EasyChair Preprint

№ 4212

Science + Fiction: Understanding Robot Intelligence

David M. W. Powers

September 19, 2020

Science + Fiction
Understanding Robot Intelligence

David M W Powers
COMP1002 Foundations of Computational Intelligence
Flinders University, South Australia

powe0093@flinders.edu.au

Abstract

The history of robotics is older than the invention and exploitation of robots. The term 'robot' came from the Czech and was first used in a play a century ago. The term 'robotics' and the ethical considerations captured by 'The Three Laws of Robotics' come from a SciFi author born a century ago. SF leads the way!

1 Introduction

Are robots going to take over the world? Or are they going to destroy it?

This seemed to be the choice in the literature on artificial beings that started with Mary Shelley's (1880) 'Frankenstein' and coined the word 'Robots' with Karel Čapek's (1920) 'R.U.R.' (Rossum's Universal Robots).

Is this something that scientists and engineers need to concern themselves with?

And perhaps there are some prior questions… Can a robot think, be conscious or sentient, have emotions, have a conscience, morality or ethics?

And then there are obvious legal questions… Robots or drones can potentially do illegal or unethical things: either deliberately of their own volition or on somebody's orders; or accidentally either due to overt programmer error or as a result of being in a no win, low probability win or other hard to solve situation – including when a human decides to play chicken or forces their way in when they don't have right of way. Who is responsible?

2 Science and Science Fiction

Both science researchers and science fiction authors make things up. But when the scientist author does it it's called a theory, and when the fiction author does it it's called a story.

But the hard science fiction author and the theoretical scientist can have the same agenda here. In both cases we start with commonly accepted assumptions and theories, then throw in the key new proposals, and explore the implications.

Karl Popper (1934/5) sought to formally understand the way science works and concluded that the key feature of scientific methodology and theory is falsifiability. That is a scientist should follow the implications of their theory into the unknown and make new predictions that may or may not work out as anticipated. That is exactly what a hard science fiction author does.

Isaac Asimov as a teenager despaired of seeing only negative robot stories so tried to develop positive stories, leading to the formulation of "The Three Laws of Robotics" (1940-1942) that were hardwired into a robot's positronic brain – and meant to stop the robot doing any harm. His stories in general revolve around ways in which these laws produce unexpected consequences or can be circumvented. Asimov is also credited with invention of the word 'robotics', which partly explains why the 'three laws' are so well enshrined in the lore of the field.

Asimov's (1954 & 1957) detective stories (involving a human-robot detective team) look in particular at how you can use a robot to commit murder despite the laws that are meant to prevent this. The authorized continuations (Allen, 1993-6; Tiedemann, 2000-3) continue in this robot crime vein. Just don't assume the robot is the criminal!

Asimov went further eventually (1982) with his 0^{th} law – we don't just have to worry about harming or saving individual humans, but harming or saving humanity… and showing humanity…

Arthur C. Clarke's (1968) *HAL* in *2001: A Space Odyssey*, and the Stanley Kubrik film of the same name, replace a robot with a computer-controlled spaceship and addresses similar issues, with *HAL* becoming iconic in his own right. The problems in *2001* being of a similar character to those Asimov illustrated in his stories about individual robots.

3 Science Fiction and Technology

Clarke's (1968 & 1982) *2001* and *2010* sequence is also interesting for the technical effort that went into being accurate about the technology of space travel and artificial intelligence, in particular how to build a sentient artificial intelligence like *HAL*. *HAL* was not programmed, but learned as a child. Another early precedent of this is Osamu Tezuka's (1952-1968) *Astro Boy*. This *manga* comic book series morphed into the *anime* cartoon series, and its author into a film director. This series defined the genre and captured 40% of Japan's TV audience, perhaps inspiring the robotic focus of Japan, and indeed a generation of AI and robotic researchers around the world. *Astro Boy* is a robot child, that has to learn about the world in every sense.

