes beyond it, to include feelings of distress, agony, sorrow, anguish, and loss. From the opposite perspective ethical treatment also involves the capacity to experience joy or well being, but that is outside the scope of this paper. Future robots' capacity to feel will affect not only our treatment them, but also their treatment of us. For we expect robots to treat us well, but that is more likely if they can empathize with our feelings of suffering and joy. This capacity is more compelling if it goes beyond intellectualized empathy to empathetic feeling (such as we have via mirror neurons). But why should we equip robots with emotions at all?

For the purposes of this paper, Rolls (2007) provides a good summary of the essential characteristics of emotion: an emotion is a state elicited by the delivery or omission of a reward or punisher (either present or remembered), which functions as positive or negative reinforcement. Thus the organism acts to seek rewards and avoid punishers, and from an evolutionary perspective these actions are adaptive, in the sense of inclusive fitness (Plutchik, 2003, pp. 218–23). Rolls (2005, 2007) lists the principal functions (evolutionary adaptations) fulfilled by emotion, and most of them are also relevant to robotics (MacLennan, 2009). Certainly, many robots will not need emotions, but animals depend on emotions for efficient, real-time behavior, and for the analogous reasons we can expect them to be valuable in some autonomous robots.

Nevertheless, there would seem to be no reason why robots could not have all the appropriate information structures and control processes to fulfill the functions of emotions, but without *feeling* them. That is, there would seem to be no contradiction in "zombie robots," who, for example, have an internal representation corresponding to pain, and react as though in pain in appropriate circumstance, but which feel no pain. Regardless of whether one thinks this possibility is likely or unlikely, it remains in the realm of opinion unless we can find some principled, and preferably scientific, approach to the Hard Problem of robot emotions.

Therefore, our ethical treatment of robots will depend upon whether they have the capacity to feel pain and to suffer in other ways. Moreover, we might want some sociable robots to feel their emotions because not doing so could have dehumanizing consequences for them and us. For example, if future

robots simulate feeling with great verisimilitude but we believe that they don't feel anything, then we may unconsciously transfer our callousness to humans or other animals (as early vivisectionists ignored the apparent agony of their victims in the belief that they were "just machines"). In other words we might unconsciously discount external evidence of internal subjective states. Conversely, human ethical treatment — especially in the moment — is enhanced by our vicarious experiencing of another's feelings, especially their pain or suffering. A merely intellectual understanding may be much less motivating; indeed, the incapacity to feel another's emotions is a disability. We might expect the same to be the case for advanced robots, who would be less likely to treat us kindly if they cannot "feel our pain." Be that as it may, our ethical relationship to robots with synthetic emotions will depend on whether or not they can feel their emotions.

2 Neurophenomenological Approach

In this paper I will apply a neurophenomenological analysis to the Hard Problem of robot suffering and, more generally, robot emotions. This approach takes for granted the practical irreducibility of phenomenal consciousness to physical processes, whether this irreducibility is epistemologically necessary (e.g., Chalmers, 1996, Pt. 2; MacLennan, 1995, 1996; Strawson, 1994, 2006) or a consequence of the conceptual limitations of contemporary theories. The method involves parallel reductions in the neurological and phenomenological domains, in which observations and investigations in each domain inform those in the other.

Phenomenological reduction is based on the obvious fact that our conscious experience is structured in subjective time, space, and quality. Therefore conscious experience is both qualitatively and quantitatively reducible. *Qualitative reduction* separates conscious experience into phenomena of different kinds, such as perceptions associated with different sensory modalities, but also on the basis of qualities such as waking, dreaming, imagination, recollection, anticipation, and desire. However, we must beware of naive qualitative analyses, and careful experimental phenomenology informed by neuroscience is required for a correct analysis (Ihde, 1986; McCall, 1983).

Quantitative reduction analyzes a phenomenon in terms of smaller phenomena of the same kind. The simplest example is provided by visual phenomena, which can be reduced to smaller visual phenomena, and eventually to localized patches of color and oriented edges and textures. Similarly, proprioceptive and haptic phenomena can be reduced to simpler phenomena localized to patches of skin, joints, muscles, etc. These phenomenological reductions are supported neurophenomenologically by our knowledge of the receptive fields of neurons in the primary sensory cortices.

