# Chapter 56

# Ethical Treatment of Robots and the Hard Problem of Robot Emotions

**Bruce J. MacLennan**
*University of Tennessee, USA*

## ABSTRACT

*Emotions are important cognitive faculties that enable animals to behave intelligently in real time. The author argues 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 people do. Under what conditions would phenomenal experience of emotions be possible for robots? This is, in effect, the "hard problem" of robot emotions. This paper outlines a scientific approach to the question grounded in experimental neurophenomenology.*

## INTRODUCTION

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 of them, but also their treatment of us. For we expect robots to treat us well, but that will be more likely if they can empathise with our feelings of suffering and joy. This capacity is more compelling if it goes beyond intellectualised 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 punishment, which might be either present or remembered, and that serves as positive or negative reinforcement (Rolls, 2007). This reinforcement leads to changes in behaviour that are adaptive in the sense of inclusive fitness (Plutchik,

2003). Certainly, many robots will not need emotions, but animals depend on emotions for efficient, real-time behaviour, 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 organising subservient activities. Also, natural emotions provide for response flexibility through a "bow tie" organisation. That is, many different stimuli may lead to a single behavioural 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 behaviourally 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 organisation, 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 dehumanising consequences for them and us. For example, if future robots simulate feeling with great verisimilitude but we believe that they do not 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.

## Related Content

The Role of Students' Personality Traits in the Effective Use of Social Networking Sites in the Educational Context
Josip Burusicand Mia Karabegovic (2016). *Psychology and Mental Health: Concepts, Methodologies, Tools, and Applications  (pp. 1205-1224).*
www.irma-international.org/chapter/the-role-of-students-personality-traits-in-the-effective-use-of-social-networking-sites-in-the-educational-context/153445

Game Theory
William Amone (2016). *Neuroeconomics and the Decision-Making Process (pp. 262-286).*
www.irma-international.org/chapter/game-theory/148119

I'm Always Online: Well-Being and Main Sources of Life Dis/Satisfaction of Heavy Internet Users
Tihana Brkljai, Filip Majetiand Anja Wertag (2018). *Psychological, Social, and Cultural Aspects of Internet Addiction (pp. 72-89).*
www.irma-international.org/chapter/im-always-online/193095

Real-Time Facial Emotion Analysis for Adaptive Teaching Strategies Using Deep Learning
V. Suganthiand M. Yogeshwari (2024). *Explainable AI Applications for Human Behavior Analysis (pp. 142-156).*
www.irma-international.org/chapter/real-time-facial-emotion-analysis-for-adaptive-teaching-strategies-using-deep-learning/347683

Some Basic Elements of Psychoanalysis According to Sigmund Freud
 (2018). *Computational Psychoanalysis and Formal Bi-Logic Frameworks (pp. 38-113).*
www.irma-international.org/chapter/some-basic-elements-of-psychoanalysis-according-to-sigmund-freud/195908