Cruelty to Robots?
The Hard Problem of Robot Suffering
Bruce MacLennan
University of Tennessee, Knoxville
“Perhaps the greatest significance of the computer lies in its impact on Man’s view of himself ... [T]he computer aids him to obey, for the first time, the ancient injunction *Know thyself.***”

— Herb Simon
Computing and Philosophy

- Closely connected from the beginning
  Leibnitz, Boole, Ada Lovelace,
  Turing, von Neumann, Markov,
  McCulloch, Newell & Simon,
  Minsky, McCarthy, et al.

- Common topics
  epistemology, ontology,
  computability, philosophy of math.,
  logic, information, language,
  philosophy of mind, aesthetics,
  ethics

- IACAP founded 25 years ago
Minds, Machines, and Morals: Some Issues

- How are moral norms learned?
- What is the logic of moral reasoning?
- Could robots be held morally responsible?
- Might we have ethical obligations to future robots?
Some papers from IACAP ‘13

- What Would it Take to Build a Moral Robot?
- Relationship Between Intelligent, Autonomously Functioning Machines and Ethics
- Designing Moral Normative Constraints on Machine Behavior
- Computational Models of Moral Perception, Conflict and Elevation
- A Computational Account of Complex Moral Judgment
- On Deception and Trust in Artificial Agent Development
- Integration of Cognitive and Metacognitive Processes with Data-driven and Knowledge-rich Structures
- Mind-, Body- and Emotion-Reading
- Attempts to Attribute Moral Agency to Intelligent Machines are Misguided
- Modal Reasoning as Simulation
- Rational Learners and Non-utilitarian Rules
- Introduction to Moral Induction Model and its Deployment in Artificial Agents
- Designing and Implementing a Model for Ethical Decision Making
- Unmanned Military Systems, the Frame Problem, and Computer Security
- Neuroethical Challenges of Brain Simulations
- Privacy as Informational Commodity
- Man, Machinery and Cryptocoin Avarice
- Why There is No Escaping Physical Identity in the Virtual World
- Just War Theory and Information Warfare
- Ethical Issues of Brain-computer Interfaces
- On the Humanity of Making a Computer that Feels Pain
- Toward a Socio-technical epistemology

<www.iacap.org/conferences/iacap2013/iacap_2013_proceedings/>
Recent Related Papers


“The Hard Problem of Consciousness”

Chalmers coined the term “hard problem of consciousness”: "The really hard problem of consciousness is the problem of experience. When we think and perceive, there is a whirl of information-processing, but there is also a subjective aspect. ... It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does.” (1995)

“The hard problem of consciousness is the problem of explaining why any physical state is conscious rather than nonconscious.” (Internet Encyclopedia of Philosophy)
Cruelty and the Hard Problem

The “Hard Problem” of robot emotions: the possibility and preconditions for a robot to experience its emotions

Ethical treatment depends on capacity so suffer

Affects our treatment of them

Affects their treatment of us

genuine empathy
Rolls’ Definition of Emotion

- A state elicited by:
  - delivery or omission
  - of a reward or punisher
  - either present or remembered
  - that functions as positive or negative reinforcement
Functions of Emotions (Rolls)

- elicit autonomic & endocrine responses
- facilitate response flexibility
- inherently motivating
- facilitate communication
- social bonding

- persistent effect on cognitive processing
- facilitate memory storage & retrieval
- coherent context for cognitive processing
- evoke retrieval of non-cortical memories
Zombie Robots?

Could a robot have appropriate information structures and control processes to fulfill the functions of emotions, but without feeling them?

A matter of opinion in the absence of principled, preferably scientific, approach to the Hard Problem of robot emotions.
Neurophenomenological Analysis

- Parallel reductions in:
  - neurological domain (external)
  - phenomenological domain (internal)

- Phenomenal consciousness is reducible:
  - qualitatively
  - quantitatively
Qualitative Reduction

- Separates conscious experience into phenomena of different kinds
- Different sensory modalities:
- Other modalities:
  - waking, dreaming, imagination, recollection, anticipation, desire, ...
- Naïve qualitative analysis corrected by:
  - experimental phenomenology
  - insights from neuroscience
Quantitative Reduction

- Analysis of phenomena into constituent phenomena of the same kind
- Examples: visual phenomena; proprioceptive and haptic phenomena
- Phenomenological analysis reinforced by knowledge of receptive fields of neurons in primary sensory cortices
Comprehensive Neurophenomenological Analysis

- Neurophenomenological analysis cannot be limited to perception
- Addresses neural correlates of all conscious experience
- Awaits progress in neuroscience and experimental phenomenology
- Goal: explanation of structure & dynamics of consciousness in terms of more elementary phenomenological processes
Protophenomena

- Practical end to neurological reduction
  - neurons? minicolumns? synapses? receptor sites?