Although the parallel neurophenomenological reduction is easiest to understand in sensory phenomena, neurophenomenological analysis cannot stop there, but must consider the neural correlates of all aspects of conscious experience. This is progressively more difficult as we move inward from the periphery of the nervous system. Therefore, in many cases a more definite analysis awaits progress in neuroscience, but also corresponding experiments in experimental phenomenology. Nevertheless, just as we expect the overall function of the brain to be explained in terms of more elementary neural processes, so also we expect the macroscopic structure and dynamics of consciousness to be explained in terms of more elementary phenomenological processes.

Although we may disagree about exactly where it lies, there is an end to neurological reduction. The individual neuron is the obvious candidate for a functional unit, but some might argue for a larger unit such as a minicolumn, while others advocate a smaller unit such as the synapse, or matching neurotransmitter and receptor molecules. In any case we hypothesize a corresponding unit of phenomenological analysis, the *protophenomenon*. The idea is that just as the interaction of vast numbers of neural units constitutes the macroscopic neurodynamics of the brain, so the interaction of vast numbers of protophenomena constitutes conscious experience. The parallel reductions of

neurophenomenology suggest that protophenomena correspond to the elementary objects and processes of neurological theory. As such, protophenomena are *theoretical entities* in neurophenomenology, just as atoms were when first hypothesized in chemistry.

While protophenomena are the hypothetical constituents of conscious experience, in most cases we are not conscious of them. This paradox can be resolved by analogy. Objects are composed of atoms interacting so that their behavior is coordinated, but atoms are not *objects* in the colloquial sense; we might call them "proto-objects." (Objects, in the ordinary sense, may be hard or soft, warm or cool, for example, but individual atoms do not have these properties.) Likewise, a phenomenon in conscious experience arises from the coordinated behavior of its constituent protophenomena. Also, as an atom can be added to or removed from an object without changing the object qua object, so also a protophenomenon can be added to or deleted from a conscious phenomenon without changing the phenomenon qua phenomenon. This conclusion is supported by neurophenomenological analysis, for we would not ordinarily notice the contribution of a single neuron to our conscious state.

Although it is an empirical question, we do not know at this time what specific neural processes correspond to protophenomena, and so we call them *activity sites*.

The protophenomena may be considered the elementary degrees of freedom — or dimensions — of a conscious state. Consideration of sensory neurons suggests that their level of activity is correlated with the degree of presence in the conscious state of the corresponding protophenomenon, which is related to the receptive field of the neuron. We refer to this degree of presence as the *intensity* of the protophenomenon. Therefore the conscious state is coextensive with the intensities of its constituent protophenomena, which are correlated with neurodynamical activity in the corresponding activity sites. Depending on what the activity sites turn out to be, protophenomenal intensity might vary continuously with a variable such as membrane potential, or might make momentary contributions to the conscious state, if the corresponding event is the generation of an action potential or the binding of a neurotransmitter molecule to a receptor.

Whatever the activity sites may turn out to be, it is apparent that it is their interconnection and consequent interdependent activity that governs the macroscopic neurodynamics of the brain. Therefore the parallel neurophenomenological reduction implies that protophenomenal interdependencies constrain their dynamics and lead to the coherent changes of protophenomenal intensity that constitute a phenomenon.

Cortex has a common architecture across the cerebrum, in particular in the sensory areas. This suggests that the neurons that serve vision, for example, are not essentially different from those serving hearing. Recent experiments support this conclusion and imply that it is the connections among neurons that determine whether they are supporting visual or auditory qualities (e.g., Sur, 2004). The parallel conclusion in the subjective domain is that the qualitative character of protophenomena is determined by their interdependencies. That is, qualia arise from protophenomenal interdependencies. (Isolated protophenomena are not qualia per se.)

Space limitations preclude addressing criticisms of neurophenomenology and the notion of protophenomena, but they are discussed elsewhere (MacLennan, 1995–2010).

3 Neurophenomenology of Human Emotion

A neurophenomenological analysis of emotion begins with a phenomenology of emotion, that is, with an investigation of the structure of emotional experience, which has proved to be difficult despite many attempts (Plutchik, 2003, pp. 3–17, 64–7; MacLennan, 2009). There is also a large body of research on the neurophysiology of emotions in the context of evolutionary biology (e.g., Panksepp, 2004; Plutchik, 2003; Prinz, 2006), which has the advantages of not being restricted to human

emotion and of considering the adaptive functions of emotion, both of which are important to the issue of robot emotion.

