Running head: ETHICAL TREATMENT OF ROBOTS

Ethical Treatment of Robots and the Hard Problem of Robot Emotions

Bruce J. MacLennan

Department of Electrical Engineering & Computer Science

The University of Tennessee

Knoxville, TN 37996 USA

Abstract

Emotions are important cognitive faculties that enable animals to behave intelligently in real time. We argue that many important current and future applications of autonomous robots will require them to have a rich emotional repertoire, but this raises the question of whether it is possible for robots to experience their emotions consciously, as we do. Under what conditions would phenomenal experience of emotions be possible for robots? This is, in effect, the "hard problem" of robot emotions. Our paper outlines a scientific approach to the question grounded in experimental neurophenomenology.

Ethical Treatment of Robots and the Hard Problem of Robot Emotions

Emotional Robots

In this paper I focus on *the hard problem of robot emotions*: the possibility and preconditions for a robot experiencing its emotions, and on its implications for the ethical treatment of robots. Everyday notions of ethical treatment depend in part on the recipient's capacity to suffer, which 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 I will not address positive emotions in this paper. Future autonomous 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 will be more likely if they can empathize with our feelings of suffering and joy. This capacity is more compelling if it goes beyond intellectualized empathy and includes empathetic feeling (such as we have via mirror neurons). But why should we equip robots with emotions at all?

An emotion may be defined as a state that is evoked by a reward or a punisher, which might be either present or remembered, and that serves as positive or negative reinforcement (Rolls, 2007). This reinforcement leads to changes in behavior that are adaptive in the sense of inclusive fitness (Plutchik, 2003, pp. 218–23). Certainly, many robots will not need emotions, but animals depend on emotions for efficient, real-time behavior, and for analogous reasons we can expect them to be valuable in some autonomous robots. More specifically, Rolls (2005, 2007) enumerates a number of functions fulfilled by emotions, which have analogues in robotics. First, emotion is motivating and directed toward action, and likewise robots need a means for selecting goals and organizing subservient activities. Also, natural emotions provide for response

flexibility through a "bow tie" organization. That is, many different stimuli may lead to a single behavioral goal, represented by the emotion, which can be achieved by a variety of means (Rolls, 2006). Likewise in robots, it is useful to identify general motivational states that can be triggered by a variety of stimuli and fulfilled in a variety of ways. Further, in animals an emotion establishes a persisting state (e.g., a mood) that biases cognitive processing to be more appropriate to the situation, and emotions can serve the same purpose in robots. In particular, emotion is crucial in memory encoding and rapid retrieval of behaviorally relevant information, which is valuable in robots as well. Further, natural emotions trigger autonomic and endocrine responses, which affect adrenaline release, heart rate, and other functions. Analogously, robot emotions might affect power management, reallocation of computational resources, adjustment of clock rates, preparatory deployment and priming of sensors and actuators, and so forth. Finally, emotions are critical in regulating interactions among animals, promoting cooperation and other forms of social organization, and in facilitating communication of mental states, attitudes, intentions, etc. through emotional expression. These functions are also important in robots that cooperate with each other or with humans (Breazeal, 2003; Breazeal, Brooks, Gray, Hoffman, Kidd, Lee, Lieberman, Lockerd & Chilongo, 2004).

The foregoing implies that some robots will benefit from having systems that are functionally equivalent to emotions, but does that imply that they will *feel* their emotions? This is the "hard problem" of robot emotions. It is a subproblem of the "hard problem" of consciousness: "The really hard problem of consciousness is the problem of experience. … It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises" (Chalmers, 1995). It would seem to be possible that a robot could have an internal representation corresponding to pain, that this representation could be created by

potentially damaging stimuli, and that these circumstances could cause the robot to behave as though in pain, but without the accompanying subjective experience of pain. However, without a solution to the hard problem, we cannot say if this is a genuine possibility or not, or the circumstances under which a robot might feel pain, fear, or other emotions.

Our ethical treatment of robots will depend in part on 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 action — especially in the moment — is enhanced by our vicarious experiencing of another's feelings, in particular, of 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.

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 neurophenomenological 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 is not limited to perception, 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 human conscious state perhaps comprises from 10^8 (for minicolumns) to 10^{15} (for synapses) protophenomena. 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 the dynamics of protophenomena 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 the topology of their interdependencies (MacLennan, 1999, 2010). That is, qualia arise from protophenomenal interdependencies. (Isolated protophenomena are not qualia per se.)