- Protophenomenon: corresponding unit of phenomenological reduction

- Phenomena arise from coordinated behavior of constituent protophenomena

- Conscious state results from interaction of vast numbers of protophenomena
  - $10^8$ to more than $10^{15}$

- Activity site: physical system corresponding to a protophenomenon
Protophenomenal Intensity

- Protophenomena are elementary degrees of freedom of conscious state

- **Intensity** of protophenomenon = strength of presence in conscious state

- Correlated to physical activity at corresponding activity site
  - Some possibilities: membrane potential, firing rate, number of occupied receptor sites
Protophenomenal Interdependencies

- Neurodynamics of brain governed by interconnection of activity sites
- Corresponding interdependencies among intensities of protophenomena
- Lead to coherent dynamics of protophenomena, which constitute *phenomena* proper
Structural Theory of Qualia

- Common cortical architecture across cerebrum
  - auditory cortex can support visual perception and vice versa

- Perceptual properties determined by neural connections

- Correspondingly, phenomenal qualities determined by protophenomenal interdependencies

- Qualia arise from protophenomenal interdependencies
  - Protophenomena are not qualia per se
Neurophenomenology of Human Emotion

- Emotions vs. feelings
- Somatic Feeling or James-Lange Theory
  - William James (1884) and Carl Lange (1885)
- Three-level hierarchy (Prinz, Damasio):
  - somatic neurons $\sim$ emotional protophenomena
  - cortical maps $\sim$ emotional phenomena
  - categories $\sim$ recognized emotions
  - “an emotion, the feeling of that emotion, and knowing that we have a feeling of that emotion” (Damasio)
- Neurophenomenological agenda is to correlate:
  - neuropsychology of emotion
  - phenomenology of emotional experience
Prinz’ Emotional Processing Hierarchy

Figure 9.2. Brain structures involved in the emotion hierarchy. SI, primary somatosensory cortex; SII, secondary somatosensory cortex; In, insula; dACC, dorsal anterior cingulate cortex; rACC, rostral anterior cingulate cortex; VMPFC, ventromedial prefrontal cortex.

1) primary somatosensory cortex (SI), pons, insula (In)
2) secondary somatosensory cortex (SII), dorsal anterior cingulate cortex (dACC), insular cortices
3) ventromedial prefrontal cortex (VMPFC), rostral anterior cingulate cortex (rACC)

(Prinz, Gut Reactions, fig. 9.2)
Empirical Issues in Robot Emotions

- Correlation of protophenomenal intensity and physical activity at activity sites treated as brute fact
  - Dual-aspect monism (Chalmer’s Type-F monism)

- Empirical question: What range of physical processes have corresponding protophenomena (in humans)?

- Inference: the same physical processes would be associated with protophenomena in robots

- Open question, but relevant experimental procedures are becoming available

- Some hypotheses...
Cook’s Hypothesis

- Intensity of protophenomenon correlated with flux of ions across cell membrane when ion channels open during action potential.
- Neuron’s sensing of its intercellular environment.
- Essential physical properties:
  - boundary separating activity site from environment
  - ability to sense environment
  - consequent modification of environment
- Possible in robots

(figure from Cook)
Chalmers’ Hypothesis

- Information spaces have both physical and phenomenal aspects
- “Differences that make a difference”
- Physical information space must have sufficient structure and causal relations
- Phenomenal information space has corresponding structure
- Combining Cook and Chalmers: increase of system mutual information between cell and environment
  - information is both physical and phenomenal
Activity Sites in Robots

- Chalmers’ hypothesis ⇒ physical information spaces will be activity sites with associated protophenomena
- Robot’s central processor physically realizes information spaces ⇒ supports protophenomena
- Protophenomena would need to be appropriately structured through interdependencies to cohere into emotional phenomena (felt emotions)
- What is this structure?
Distributed Interoceptors

- Primary function of robot emotions:
  - make rapid assessments of external or internal situations
  - ready the robot to respond with action or information processing
  - may involve power management, adjustment of clock rates, deployment and priming of specialized actuators and sensors, initiation of action, etc.

- Interoceptors: internal sensors

- Distributed around robot’s body

- Bodily organization
  - ⇒ coherent somatic phenomena
  - ⇒ felt emotions
Phenomenological Structure of Robot Emotions

- Bodily organization of protophenomena
- Each interoceptor has a response function defined over its input space
- Interoceptors are connected to each other and into higher-order sensory areas
- Creates topology over composite input space
- Defines qualitative structure of corresponding emotional phenomena
Robot vs. Human Emotions

- Some of robot’s somatosensory spaces will be structured similarly to ours
  - e.g., skin and joint sensors

- In these cases, emotional experiences will be similar to ours

- Other interoceptor systems will be very different from ours and have different phenomenological structures
  - may be understandable but not imaginable
  - peculiar to the robot’s “form of life”

CISML — 28 Feb. 2014
Why call these phenomena ‘emotions’?

- They fulfill similar functions to natural emotions
  - reflect general goals of critical importance
  - directly motivating
  - persisting, pervasive, appropriate effects on physical state of robot

- Effect on consciousness is similar to natural emotions
  - modulate structure & dynamics of conscious state
  - relevantly to particular behavioral ends
Conclusions

- Future robots could have conscious emotional experiences homologous, but not identical to ours

- Detailed neurophenomenological experiments will be required to determine the precise conditions

- In particular, empirical investigations are required for
  - determining the range of physical processes with corresponding protophenomena
  - understanding the neural structures underlying emotional experience

- The result will be a principled answer to whether and under what conditions robots could feel their emotions