It might seem axiomatic that we feel our emotions, but it is important to distinguish emotion and feeling (Damasio, 1999, pp. 42–9). The primary emotional response takes place in the limbic system and is unconscious, whereas conscious emotional experience is confined to cortical areas. William James (1884) and Carl Lange (1885) independently made the surprising claim that the foundation of emotional experience is sensation of prior bodily change. Although there have been objections, an increasing body of neuropsychological data supports various modifications and extensions of this theory (e.g., Damasio, 1994, 1999; Prinz, 2006). Prinz (2006, ch. 9) suggests that a three-level emotional processing hierarchy underlies emotional consciousness, a view that is quite compatible with protophenomenal analysis. At the lowest level are neurons with small receptive fields responding to local conditions in skeletal muscles, visceral organs, hormone levels, etc.; these correspond to emotional protophenomena. At the intermediate level neurons integrate the protophenomena into coherent patterns of activity, that is, into emotional phenomena. These seem to be similar to the first- and second-order maps described by Damasio (1999, ch. 6). At the third level these patterns are classified and specific emotions are recognized (Prinz, 2006, p. 214). Prinz's hierarchy can also be compared to Damasio's (1999, ch. 9) three-level hierarchy of emotion, feeling, and *feeling* feeling: "*an emotion*, *the feeling of that emotion*, and *knowing that we have a feeling of that emotion*" (Damasio, 1999, p. 8, italics in original). Thus, like other conscious phenomena, the qualitative character of an emotional phenomenon consists in the interdependencies among its constituent protophenomena.

Therefore, to obtain a comprehensive explanation of the structure of conscious emotional experience, we need to investigate the representation and integration of information in specific cortical areas, especially those aspects related to emotional response, and to correlate these neuropsychological investigations with phenomenological investigations into the structure of conscious emotional experience. By these means we will be able to identify the neuronal processes correlated with emotional protophenomena, and the interdependencies among them that define the qualitative structure of felt emotion. This is, of course, an ongoing and long-term research program, but neurophenomenological research into human emotional experience already provides a basis for understanding the determinants of the phenomenology of robot emotion.

4 Empirical Issues in Robot Emotions

Based on the forgoing analysis of conscious and unconscious emotional response in humans, we can address the problem of conscious emotional response in robots in a more focused way. Protophenomena are elementary subjective degrees of freedom, which correspond to activity sites in the brain, so that physical processes at these sites are correlated with the intensity of the corresponding protophenomena in conscious experience. This correlation is established empirically and treated as a brute fact of nature, so protophenomenal theory is a kind of dual-aspect monism (Atmanspacher, 2012).

Is it possible that physical processes in a robot's "brain" (central information processor) could constitute activity sites with associated emotional protophenomena? Since the only direct evidence of phenomenal consciousness that we have is our own, human consciousness, the issue is whether the robot's information processing devices are sufficiently similar to the human brain's *in the relevant physical ways*. The goal is to determine empirically the sorts of physical properties that are invariably associated with protophenomena in brains, from which we may infer that these physical properties will be associated with protophenomena in other systems.

Unfortunately, at this stage of the scientific investigation of consciousness we cannot say what physical properties of material systems are sufficient for them to be activity sites and to support protophenomena. Nevertheless, the question is empirical, since it can be addressed by controlling physical quantities and substances in individual neurons and observing (through introspection) their effects on conscious experience. The technology for conducting these experiments is improving (e.g., Petit et al., 1997; Losonczy, Makara, and Magee, 2008; Service, 2013). Therefore we anticipate that it is just a matter of time before we have a better understanding of the essential physical properties of activity sites. In the meantime we may entertain several hypotheses.

First, the activity sites could be the somatic membranes of neurons and protophenomenal intensity might correspond to membrane potential relative to its resting potential. Before we could decide whether similar activity sites could be constructed in an artificial system, we would need to have a more detailed understanding of the relation of membrane potential to protophenomenal intensity. The questions are empirical, and their answers will constrain the sorts of artificial physical devices that could support protophenomena.

Cook (2000, 2002, chs. 6–7, 2008) argues that the intensity of a protophenomenon correlates with the flux of ions across the cell membrane when the ion channels open during an action potential, and the protophenomenon is in effect the cell's sensing of its (intercellular) environment. If this is correct, then the essential properties of an activity site might include a boundary separating it from its environment, the ability to sense its environment, and the ability to modify the environment as a consequence. In this case, it would seem to be possible to construct an artificial device supporting protophenomena, but the specific requirements would have to be determined empirically.