The foregoing must suffice as a summary of the neurophenomenological approach. Space limitations preclude addressing criticisms of neurophenomenology and the notion of protophenomena, but they are discussed elsewhere (MacLennan, 1995–2011).

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). Especially useful for our purposes is work on the neurophysiology of emotions by evolutionary psychologists (e.g., Panksepp, 2004; Plutchik, 2003; Prinz, 2006), for they investigate emotions in a variety of species. Therefore it is easier to separate the general features of emotion from the specifics of human emotion and to transfer the general features to robots.

Neurophenomenology requires that the neurophysiological emotional response be distinguished from the feeling of an emotion (Damasio, 1999, pp. 42–9). The former takes place in the limbic system and is largely unconscious, but causes physiological changes that affect conscious experience and are interpreted as emotions. This *visceral* or *somatic* theory of

emotion, which was proposed independently by William James (1884) and Carl Lange (1885), is the basis (with modifications) of most contemporary theories (e.g., Damasio, 1994, 1999; Prinz, 2006).

Neurophenomenological investigations support a three-level hierarchy of emotional processing that is compatible with a protophenomenal account of emotional experience. According to Prinz (2006, ch. 9), the lowest level comprises neurons with small receptive fields that respond to visceral organs, hormone levels, skeletal muscles, etc. We expect these neurons to be activity sites corresponding to emotional protophenomena, which are insufficiently interconnected to cohere into full-fledged emotional phenomena. This integration is accomplished in cortical maps at the second level, where emotional phenomena are experienced, that is, where feelings are felt (Damasio, 1999, ch. 9). At the third level of the hierarchy the emotion is consciously categorized and recognized (Prinz, 2006, p. 214), which expands it from a purely emotional phenomenon into a conscious cognitive state.

Our understanding of the neurophenomenology of human emotion is still at an early and incomplete stage. As research continues, we expect to be able to establish more detailed correlations between the activity of neurons in specific cortical regions and qualitative and quantitative aspects of emotional experience. This will enable us to predict and craft the emotional lives of future robots.

Empirical Issues in Robot Emotions

With these insights from the neurophenomenology of human emotion, we can begin to address the possibility of conscious emotions in robots. In humans conscious experience consists in variations in protophenomenal intensity correlated to specific physical activity in neural tissue. 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). The question of robot consciousness hinges, then, on whether the processes in the robot's "brain" are sufficiently similar to the human brain's processes *in the relevant physical ways*. (This is an issue of physical properties, not functionalism.) Determining the relevant physical properties will require detailed investigation of the affect on conscious experience of specific physical procedures (e.g., control of membrane potentials, ion channels, neurotransmitter receptors, etc.). Fortunately, there has been remarkable progress in our ability to conduct such experiments (e.g., Petit et al., 1997; Losonczy, Makara, and Magee, 2008; Service, 2013). On this basis we will be able to make empirically justified predictions of the sorts of physical processes in robots that would have accompanying protophenomena.

What sorts of physical systems might have associated protophenomena? Chalmers (1996, ch. 8) speculates that all *information spaces* have protophenomena, for they can be simultaneously physical and phenomenological. The essence of an information space is a "difference that makes a difference"; that is, a discrimination that affects subsequent behavior. Such an information space can be understood in terms of physical causality, but also in cognitive terms as discrimination and action. The physical and protophenomenal aspects are tightly correlated and nearly coincident. If Chalmers is correct, then we should expect that the information spaces instantiated in robots should have associated protophenomena, which appropriate interdependencies could integrate into full-fledged conscious phenomena.

Norman Cook (2000, 2002, chs. 6–7, 2008) has suggested that the intensity of a protophenomenon is correlated to the ion flux across a neuron's membrane when it generates an action potential. A related idea is that protophenomenal intensity correlates with the binding of neurotransmitters to receptors, which affects the neuron's behavior (generation of an action

potential). This is, in effect, the neuron's sensing of its extracellular environment, with consequent action: a difference that makes a difference. In both physical and information theoretic-terms, there is an increase of *system mutual information* (effectively, negative entropy) of the neuron and its environment (MacLennan, 2011). Here again, there does not seem to be a reason why an artificial neuron or similar information-processing device would not have an associated protophenomenon.

