Review of: "[Essay] Not Quite Like Us? — Can Cyborgs and Intelligent Machines Be Natural Persons as a Matter of Law?"

John Bishop[1]

1 Goldsmiths College, University of London

Potential competing interests: No potential competing interests to declare.

I thoroughly enjoyed Daniel Gervais' paper and, as I am broadly in agreement with his arguments and conclusions, I have little to add to those. In contrast, my aim herein is merely to draw the author's attention to other ideas; work that I believe is germane in finessing the "difference in law between human and machine":

1. ***On robots becoming 'more human-like' and 'humans becoming more robotic'.***

   Selmer Bringsjord's 1992 volume "*What robots can and can't be*" explicitly addresses the case of robots passing the Turing Test and becoming ever more 'human-like' and humans having implants and becoming ever more 'robotic':

   Cf. "*The gradual blurring of the traditional boundaries between human persons and robots (or androids). And the position to be advocated below is that though this blurring will happen, significant differences between robots and humans will persist*", (Bringsjord, S., (1992), *What robots can and can't be,* p.4. Springer: Studies in Cognitive Systems).

   At the heart of this position, Bringsjord critiques the notion that robots, qua computational systems (i) can be 'consciously aware of emotion'; (ii) have 'free will' and (iii) can make 'moral decisions'.

   More recently, Selmer revisits and develops this position in Bringsjord, S., (2007), *'Ethical robots: the future can heed us*", AI & SOCIETY: 22 , pp. 539–550 (Springer 2008), wherein he emphasises that a robot can only perform actions as a result of its program. For example, a robot might perform both 'moral' and 'immoral' actions by, say, holding onto, or dropping, a ball 'representing Earth'. However, the robot is unable to choose between such actions on the basis of its own 'morality', as the individuals programming and controlling the robot are ultimately the ones who made those decisions. Furthermore, even if, to enable the robot to do something "unexpected' with respect to its programmers, 'random factors' were included in its software, its subsequent actions would merely be 'determined by the random factor'; certainly not 'freely chosen' by the machine.

   In an online article posted on the Montreal AI Ethics Institute website, Florence Simon discusses Bringsjord's article

[ibid] in the context of J.P. Sullins, who argues for 'the moral agency of robots'. In her article "*The Nonexistent Moral Agency of Robots – A Lack of Intentionality and Free Will*", Florence suggests that robots are not, and cannot be, moral agents fundamentally because "*they lack the* ***intentionality*** *and* ***free-will*** *necessary for* ***moral agency****, because they can only make morally charged decisions and actions based off of what they were programmed to do*".

If Florence is correct, it is axiomatic the personhood implies 'intentionality' and 'free will' and if Bringsjord is correct, no robot, qua computation, can ever exercise 'free will'.

2. ***But could a robot, in its physical interactions with the world, ever have genuine intentional states?***

Although Gervais mentions, and briefly dismisses Searle's Minds, Brains, and Programs, I Suggest the work warrants closer attention. Stemming from the work of Brentano and Husserl for Searle the term "intentionality" refers to the property of certain mental states and events that are directed towards or about objects and states of affairs in the world. This property is limited to mental states like desires, fears, hopes, beliefs, or any other state that refers to something. In essence, intentional states of mind are those that attend to objects. Such states of intentionality are an outcome of, and are fulfilled in, structures in the brain. I.e., They are in causal relations with neurophysiological phenomena and are realized in the neurophysiology of the brain. Thus, mental states of intentionality are caused by certain biological phenomena and produce [biological] actions. It can be deduced that an 'intentional action' is a means by which an individual attempts to achieve, or fulfil, the "conditions of satisfaction" of an intention, and is, certainly for Searle, a paradigmatic 'mark of the mental':

"*Intentionality in human beings (and animals) is a product of* ***causal features of the brain*** *I assume this is an empirical fact about the actual causal relations between mental processes and brains It says simply that certain brain processes are sufficient for intentionality*", (Searle (1980, "Minds, Brains, and Programs', Behavioral and Brain Sciences 3 (3): pp. 417-457).

It was in this context that Searle (ibid) outlined his famous "Chinese Room Argument" to demonstrate that:

"*Instantiating a computer program is never, by itself, a sufficient condition of Intentionality*".

If Searle is correct, a robot can never possess genuine intentional states, or generate genuine intentional actions.

2. ***Can a robot be 'phenomenally' conscious?***

… where, 'phenomenal consciousness' refers to the subjective aspect of conscious experience, involving the qualitative and subjective nature of our perceptions, thoughts, emotions, and sensations. I.e., The "What it is like .." aspect of conscious experience that cannot be reduced to or explained by physical or functional properties alone.

In Bishop (2009), "*Why computers can't feel pain*" I expand my original reductio ad absurdum argument (cf. "Dancing with Pixies (DwP)" (cf., Bishop, 2002; Appendix to Putnam (1988)) to show that, '*conceding the 'strong AI' thesis for Program Q (crediting it with mental states and consciousness) opens the door to a vicious form of panpsychism whereby all open systems, (e.g. grass, rocks etc.), must 'instantiate conscious experience*' and hence that '*disembodied minds lurk everywhere*'.

