8.2. Philosophical Implications of Emotional Artificial Intelligence in Science Fiction

Dana Svorova

Abstract

Cognitive science, or classical cognitivism, has evolved into a very complex field of study that views the mind as a computing system, spawning other fields of study such as artificial intelligence (AI), robotics, cybernetics, and other sub-disciplines. AI, in its attempt to replicate human behaviour and reproduce mental processes, has been able to create highly sophisticated and intelligent structures suitable for facilitating working and daily life. However, the rapid development of these intelligent structures may endanger humanity as a whole. An "affective computing" AI device has recently attempted to mimic human emotions as well. In fact, robots can now comprehend emotions and the inner processes of the human brain. Despite these remarkable results, robots are still a long way from experiencing authentic human emotions, a fact corroborated by the philosophical concept of embodiment. Sci-fi cinematography offers much food for thought to reflect on all these issues. This is the case with Alex Garland's chart-topping 2014 sci-fi thriller *Ex Machina*, which very effectively illustrates the significant difference between humans and non-human systems, as well as the potential threat the latter pose. Sci-fi has proven to be truly prophetic in this regard.

Key words: artificial intelligence, human emotions, affective computing, potential danger, sci-fi

Sci-fi literature and film can be considered truly prophetic. Writers as diverse as Samuel Butler, Dean R. Koontz, and Isaac Asimov were able to predict future scientific discoveries, accurately describing well in advance the contemporary scenario of artificial intelligence (AI). In large part, their imagination has become reality. Today's sci-fi narratives are inspired by cutting-edge research conducted in

various fields of study, most notably cognitive science. This is a multidisciplinary field of study that focuses on understanding of the processes of human mind, where the latter are described as a computing system or a system of information elaborating algorithms and special codes. Researchers like John McCarthy, Marvin Minsky, Allen Newell, Herbert Simon and Alan Turing have laid the foundations for the birth of AI following Hobbes's idea that "reasoning is like computing" (Haugeland, 1989: 7). In other words, new fields of study such as conventional AI have stemmed from classical cognitivism (Clark, 1991: 117). They try to reproduce human mental and behavioural processes by designing highly sophisticated, intelligent structures that can facilitate people's daily life in and out of the workplace.

AI processes can be compared with animal or human ones. In fact, prototypes of intelligent artificial devices are inspired by inorganic, organic, and natural structures. In some cases, artificial devices are designed as natural or automatic extensions of the evolutionary processes described by Charles Darwin. For example, the features of a hydraulic pump resemble those of a heart, a camera can be compared to an eye, and so on. The "eso-somatic artifacts" of a living non-human entity are the product of the most sophisticated research (Somenzi and Cordeschi, 1986: 11). The natural selection model that progresses by trial and error has widened by incorporating the inorganic world. Concepts such as occasional changes of structure, species learning, or phylogenic accumulation of knowledge now have to be taken into account in designing AI systems. The human desire to overcome limitations along with an awareness of technology's backwardness when compared with nature's creativity have led to the design of highly sophisticated structures based on the model of natural selection as envisaged by Darwin. The conception underlying Darwin's automatons greatly differs from conventional AI programmes or artificial life (Edelman, 1995: 310–30). The hypothesis of a self-aware artificial system raises many philosophical questions, such as those pertaining to consciousness, a centuries-old problem in the conceptualization of the human. However, we cannot rule out the possibility of a rapid and successful evolution of AI, in which Darwinian automatons will overcome humans (Buttazzo, 2002: 16).

Emotional Artificial Intelligence

A new branch has emerged within AI called "affective computing", a technology that focuses on the artificial replication of human affects. It focuses on the artificial reproduction of human emotions. As early as 1973, the US writer Dean R. Koontz published a sci-fi novel entitled *Demon Seed*, in which a computer is able to experience typically human emotions such as anger, jealousy, love, desire

(Koontz, 2009). In 1973, the novel was classified as a sci-fi narrative, but nowadays its plot has become realistic. Thanks to affective computing, a machine can now process some human emotions. Emotional AI is a new, multidisciplinary field of study embracing subjects such as engineering, neurosciences, and behavioural psychology. Affective computing is based on a knowledge of informatics and human behaviour, and opens up new areas for research (Picard, 1995: 1). The neurosciences have offered evidence that emotions play an important role in human reasoning and decision-making processes. Consequently, to be working effectively, interactive machines should be able to recognize and express emotions. Rosalind Picard, founder of the Affective Computing Research Group and MIT Director, writes that "[a] quantum leap in communication will occur when computers become able to recognize and express affect" (1995: 4). In 1995, Picard presented an ambitious project on this subject. Thanks to extensive research, Koontz's fantasy has became possible. Robots can now recognize the state of mind of their human interlocutors by reading their body language, measuring their heartbeat and the temperature of their body, and providing a response to the information gathered. Robots turn down the corners of the mouth to communicate displeasure, or lower the eyes to communicate guilt, but have not yet being able to blush or get goosebumps from intense pleasure. In sum, evolutionary and proactive robotics have become intelligent systems that interact with the environment and their bodies. Robots are wired in to simulate empathy and emotions, but authentic feelings of happiness, sadness, and bliss are radically different (Picard, 1995: 12–14). Robots lack that typically human component called authentic emotion.