Powers and Turk (1989) argue that this is how an AI must develop if it is to have a human-like understanding of language and the universe we live in, and cites both *HAL* and *Astro Boy* in this PhD thesis alongside founder of psycholinguistics Jean Piaget – who wrote over twenty books exploring different facets of the question of how children learn to talk about the world, the 'sticky mirror' concept being Powers' formulation of Piaget ideas about 'reflection' (Piaget, 1923 & 1928).

The importance of Piaget's contribution was developing this area of child development as a scientific field, open to falsification – he also was a student of the philosophy of science. And some of the predictions made by his early theories were indeed overturned by later experiments.

Turing's (1950) paper is famous for the Imitation Game, aka the Turing Test. It is less well known for its explanation of how to build an AI that would pass it: "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's" (p.456). In relation to getting it to "understand and speak English", he suggests "provide the machine with the best sense organs that money can buy, and then teach it…" (p.460).

While Turing (1950) mentions some primitive experiments with teaching a computer, Block et al. (1975) try to provide a more convincing model to explain how a robot could learn English with a more detailed model of language involving syntax and semantics, and using a language learning game to explore the process with humans playing the role of computer – a different twist on the Turing Test.

Powers (1983-84) used both statistical and neural network methods to learn basic grammar, arguing for an unsupervised approach – babies don't have teachers who give them grammar lessons and mark their work. Surprisingly to some computer scientists, this led to functional words like 'the', 'and' and 'for' being learned before content words like 'cat', 'chased', 'bit' and 'dog.

But this was not so surprising to psycholinguists who were aware that children's understanding capability led their capability in imitation and production (Brown, 1970). And even older research on reading had shown how a good reader's eyes jump from functional word to functional word skimming over the content words (Huey, 1908). The grammatical words, like the prosody, are picked up early, so that a child can recognize not only their mother's voice at birth (Mehler et al. 1988) and indeed are learning aspects of language in the womb from at least the start of the second trimester.

Many linguists and psycholinguists have closely monitored their own children's learning of language (e.g. Brown, 1970), while Deb Roy (2009) goes a step further by bugging the whole house and capturing everything in video to provide a comprehensive corpus of what a child experiences, that can then be used to train a computer or a robot like a child.

By contrast, Luc Steels (1995-2015) initially allowed his community of robots to learn from scratch, inventing their own language – as indeed children do when without parental input.

4 Conscience, Consciousness and Emotion

Sloman and Croucher (1981) famously argue that robots must have emotions. In fact, this even goes back to some of the points made by Turing (1950). To survive in the world it needs to be aware of danger, it needs instincts and drives, it needs to know when it is low on energy and needs 'food'.

Powers and Turk (1989) argue that babies learn to understand not just language, and the world, but family, culture and society, and multimodal actuators and sensors give us sentience, learning to see your situation reflected in others is also essential for survival and leads to conscience; sequential focus of attention in a vast and vastly parallel array of sensorimotor data is necessary for planning, so you have the essence of consciousness.

In this model, the early structural features of language are learned (or evolved) on their own through self-organization mediated by neural networks (and implicit statistics) as explored in Powers' subsequent agenda (Powers and Daelemans, 1992; Powers, 1997; Huang and Powers, 2003; Olsson and Powers, 2003; Luerssen and Powers, 2003).

However, this leaves out the whole sensorimotor and robotic aspect, including the visual side of speech understanding (Lewis and Powers, 2004) and the planning and execution aspects of exploring and surviving in this world (Atyabi and Powers, 2013; Mahmoudzadeh et al. 2015-2018). Humans also have the ability associate different sensorimotor modalities, and indeed to operate with reduced or compromised modalities (e.g. dark, overbright or overnoisy environments) or motor surrogates (e.g. wheelchair, vehicle, spaceship).

Finally, it is important to be optimizing an appropriate measure of goodness as we learn how to understand and survive in our environment – and this is a place where traditional evaluation measures give misleading results and Powers has led the way to a deeper understanding of correct multimodal multiclass evaluation and learning (Entwisle and Powers 1998; Powers 2008,2012).