Chalmers (1996, ch. 8) considers the possibility that *information spaces* may provide the link between the physical and the phenomenal, since they can be realized either as physical systems or as phenomenological structures. In particular, he suggests that quite simple physical systems might have associated protophenomena (p. 298). An information space is characterized by "differences that make a difference," that is, by distinctions that causally affect behavior. The physically realized information space must have a sufficient number of states to support the distinctions and must have the appropriate causal relations. The structure of the phenomenal space corresponds to the structure of the information space (due, in protophenomenal terms, to the protophenomena having interdependencies that correspond to the causal relations in the physical system).

Further, Chalmer's hypothesis and Cook's theory seem to be compatible. The binding of neurotransmitters to their receptors conveys information to a neuron about its extracellular environment, which can be quantified as an increase in the *system mutual information* between the cell and its environment (MacLennan, 2010).

In summary, if Chalmer's suggestion is correct, then many physically realized information spaces will be activity sites with associated protophenomena. In particular, since a robot's processor is devoted to the physical realization of information spaces, it would be reasonable to suppose that its constituent devices would have associated protophenomena. This would not, of course, imply that a robot is conscious, for protophenomena are not yet phenomena, but if the information processing were organized to create the appropriate protophenomenal interdependencies so that they cohered into phenomena and created a phenomenal world, then we could say that the robot is conscious. In particular, an appropriate structure among the protophenomena would produce emotional phenomena (felt emotions).

In robots, as in animals, a primary function of emotion is to make rapid assessments of external or internal situations and to ready the robot to respond to them with action or information processing. This may involve power management, shifting energy to more critical systems, adjustment of clock rates, deployment and priming of specialized actuators and sensors, initiation of action, and so forth. These processes will be monitored by interoceptors (internal sensors) that measure these and other physical properties (positions, angles, forces, stresses, flow rates, energy levels, power drains, temperatures, physical damage, etc.) and send signals to higher cognitive processes for supervision

and control. Therefore, many of these interoceptors will be distributed around the robot's body and this spatial organization will be reflected in somatosensory maps or other information structures. As a consequence coherent patterns among the interoceptive signals will be represented, and the associated protophenomena will cohere into bodily-organized phenomena.

In this way emotional phenomena are structured spatially in relation to the body, but these phenomena also have a qualitative structure, which may vary depending on the input space of the interoceptors. Each interoceptor will have a response function defined over its input space, but connections among the interoceptors in a region and connections to higher-order sensory areas will stitch together a topology representing the composite input space (MacLennan, 1995, 1999). This topology defines the qualitative structure of the resulting emotional phenomena.

Some of a robot's sensory spaces will be similarly structured spatially and qualitatively to ours, and in these cases we can expect the robot's emotional experiences (its feelings) to be similar to our own. Examples might include pressure sensors in the skin and angle and torque sensors in the joints. On the other hand, other interoceptors will be quite different from humans'. For example, a robot is unlikely to experience a quickened heartbeat or shallow, rapid breathing (because it is unlikely to have a heart or lungs), and we, in contrast, do not experience a redistribution of electrical power, which a robot might. These interoceptive spaces will have their own topologies, which determine their phenomenal structures, and so in these cases we must expect the robot's emotional experiences to be significantly different to ours and alien to us. Although we may be unable to imagine them, we will be able to understand their abstract structure, which will give us some insight into the robot's experience. In general, a robot's emotions will be peculiar to its "form of life," as ours are to ours (MacLennan, 1996).

If this is the case, one might question why these robotic experiences should be considered emotions at all. One reason is their similar function to natural emotions (explained above). For example, they will reflect general goals of critical importance to the robot's behavior, which are therefore directly motivating, and that consequently have a persisting, pervasive, appropriate effects on the physical state of the robot (by controlling sensors, effectors, and information processing). Another reason is that, due to the need for rapid, pervasive response, these experiences will have unconscious roots (i.e., below the level of coherent phenomena); conscious experience will be secondary and modulated by the already activated emotion.

5 Conclusions

In conclusion, I have argued that it is by no means impossible that some future robots may feel their emotions, that is, that they may have subjective emotional experiences homologous, but not identical, to ours. To determine the precise conditions sufficient for robot feelings it will be necessary to conduct detailed neurophenomenological investigations of subjective experience in order to isolate the physical processes correlated with the smallest units of that experience. This will enable us to formulate empirically testable hypotheses about the sorts of nonliving physical systems (if any) that may support protophenomena, and therefore conscious experience. This, in itself, is not sufficient to imply that robots could feel their emotions, for it is also necessary to understand neural structures underlying emotional experience, and the corresponding interdependencies among emotional protophenomena. The results of these neurophenomenological investigations will show us how to structure the emotional protophenomena of robots so that they cohere into emotional phenomena, that is, so that the robots feel their emotions. Thus, although significant questions remain unanswered, they can be addressed empirically, and their answers will allow us to decide whether robots could suffer, and thus deserve ethical treatment.