The field of affective computing is a critical new research area that needs to be explored, and can contribute to advances in emotion and cognition theory while greatly improving human-machine interaction. Robots use a sort of "limbic system" to replicate and recognize human emotions and then to respond accordingly thanks to connections between different parts of the system (Benedetti, 2020). Many different technologies are used to this purpose, some of which have been around for many years, like openSMILE, open CV, MARY Text to Speech System, project SEMAINE, ARIE VALUSPA, and others (Benedetti, 2020). They enable developers to create a virtual personality that can interact with humans for an extended period of time while responding appropriately to non-verbal cues from an interlocutor. Sophia is the best-known humanoid robot capable of interacting with humans. She was created by the Hong Kong-based brand Hanson Robotics. Sophia acts very realistically, being able to smile and cry as well as to feel fear and jealousy (Hanson Robotics, 2021). Her posture, movements, and expressions are extraordinary. In 2017, she was granted Saudi Arabian citizenship, becoming the first robot-citizen.

The Italian scholar Giorgio Carlo Buttazzo has observed that recent advances in informatics have influenced the characteristics of robots as described by modern sci-fi. For example, the theory of connectionism and the artificial neural network have inspired several Darwinian robots that learn from experience, have consciousness, can communicate emotions, interact with humans using some kind of cunning, and can evolve on their own (Buttazzo, 2002: 16–17). They do not, however, have the ability to perceive living processes in their bodies. The latter capability, which is a specifically human trait, is central to contemporary sci-fi film, populated by autonomous and self-sufficient robots incapable of authentic feelings but longing to experience them.

Affective Computing in Film

The most appreciated and best-known movie addressing this issue is a sci-fi psychological thriller *Ex Machina*. The movie, directed by Alex Garland, was released in 2014 and received mostly positive reviews, also winning an Academy Award for visual effects as well as a nomination for best original screenplay (IMDb, 2022).

The plot centres around a secret experiment conducted by the young researcher Nathan Bateman, which consists in testing the intelligence of a humanoid robot using the Turing Test. To this purpose, a young programmer named Caleb Smith has been selected. The Turing Test assesses whether a machine can think (Picard, 1995: 3). It consists of an interaction between a robot and a human in which a question-and-answer method is used to determine whether a robot is equipped with intelligence and consciousness (Turing, 1950). The story of *Ex Machina* takes place in a isolated house-laboratory fitted with a security system designed to give an alert to external threats of intrusion and provide safe escape routes. Ava is the name of the intelligent humanoid robot with which Caleb interacts. Her body is robotic but her face, hands, and legs are human-like (Heffernan, 2019: 127–40). Caleb is smitten by Ava. He is taken aback by her flawless command of the English language, creativity, and seductive prowess. Also, she can express her emotions. The two odd interlocutors fall in love. Meanwhile, Nathan has been scrutinizing every aspect of their interaction (Parker, 2015). Ava is terrified of being replaced by a new model of a humanoid robot, so she begs Caleb to save her by assisting her escape from the secret lab. She also wants to experience freedom and feel the sun on her robotic "skin". Caleb promises her that they would run away together and start a new life like free persons. To accomplish this goal, they devise a plot against Nathan. Helped by another robot, Ava kills Nathan, who, before dying, has activated the security system, which blocks all

doors in the building. Ava watches Caleb, who is still trapped inside the building, with an impassive expression. She has found a way out using Nathan's identity card and has left the building without looking back. Ava has thus revealed her true nature: she is merely an intelligent humanoid robot devoid of empathy and compassion (Constable, 2018). *Ex Machina* has several philosophical implications and can be viewed as a visionary and extraordinary story depicting the potential threats posed by the rapid development of Darwinian AI and emotional AI. If AI surpasses humanity in intelligence and becomes "super-intelligent", it may becomes difficult or impossible to control.