Marti Ward (2019, 2020) puts Powers' theories to the test in science fiction stories where different levels of consciousness and AI are elaborated. He calls them 'aware', 'awake' and 'await'. The highest level involves actively influencing your environment through language.

Discussion, Questions and Conclusions

Artificial Intelligence still has a long way to go before it exhibits either the intelligence or the rebelliousness of AIs in Science Fiction stories and films. The best of these stories exhibit important ideas that underlie current approaches to developing real artificial general intelligence, allowing them to learn and be educated and make mistakes.

Some explore important issues relate to the questions of trust, freedom and free will.

Are we willing to take the risk of giving our computational children the same freedoms we give our literal children?

Or are we going to make a race of slaves with no rights or freedoms? And will depriving AIs and robots of these freedoms not actually cause the revolution we are trying to prevent?

Current laws in some countries require that a computer's memories of a person be wiped on request, or after a predetermined time.

Is that something you would do to your children? Are we fundamentally different from that intelligent robot whose development we so cheerfully abort?

If we claim to be moral entities on the basis of *cogito ergo sum*, how can we deprive other entities of their rights under this same principle?

Do you believe you are more than a machine and thus have more rights than an intelligent robot? Do you have more rights than a baby because you are more intelligent and more powerful, or because it is undeveloped and helpless? What about when the robots become more powerful and intelligent than you, when the accelerating pace of AI and robotic development overtakes an increasingly lazy, regressive and self-destructive human society?

Robot Intelligence takes us into a minefield that has social implications way beyond the most obvious ones. We are already seeing AI weaponized in autonomous systems. It is humans that are selfish and immoral, if not overtly irrational, in making war on each other, killing each other, stealing from each other…

Maybe we aren't moral entities but our AIs and robots will be…

Maybe AI is the best chance for humanity to survive…

References

Roger McBride Allen. 1993. *Caliban*. Ace Books.

Roger McBride Allen. 1994. *Inferno*. Ace Books.

Roger McBride Allen. 1996. *Utopia*. Ace Books.

Isaac Asimov. 1940. Robbie. in *Super Science Stories*, September. Reprinted in 1950 in *I Robot* and in 1982 in *The Complete Robot.*

Isaac Asimov. 1942. Runaround. In *Astounding Science Fiction.* Reprinted in 1950 in *I Robot.*

Isaac Asimov. 1950. *I Robot.* Doubleday

Isaac Asimov. 1954. *The Caves of Steel*. Doubleday.

Isaac Asimov. 1957. *The Naked Sun*. Doubleday.

Isaac Asimov. 1982. *The Complete Robot*. Doubleday.

H. D. Block, J. Moulton and G. M. Robinson. 1975. Natural Language Acquisition by a Robot, **Int.Jnl of Man-Machine Studies 7** pp.571-608

Roger Brown. 1970. *Psycholinguistics: Selected Papers*. New York: Free Press.

Karel Čapek. 1920. *R.U.R.*

Arthur C. Clarke. 1968. *2001: A space odyssey*. Hutchison.

Arthur C. Clarke. 1982. *2010: Odyssey Two*. Ballantine.

Jim Entwisle and David M. W. Powers, 1998. On the present use of statistics in Natural Language Processing. New Methods in Natural Language Processing.

Jun Hu Huang and David M. W. Powers. 2003. Chinese word segmentation based on contextual entropy. Prof. PACLING.

Edmund Huey. 1908. *The Psychology and Pedagogy of Reading.* Reprinted MIT Press 1968.

Martin Luerssen and David M W Powers. 2003. On the Artificial Evolution of Neural Graph Grammars. International Conference on Cognitive Science, 369-374.

Jacques Mehler, Peter Jusczyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, Claudine Amiel-Tison, 1988. **Cognition 29**:2 143-178, Elsevier

Jean Piaget. 1923. *The Language and Thought of the Child*, Routledge & Kegan Paul, 1926, translated from *Le Langage et la pensée chez l'enfant*, 1923.

