The Computer's Conception of the World: Using the Psychology of Piaget and Vygotsky to Prevent Sociopathy in AIs

J.N.A. Brown
LinkedIn Corporation
jnabrown@gmail.com

L. Esterle
Aston Univeristy
l.esterle@aston.ac.uk

As AIs become our partners in more and more real-world interactions, it will be increasingly important for them to gain and hold public trust and confidence. Currently, faith in AI, Big Data, and the corporations associated with them is at an all-time low [1]. Even those who embrace technology are concerned that empowering AIs embodied in cyber-physical systems and interacting physically with other machines and with humans could prove very dangerous [2]. We propose that the same processes that enable humans to become empathic, and the same therapies that alleviate human sociopathy could be applied to AIs. To begin, we must accept that, to date, attempts in our field to model self-awareness in AIs have not gone far enough [3]. We propose to improve current models of computational self-awareness with awareness of self and of the self-awareness of others, based on the foundational work of Piaget [4] and Vygotsky [5]. The initial sensory experiences of an infant allow her to begin to form a model of the world. She develops her skills, and learns to apply them in new ways, and also acquires new skills [6]. These skills combine and recombine as she and they develop, enabling her to expand and build upon her experiences and perceptions; improving her ability to sense and to use what she has sensed. In normal psychological development, improved perception and understanding propels her through several radically-different world views, each one allowing her to better cope with the world around her [7]. She starts as the only thing in the universe and, at about 18 months of age, suddenly perceives that the universe is something much bigger; a space that contains more than just her [6]. From there, she slowly develops an understanding that the world exists beyond her direct experiences, and that it is populated by others. This shift from Piaget's "unconscious egocentrism of thought" is not the final step in the development of self-awareness in psychologically-healthy humans. The next stages are vital in terms of skill development, for this is where we move from internal "psychogenesis" to "sociogenesis", the ability to learn with and from others. Humans who do not move from simple self-awareness to the awareness that others are equally self-aware, may learn to interact with others, to compete and to cooperate, but they cannot be empathic [8]. They can learn to appear considerate, but they are manipulative and deceptive, and feel no remorse or guilt. They are sociopaths, unable to fully participate in shared social or private relationships, seeing every experience only from their own perspective and are entirely unaware of how their behaviour affects others [9]. Models of self-awareness in AIs usually stop at the stage of individual self-awareness [10]. Do we want to raise AIs to be Machiavellian, narcissistic, or even sociopathic, or do we want to raise them to be sensitive to the perspectives of others? To meet the short-term goals of our field, let us remember that it is usually not the sociopath who works best with others towards individual or shared goals, but those who are innately able to collaborate and cooperate [11]. To meet the rather longer-term goal of successful relations between organic and artificial intelligences in the real world, let us remember that sociopaths see no problem with lying, cheating, fraud or murder, because they consider others as nothing but pawns, dupes, or obstacles. They don't have to be our friends, but when we are working alongside machines powered by AIs, or riding in trains, planes, and automobiles piloted by them, we will want to be able to trust that our well-being and our very existence matters to them. When they are interacting with us in the real world, we will want our new partners to be mentally-stable and well-socialized, or at least capable of consideration and sympathy. We are delighted to say that this new model and the theories underlying it open up two new fields that are as old as the fiction of Isaac Asimov [12]; the fields of *AI Psychology* (for the careful rearing of healthy AIs) and *AI Psychotherapy* for the remediation of AIs that are suffering from the effects of having been either abused [13] or taught to be abusive [14].

[01] P Andras, L Esterle, M. Guckert, TA Han, PR Lewis, K Milanovic, T Payne et al. "Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems." *IEEE Technology and Society Magazine* 37, no. 4 (2018): 76-83.

[02] D Amodei, C Olah, J Steinhardt, P Christiano, J Schulman, and D Mané. "Concrete problems in AI safety." *arXiv:1606.06565* (2016).

[03] KL Bellman, C Landauer, P Nelson, N Bencomo, S Götz, PR Lewis, L Esterle. Self-modeling and Self-awareness. In *Self-Aware Computing Systems*. Sringer, 279-304

[04] J Piaget (2005) *Language and Thought of the Child: Selected Works, vol 5*. Routledge.

[05] LS Vygotskiĭ. *Thought and Language*. MIT Press, 2012.

[06] J Piaget, G Vakar, and E Hanfmann. Comments on Vygotsky's Critical Remarks Concerning "The Language and Thought of the Child", and "Judgment and Reasoning in the Child" (E Hanfmann & G Vakar, trans.). MIT Press, 1962.

[07] LS Vygotsky *Thought and language* and *Judgment and Reasoning in the Child* (E Hanfmann & G Vakar, trans.). (1962).

[08] L Esterle, and JNA Brown. "Levels of Networked Self-Awareness." In 2018 *IEEE 3rd International Workshops on Foundations and Applications of Self* Systems (FAS- W)*, pp. 237-238. IEEE, 2018.

[09] PR Perez (2012). The etiology of psychopathy: A neuropsychological perspective. *Aggression and Violent Behavior*, 17, 519-522.

[10] L Esterle and JNA Brown (under review) "I Think Therefore You Are: Models for Interaction in Collectives of Self-Aware Cyber-physical Systems". In ACM Transactions on Cyber-Physical Systems…

[11] S Baron-Cohen. *Zero Degrees of Empathy: A new theory of human cruelty*. Penguin UK, 2011.

[12] I Asimov. *I, robot*. Vol. 1. Spectra, 2004.

[13] M Brundage, S Avin, J Clark, H Toner, P Eckersley, B Garfinkel, A Dafoe et al. "The malicious use of artificial intelligence: forecasting, prevention, and mitigation." *arXiv:1802.07228* (2018).

[14] J Zou and L Schiebinger. "AI can be sexist and racist---it's time to make it fair." *Nature* 559 (2018): 324-326