

# Uncanny Moral Behavior

Carson Reynolds

*University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*

---

## Abstract

In trying to imitate human moral behavior, will we venture into the moral uncanny valley where the decisions appear superficially moral and yet disgust us? The uncanny valley describes a feeling of eeriness and ominousness which accompanies near-human like appearances and behavior in robots. As the sophistication of physical and virtual actroids has increased, under controlled circumstances they can be mistaken for humans. But outside of movement and appearance, I hypothesize here that as robotic, transhuman, or artificial entities behave more explicitly in a functionally moral manner humans will respond with feelings of revulsion.

The paper describes several examples of functionally moral behavior and for each plot acceptability versus culpability. I assert that the most human-similar moral behavior is that which is identical in terms of culpability. This paper's hypothesis is that when systems are very similar to humans in terms of culpability, their actions may become repugnant.

In addition to traditional cases like industrial robot, corpse, zombie, and healthy person we will consider autopilots, fire control systems, an organ-swapping robotic hospital, an automatic trolley switch, and automatic quarantine nanobots. In these cases, some drawn from actual incidents, others from speculation we find it useful to draw parallels between the culpability of a human versus progressively more sophisticated moral systems.

The argument is wrapped up through examination of the questions this model poses, such as whether a transhuman or cyborg who fuses human and artificial moral deliberation is less or more repugnant than an ordinary human. Or whether moral deliberation can be meaningfully distilled down into emotive notions like acceptability versus repugnance.

*Key words:* Uncanny Valley, Roboethics, Transhumanism, Emotivism

---

## 1 The Uncanny Valley

Upon encountering forms which are human-like and yet viscerally dissimilar it has been theorized humans react with feelings related to the uncanniness

*Article submitted to the European Conference on Computing and Philosophy, 2008*

of the form. This hypothesis (which is widely known to roboticists) was well illustrated in an article by (Mori, 1970) and was also anticipated in an article originally published in 1906 by Jentsch:

...one of the most successful devices for easily creating uncanny effects is to leave the reader in uncertainty whether a particular figure in the story is a human being or an automaton... Jentsch (1996)

Although Jentsch is referring to fictional writing, both robotics researchers and artists (such as Ron Mueck) have explored uncanniness with physical objects. Geiger et al. (2003) demonstrated synthetic video animations which were not distinguishable from authentic video captures of a live speaker. More notably, the android projects of MacDorman and Ishiguro (2006) have become oft-cited examples of uncanniness. Other work involving these robots by Arita et al. (2005) showed that 10-month old infants who are exposed to the robots for a brief period of time gaze at interactive robots and humans for similar periods of time when compared to none-moving robots.

The concept of the uncanny valley is not accepted universally by robotics researchers. Hanson et al. (2005) dispute the existence of the uncanny valley and suggest that alternative models are needed.

However, these works focus on projects which seek to approach human appearance and movement. If we instead considered the moral behavior of the robot does such a valley appear? The concept of functional morality is developed in *Moral Machines* by Wallach and Allen (2008). I shall proceed by describing some examples of functionally moral behavior:

- Misinformed Autopilot: in 2005 an autopilot system on Malaysia Airlines Flight 124 failed causing the plane to climb and nearly stall. Culpability: low. Acceptability: low.
- Fire-Control System: in 2007 a robotic anti-aircraft system misfired killing nine. Culpability: medium. Acceptability: low.
- Organ Swapping Robotic Hospital: an extension of the utilitarian forced organ donation thought experiment. Culpability: medium-high. Acceptability: very low.
- Automatic Trolley Switch: an extension of the trolley thought experiments in which an agent throws the switch. Culpability: medium-high. Acceptability: low.
- Nanobot Quarantine: implanted nanobots monitor their host's condition and quarantine by inducing a coma when extremely abnormal and contagious conditions are detected. Culpability: high. Acceptability: neutral.

## 2 The Uncanny Moral Valley

If we are to make a plot with culpability ranging along the horizontal axis and acceptability along the vertical an interesting impression appears. Assuming that industrial robots and humans (in the absence of other details) are neutral in terms of acceptability we find them along the base line. However, as the various scenarios described above increasing in the degree of culpability which can be attributed to the entity in question we see the familiar trend of a dip in acceptability as systems become more culpable but not quite fully responsible.

It has argued by Kohlberg that humans develop their moral sense over time. If this is the case then we might infer that some point along the path to full adult culpability, children will exhibit repugnant moral behavior. Exactly when this might occur, as implied by this paper's theory, is left for speculation.

### References

- Arita, A., Hiraki, K., Kanda, T., and Ishiguro, H. (2005). Can we talk to robots? ten-month-old infants expected interactive humanoid robots to be talked to by persons. *Cognition*, 95(3):B49–B57.
- Geiger, G., Ezzat, T., and Poggio, T. (2003). Perceptual evaluation of video-realistic speech. Technical Report CBCL Paper 224 / AI Memo 2003-003, Massachusetts Institute of Technology.
- Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., Fernandez, R., and Stephanou, H. (2005). Upending the uncanny valley. In *AAAI'05: Proceedings of the 20th national conference on Artificial intelligence*, pages 1728–1729. AAAI Press.
- Jentsch, E. (1996). On the Psychology of the Uncanny. *Trans. Roy Sellars. Angelaki*, 2:7–16.
- Kohlberg, L. *The Philosophy of Moral Development: Moral Stages and the Idea of Justice (Essays on Moral Development, Volume 1)*. Harper & Row, 1st edition.
- MacDorman, K. and Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3):297–337.
- Wallach, W. and Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, USA.