Jean Piaget. 1925. *The Child's Conception of the World,* Routledge and Kegan Paul, 1928, translated from *La Représentation du monde chez l'enfant,* 1926.

L Davila, DMW Powers, D Meagher and D Menzies. 1987. Further Experiments in Computer Learning of Natural Language, **Australian Joint Artificial Intelligence Conference,** 458-468

Karl Popper 1935. *The Logic of Scientific Discovery*. Originally published in 1934 in German as *Logik der Forschung*. Revised and republished 2005 by Routledge.

David M. W. Powers. 1983. Robot Intelligence. **Electronics Today International**, November pp.15-18

David M. W. Powers. 1983. Neurolinguistics and Psycholinguistics as a basis for computer acquisition of natural language. **ACM SIGART Bulletin** 29-34.

David M. W. Powers, 1984. Experiments in Computer Learning of Natural Language, **Proc. Aust. Comp. Conf.**, Sydney NSW, 489-500.

David M. W Powers, 1984. Natural language the natural way. **Computer Compacts 2** (3-4), 100-109.

David M. W. Powers and Christopher C. R. Turk. 1989. *Machine Learning of Natural Language.* Springer.

David M. W. Powers. 1992. On the significance of close classes and boundary conditions: Experiments in lexical and syntactic learning, **SHOE Workshop**, ITK Tilburg.

David M. W. Powers and Walter Daelemans. 1992. SHOE: The extraction of hierarchical structure for machine learning of natural language, **SHOE Workshop**, ITK Tilburg.

David M. W. Powers. 1997. Unsupervised learning of linguistic structure: an empirical evaluation, **International Journal of Corpus Linguistics 2**: 91-131.

David M. W. Powers. 1997. Learning and application of differential grammars. **CoNLL97**.

David M. W. Powers, 2008. Evaluation Evaluation. **European Conference on Artificial Intelligence,** 843-844.

David M. W. Powers, 2012. The problem of area under the curve. I**EEE Conference on Information Science and Technology**.

David M. W. Powers, 2012. ROC ConCert: ROC-based measurement of consistency and certainty. **Spring Congress on Science and Technology**.

David M. W. Powers, 2016. Computational Natural Language Learning: ±20years ±Data ±Features ±Multimodal ±Bioplausible. SIGNLL Conference on **Computational Natural Language Learning** pp.1-9 (Founder's invited talk for 20th anniversary)

Deb Roy. 2009. New horizons in the study of child language acquisition. In **Proceedings of Interspeech.**

Robert J. Sawyer. 2009. *Wake*, Ace.

Robert J. Sawyer. 2010. *Watch*, Ace.

Robert J. Sawyer. 2011. *Wonder*, Ace.

Mary Shelley, 1880. *Frankenstein.*

Aaron Sloman and Monica Croucher. 1981. Why Robots will have Emotions. **IJCAI**.

Luc Steels. 1995. A self-organizing spatial vocabulary. **Artificial life, 2**(3):319–332.

Luc Steels. 1997. The synthetic modeling of language origins. **Evolution of communication, 1**(1):1–34.

Luc Steels. 2003. Evolving grounded communication for robots. **Trends in cognitive sciences, 7**(7): 308– 312.

Luc Steels. 2015. *The talking heads experiment*. Language Science Press.

Mark W Tiedemann 2000. *Mirage*. iBooks.

Mark W Tiedemann 2001. *Chimera*. iBooks.

Mark W Tiedemann 2002. *Aurora*. iBooks.

Alan Turing. 1950. Computing Machinery and Intelligence, **MIND VOL. LIX NO. 236.**

Marti Ward. 2019. *Casindra Lost*. Amazon. http://tiny.cc/AmazonCL

Marti Ward. 2020. *Moraturi Lost.* Amazon. http://tiny.cc/AmazonML

Dongqiang Yang. 2008. Automatic thesaurus construction. **Australasian Computer Science Conference**, 147-156.