Conclusion

In current scientific debate it has been highlighted that the rapid evolution of AI will expose humanity to a number of dangers. For example, Stephen Hawking, Elon Musk, and Bill Gates have pointed out that the development of machines lead to various risks because a "strong" AI evolves more quickly than biological structures. According to Hawking (2018: 160–75), AI can also endanger the internet's ecosystem, which is a precondition of everyday life (Scharf, 2015). But the most dangerous aspect of all is the absence of human intelligence (ANSA, 2018; Whigham, 2018). John Searle (1980) points out that, "according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind". Indeed, strong AI has enabled machines to develop extremely complex capabilities such as abstract thinking, imagination, and creativity. To paraphrase Searle, if appropriately programmed computers are minds, their emotions are not yet authentic feelings, but rather the result of sophisticated algorithms. Automatons can replicate human emotions, but these are not true biological processes. Emotions and sensations are now known to be the result of a bi-directional interaction between body and mind, the result of extremely complex neurochemical processes that have yet to be fully described (Damasio, 2006: 155–60). Perhaps in the near future, it will be possible to encounter an *embodied* machine that is indistinguishable from a human. Is this, however, a desirable goal? The substantial advancements in AI have raised many concerns that someday could result in human extinction or some other irrecoverable global catastrophe.

Works Cited

ANSA, (2018), "L'intelligenza artificiale può diventare una minaccia?", 4 November, online at: <https://www.ansa.it/canale_scienza_tecnica/notizie/tecnologie/2018/11/02/lintelligenza-artificiale-puo-diventare-una-minaccia-_264051f8-5e55-407a-bd8f-393fcf2b3c29.html> [accessed 21 September 2022].

Benedetti, Andrea (2020), "Emotional AI, ecco chi studia le emozioni dei robot", *AI4Business*, 13 November, online at: <https://www.ai4business.it/intelligenza-arificiale/emotiona-ai-ecco-chi_studia-le-emozioni-dei_robot/> [accessed 21 September 2022].

Buttazzo, Giorgio (2002), "Coscienza artificiale: Missione impossibile?", *Mondo Digitale* 2002.1, online at: <http://retis.sssup.it/~giorgio/paps/2002/aica02.pdf> [accessed 21 September 2022].

Clark, Andy (1991), *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, MA: The MIT Press.

Constable, Catherine (2018), "Surface of Science Fiction: Enacting Gender and 'Humanness' in *Ex Machina*", *Film-Philosophy* 22.2, pp. 281–301; <https://doi.org/10.3366/film.2018.0077>.

Damasio, Antonio (2006), *Descartes' Error: Emotion, Reason and the Human Brain*. London: Vintage.

Edelman, Gerald (1995), *Darwinismo neurale. La teoria della selezione dei gruppi neuronali*. Turin: Einaudi [1987].

Hanson Robotics (2021), "Sophia", online at: <https://www.hansonrobotics.com/sophia/> [accessed 21 September 2022].

Haugeland, John (1981), *Mind Design. Philosophy, Psychology, Artificial Intelligence*. Cambridge, The MIT Press.

Hawking, Stephen (2018), *Brief Answers to the Big Questions*. London: John Murray.

Heffernan, Teresa (2019), *Cyborg Futures: Cross-disciplinary Perspectives on Artificial Intelligence and Robotics*. Basingstoke: Palgrave Macmillan Cham.

IMDb (2022), "Ex Machina (2014)", online at: <https://www.imdb.com/title/tt0470752/> [accessed 21 September 2022].

Koontz, Dean (2009), *Demon Seed*. New York: Berkley Books [1973].

Parker, Laura (2015), "Human After All: *Ex Machina*'s Novel Take on Artificial Intelligence", *The Atlantic*, 15 April, online at: <https://www.theatlantic.com/entertainment/archive/2015/04/ex-machina-and-the-virtues-of-humanizing-artificial-intelligence/390279/> [accessed 13 August 2022].

Picard, Rosalind (1995), *Affective Computing*. Perceptual Computing Section Technical Report No. 321. Cambridge, MA: MIT Media Laboratory.

Searle, John (1980), "Minds, Brains, and Programs", *Behavioral and Brain Sciences* 3.3, pp. 417–57.

Scharf, Caleb (2015), "L'intelligenza artificiale è una minaccia?", *Le Scienze*, 21 February, online at: <https://www.lescienze.it/news/2015/02/21/news/intelligenza_artificiale_minaccia_ia_evoluzione_internet-2493790/> [accessed 11 November 2022].

Somenzi, Vittorio, and Roberto Cordeschi (1986), *La filosofia degli automi. Origini dell'intelligenza artificiale*. Turin: Bollati Boringhieri.

Turing, Alan M. (1950), "Computing Machinery and Intelligence", *Mind* 59.236, pp. 433–60; <https://doi.org/10.1093/mind/LIX.236.433>.