References

Atmanspacher, H. (2012). Dual-Aspect Monism à la Pauli and Jung. *Journ. Consciousness Studies*, 19, 96–120.

Chalmers, D. J. (1996). *The conscious mind*. New York: Oxford Univ. Press.

Cook, N. D. (2000). On defining awareness and consciousness: The importance of the neuronal membrane. In *Proceeding of the Tokyo-99 conference on consciousness*. Singapore: World Scientific.

Cook, N. D. (2002). *Tone of voice and mind: The connections between intonation, emotion, cognition and consciousness.* Amsterdam: John Benjamins.

Cook, N. D. (2008). The neuron-level phenomena underlying cognition and consciousness: Synaptic activity and the action potential. *Neuroscience*, 153, 556–70.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Avon.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt, Brace & Co.

Ihde, D. (1986). *Experimental phenomenology: An introduction.* Albany: State Univ. New York Press.

Losonczy, A., Makara, J. K., & Magee, J. C. (2008). Compartmentalized dendritic plasticity and input feature storage in neurons. *Nature*, 452, 436–40.

MacLennan, B. J. (1995). The investigation of consciousness through phenomenology and neuroscience. In J. King & K. H. Pribram (Eds.), *Scale in conscious experience: Is the brain too important to be left to specialists to study?* (pp. 25–43). Hillsdale: Lawrence Erlbaum.

MacLennan, B. J. (1996). The elements of consciousness and their neurodynamical correlates. *Journal of Consciousness Studies*, 3, 409−24. Reprinted in J. Shear (Ed.), *Explaining consciousness: The hard problem* (pp. 249–66). Cambridge, MA: MIT, 1997.

MacLennan, B. J. (1999). *The protophenomenal structure of consciousness with especial application to the experience of color: Extended version* (Technical Report UT-CS-99-418). Knoxville: University of Tennessee, Knoxville, Department of Computer Science. Retrieved March 24, 2013 from web.eecs.utk.edu/~mclennan

MacLennan, B. J. (2008). Consciousness: Natural and artificial. *Synthesis Philosophica*, 22(2), 401–33.

MacLennan, B. J. (2009). Robots react but can they feel? A protophenomenological analysis. In J. Vallverdú & D. Casacuberta (Eds.), *Handbook of research on synthetic emotions and sociable robotics: New applications in affective computing and artificial intelligence* (pp. 133–53). Hershey, NJ: IGI Global.

MacLennan, B. J. (2010). Protophenomena: The elements of consciousness and their relation to the brain. In A. Batthyány, A. Elitzur & D. Constant (Eds.), *Irreducibly conscious: Selected papers on consciousness* (pp. 189–214). Heidelberg & New York: Universitätsverlag Winter.

McCall, R. J. (1983). *Phenomenological psychology: An introduction. With a glossary of some key Heideggerian terms.* Madison: Univ. Wisconsin Press.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. New York: Oxford Univ. Press.

Pettit, D. L., Wang, S. S., Gee, K. R., & Augustine, G. J. (1997). Chemical two-photon uncaging: a novel approach to mapping glutamate receptors. *Neuron*, 19, 465–71.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution*. New York: American Psychological Assoc.

Prinz, J. (2006). *Gut reactions: A perceptual theory of emotion*. New York: Oxford Univ. Press.

Rolls, E. T. (2005). *Emotion explained*. Oxford: Oxford Univ. Press.

Rolls, E. T. (2007). A neurobiological approach to emotional intelligence. In G. Matthews, M. Zeidner & R. D. Roberts (Eds.), *The science of emotional intelligence* (pp. 72–100). Oxford: Oxford Univ. Press.

Service, R. F. (2013). The cyborg era begins: Advances in flexible electronics now makes it possible to integrate circuits with tissues. *Science*, 340, 1162–5.

Strawson, G. (1994). *Mental reality*. MIT Pr., Cambridge.

Strawson, G. (2006). Realistic monism. In A. Freeman (Ed.), *Consciousness and its place in nature: Does physicalism entail panpsychism?* (pp. 3–31). Imprint Academic, Charlottesville, VA.

Sur, M. (2004). Rewiring cortex: Cross-modal plasticity and its implications for cortical development and function. In G. A. Calvert, C. Spence & B. E Stein (Eds.), *Handbook of multisensory processing* (pp. 681–94). Cambridge: MIT Press.