Artificial Moral Agents: AL.Gorithm v RP (real people)*

Patrick McNutt FRSA

www.patrickmcnutt.com and Follow @tuncnunc

This presentation builds on earlier work in progress with neotenic data patterns where algorithms are *emotionally connected* to code. Further arguments in support of interpreting data as *data with emotions* involved a game between online consumers (henceforth 'onsumers') and a personalized 'act-like' a human Al. Gorithm.

The related materials are available on my webpage.
http://www.patrickmcnutt.com/news/neoteny-something-abstract-is-thinking/

Please note that the mathematics and the mathematical philosophy behind the narrative are work in progress. Please do contact the author with comments on LinkedIn or on my personal webpage.

*Preliminary speaking notes by Patrick McNutt to accompany the Presentation at Regal Hotel, Hong Kong, September 20th and at AMBS Shanghai Centre, Nanjing Road, Shanghai, October 9th 2019 under the auspices of the Manchester Business School.

Hypothesis: For biotechne algorithms to evolve as artificial moral agents (AMA), algorithms would have to be programmed with an 'act-like' real people (RP) set of ethical values. The ethical values can only be processed within a 'de-self' coded pattern of memory behavior. This short narrative is a dialectic on robots and ethics. Real people, you and I, through our isolated digital patterns of behavior, are becoming more self-reliant as humans but more integrated with sufficiently intelligent algorithms, (AL. Gorithm).

Sentient Robots

We are familiar with robots in science fiction, collectively referred to here as the *biotechne* algorithms[1] in our presentation. From the sentient computer HAL in *Space Odyssey* (1968) to the triptych of 'replicants, facsimiles of humans & born humans' in *Blade Runner* (1982, 2017). In the literature on robots and ethics there is critical discussion on whether or not robots can be programmed with the ethical values of real people (RP). The artificial moral agents, AMAs, would be programmed to 'act-like' a human real person, RP, with an ethical code of conduct. Science fiction writers such as Philip K. Dick, HG Wells, Arthur C Clarke and Isaac Asimov created the fiction of humanoid robots so we explore the perception that robots do evolve with human-like emotions such as love, honesty, trust, and even, terror.

The sentient computer HAL of the twentieth century and the twenty-first century humanoid Pepper and the android Erica affirm this perception of the biotechne algorithm that not only 'thinks' like a RP but evolves with human-like emotions. But are they thinking in the absence of an ethical boundary? In this presentation we explore the boundary of an ethical framework within the contours of a game design. To possess an ethical compass, algorithms have to become responsible. In other words, AL. Gorithm needs a conscious and wisdom. Towards that end, given that AL is programmed to replicate human behaviour and adapt human patterns of behaviour to act-like a RP, a robotic conscious, paradoxically, requires RP to de-self. We argue that this presents a hurdle. It is rational for RP to cheat, to be dishonest and exploit trust. If RP's behavior is framed within the classic

[1] [1] Borrowed from Adrienne Mayor (2018): *Gods and Robots* Princeton University Press

Prisoners' dilemma there is an incentive to cheat, betray trust, act dishonestly. Once cheating is detected, both players realise that each is better off by co-operating. So now we ask: are RPs *consciously aware* of a game? Is there[2] a Gestalt switch? If so, at what node in the coded pattern of behavior does the RP perception of the game change?

Meaning of 'de-self'

To find an answer, the classic Prisoners' dilemma allows us to frame a game design in our search for the meaning of de-self. It is not empathy, not altruism, not co-operation *per se* but a more basic 'zero transaction costs' co-existence. It is as if the categorical imperative of the Kantian equilibrium (KE) is attainable from the data patterns. But it is not a self-enforcing equilibrium outcome. To act-like a RP who cheats then it is rational for AL to cheat. Conversely to act-like a RP who de-selfs it is rational for AL to think as a player at a programmed KE. In this presentation the KE is a payoff-dominant Nash equilibrium, a point of zero transaction costs co-existence displaying altruism, honesty, trust, co-operation, no regrets..

(Sufficient) Corollary: At each node of the data pattern if AL's behaviour is equivalent to the ethical behaviour of a RP and if a RP's behaviour is equivalent to the ethical behaviour of AL, then AL and RP are *isomorphic* and at that point AL. Gorithm is a AMA.

No Meaning No Conscious

Artificial machine intelligence (machine learning, deep learning) ultimately relies on real people (RP) patterns and inference. Algorithms 'think' in real time and robots adjust to changes in their programmed environment. To have a conscious AL has to satisfy 'the big equation', that is, to be programmed with a gesture, a code, if you will, of human understanding, knowledge, reasoning and wisdom. We explore 'the big equation' as a cognitive awareness (visual) experiment[3] in order to reveal a link between meaning and conscious and to demonstrate with cognitive simple visual aids, an axiom of no meaning no conscious. The axiom holds unless and

[2] In other words does the RP perception of the game change when they realise that they are bidding against themselves. Is there a Gestalt shift in online behaviour once they realise that BIN < END? Or is the time spent searching online or the monetised value of personal data 'cookied' to their parties?

[3] By looking at the Winograd semantics and perception in terms of Kanizsa patterns.

until AL. Gorithm as an artificial moral agent (AMA) connects human thought and emotions. Ultimately for biotechne algorithms to evolve as AMAs, AL would have to be programmed with an 'act-like' RP set of ethical values processed within a 'de-self' pattern of memory behavior.

(Necessary) Corollary: For biotechne algorithms to evolve as AMAs, AL would have to be programmed with act-like RP ethical values with a coded 'de-self' pattern of behavior integrating eidetic, emotional conscious memories subject to a non-empty well-defined core of the big equation.

Neural Networks

The convolutional neural networks (CNN) do specialise in processing data by translating an image into a binary set of visual data. Facial recognition software, for example, builds on the architecture of CNN. Nonetheless, programmed codes, however mathematically robust, will not acquire RP wisdom. Data patterns have no meaning for AL. No meaning, no conscious. However, data patterns of neotenic behavior as measured and programmed and coded into an abstract 'thinking' biotechne algorithm will continue to inform the feedback loop of AL. Gorithm. But until a gesture of human understanding, knowledge, reasoning and wisdom can be ascribed to AL there is unlikely to be a robust ethics foundation in our robot age.

Prognosis

In other words, programming with ethical values is subject to a non-empty well-defined core of what we have called 'the big equation', the metric equivalent of a gesture of human understanding, knowledge, reasoning and wisdom. A game design with ethically coded AMAs for the robot digital age requires human RP to de-self. There are 2 players: RP v Al. Gorithm, there is an asymmetric game, and if both players believe that they are playing PD in the online digital space, RP's optimal strategy is to obtain a zero transaction costs co-existence, that is, to de-self. The ethical boundary of Al. Gorithm and of all future *biotechne* algorithms will evolve from the edges of a 'de-self' pattern of human online digital behavior that is processed, coded and programmed.