In summary, plausible ideas about the nature of activity sites, which identify them with elementary physical information processes, would seem to be compatible with protophenomena in artificial systems. Nevertheless, these theories await empirical verification.

Even if we can construct a robot whose physical devices support protophenomena, that does not imply that the robot will experience emotional phenomena, that is, felt emotions. For the latter, the protophenomenal dependencies will have to be so structured that the protophenomena cohere into phenomena, which requires a corresponding structure in the physical information processing system.

The somatic theory of emotion and its variants imply that fundamental components of emotional experience are feelings of bodily changes resulting from unconscious emotional preprocessing. Analogously in robots, we expect feelings to be based upon *interoceptors* distributed around the robot's body. They might include sensors for temperature, power drain, energy levels, flow rates, excessive signal strength, joint angles and torques, surface pressure, stresses, physical damage, and so forth. These very specific, localized sensors correspond to the first level of Prinz's hierarchy. In order to make sense out of these sensations they will be organized at the second level in computational maps to reflect their bodily organization and

relevance to bodily systems. This is the level at which protophenomena are assembled into bodily-organized emotional phenomena and experienced consciously.

The qualitative phenomenology of perceptions depends on the topology of the interdependencies of their constituent protophenomena (MacLennan, 1995, 1999, 2010). In the case of robot emotions, these topological relationships will be a function of the input spaces of the interoceptors and of their interconnections in second- and higher-order representations. The topological relationships in protophenomenal space will determine what the emotions feel like. For example, sensations of pressure distributed across the body surface and stress in the joints might trigger reorientation of attention and redistribution motive power to escape from a confining situation; at higher levels, this might be experienced analogously to fear, which might trigger fight-or-flight reactions.

We can expect future robot bodies to be significantly different from the bodies of humans and other animals, but homologous in many respects. As a consequence, some robot emotions will have phenomenological structures comparable to human emotions, but some others could be radically different (MacLennan, 1996). In the latter cases, we will be able to describe the qualitative structure of the synthetic emotions in terms of the topology of the phenomena, but we may be unable to imagine what they feel like to the robots. Nevertheless, these robotic experiences should be classified as emotions if they fulfill the functions enumerated by Rolls (2005, 2007), which are summarized at the beginning of the present article.

Conclusions

Emotions are critical to the survival of mammalian species, and we expect synthetic emotions to be essential in future autonomous robots, especially in those that interact with humans or each other. However, this expectation leaves open the question of whether such robots

will feel their emotions, which is the "hard problem" of robot emotions. The solution to this problem depends on the neurophenomenological investigation of human emotions, which is still in its infancy. Fundamental to this investigation is the determination of the sorts of physical systems that have associated protophenomena, which are the elementary constituents of conscious emotional phenomena (felt emotions). This is an empirical question, but one attractive hypothesis is that protophenomena are associated with fundamental physical processes of discrimination and action, which occur in robots as well as animals.

The presence of protophenomena is necessary but not sufficient for full-fledged emotional phenomena (feelings), which are the coherent activity of masses of protophenomena. In mammals this coherence arise from the bodily organization of the constituent protophenomena, consistent with the somatic theory of emotions. Although robot bodies will be very different from mammalian bodies, robotic emotions are homologous to natural emotions, and will have homologous bodily organization. Therefore, if the necessary physical conditions for emotional protophenomena were satisfied, we would expect them to cohere into emotional phenomena, that is, that the robot would feel its emotions. Such robots would be capable of feeling pain and suffering in other ways, and therefore, if for no other reasons, deserving of ethical treatment.

References

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

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journ. Consciousness Studies* 2, 200–19.

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 K. Yasue (Ed.), *Proceeding of the Tokyo-99 conference on consciousness* (pp. 19–20). 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. *Journ. 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, 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.

MacLennan, B. J. (2011). Protophenomena and their physical correlates. *Journal of Cosmology*, 14 (April–May, 2011), 4516–4525. Reprinted in R. Penrose & S. Hameroff (Eds.), *Consciousness and the universe: Quantum physics, evolution, brain & mind* (pp. 228–39). Cambridge, MA: Cosmology Science Publishers.

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. (2006). Brain mechanisms of emotion and decision-making. *International Congress Series*, 1291, 3–13. Amsterdam: Elsevier.

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.