Thus, if we reject panpsychism, we must reject the notion that machines, qua computation, can be [phenomenally] conscious.

NB. For recent discussion of the DwP, Searle and Penrose [on mathematical insight] see Bishop (2021) "*Artificial Intelligence is stupid and Causal Reasoning will not fix it*", Frontiers in Psychology 2021.

1. ***Can a robot exercise genuine teleological actions?***

   Teleological action refers to 'purposeful action' or 'action directed towards a specific goal or end'. Genuine teleological action involves intentional actions that are based on conscious desires, goals, or intentions. Phenomenal consciousness, on the other hand, refers to the subjective experience of conscious awareness, including perceptions, thoughts, and feelings.

   Lacking phenomenal consciousness, we cannot have genuine teleological action *because* genuine teleological action requires intentional action based on conscious awareness. Without phenomenal consciousness, our actions would be 'robotic' - driven purely by instinct, reflex, or unconscious processes, lacking intention or direction. While such actions *might appear* goal-directed, (instantiating, to paraphrase Dennett, an "*as-if teleology*"), they lack the intentional and conscious aspect that characterizes genuine teleological action.

   Furthermore, phenomenal consciousness provides the 'subjective experience of striving towards a goal', which is an essential aspect of teleological action. The subjective experience of effort, attention, and intentionality that is associated with phenomenal consciousness, is crucial for teleological action to be genuine and purposeful (Cf. Nasuto & Bishop, 2011).

   Hence, if a machine cannot instantiate phenomenal consciousness via computation, it cannot instantiate teleological action.

2. **Could an animat - *an autonomous robot controlled by cultures of living neural cells, which in turn are directly coupled to the robot's actuators and sensory inputs* – or an animal with it's actions externally taught, have**

genuine intentional states or understanding?

To investigate the extent to which blending the biological with the computational results in genuine intentionality and understanding, consider these two cases: (a) wherein a mechanical robot becomes 'more human' by having a real, biological brain control its actions and (b) wherein a living animal, in our case a mouse, becomes 'more robotic' by having real life behaviour(s) imposed by external operant conditioning (via optogenetics).

In 2010, at the University of Reading, UK, a team led by Prof Kevin Warwick and my colleague Prof Slawomir Nasuto, built an autonomous robot controlled by cultures of living neural cells, directly coupled to its actuators and sensory inputs (Cf. Warwick et al, 2010); a device one step closer to the physical realisation of the well-known *'brain in a vat* philosophical thought experiment (Cf. Cosmelli & Thompson, 2010).

Conversely, in 2011, Struber et al, describes a study that demonstrated possibility to perform operant conditioning on a mouse, via optogenetics, and hence how to condition mouse behaviour at the whim of the [external] experimenter.

Subsequently, in a paper, co-authored with Professor Nasuto, we argued that the use of (a) effectively resulted in the creation of a "zombie mouse". I.e., An animal with its behaviour conditioned externally, expresses the experimenters' and not the animal's, will, and powerfully demonstrates that providing an appropriate embodiment alone is not sufficient to account for the emergence of meaning, grounding, and teleology, and (b) that without appropriate consistency between brain and environment, and with its behaviour 'externally engineered', Warwick's animat, with no actual understanding or ownership of the externally engineered actions imposed on it, is also, effectively, a "zombie".

Hence, in Nasuto & Bishop (2011), we conclude that technological advancements in blending biological beings with computational systems have not yet provided such systems with either genuine intentionality or understanding.

Prof J. Mark Bishop, London, May 2023.

References:

1. Bringsjord, S., (1992), *What robots can and can't be*, Springer: Studies in Cognitive Systems.
2. Bringsjord, S., (2007), *Ethical robots: the future can heed us*, AI & SOCIETY: 22 , pp. 539–550 (Springer 2008).
3. Sullins, J. P. (2011). *When Is a Robot a Moral Agent?* In Anderson, M. & Anderson, S. L. (Eds.), *Machine Ethics* (pp. 151–161). Cambridge: Cambridge University Press.
4. Simon, F., *The Nonexistent Moral Agency of Robots – A Lack of Intentionality and Free Will*, Montreal AI Ethics Institute website URL: https://tinyurl.com/4byznp35
5. Searle (1980, "Minds, Brains, and Programs', Behavioral and Brain Sciences 3 (3): pp. 417-457.
6. Bishop, J.M., (2002), *Dancing With Pixies*, in Preston, J. & Bishop, J.M., (eds), Views into the Chinese Room, pp. 360-379, Oxford University Press.

7. Bishop, J.M. (2009), *Why robots can't feel pain*, Mind and Machines: 19(4), pp. 507-516.
8. Bishop, J.M., (2021), *Artificial Intelligence is stupid and causal reasoning will not fix it, Frontiers in Psychology*(11)
9. Putnam, H., (1988), *Representation and Reality*, (Cambridge MA: The MIT Press/Bradford Books).
10. Nasuto, S.J. and Bishop, J.M., (2011), *Of (zombie) mice and animats*, in Muller, V.C., (ed.), (2013), *Theory and Philosophy of Artificial Intelligence*, pp. 85-107, (SAPERE; Berlin: Springer).
11. Cosmelli, D. and Thompson, E., (2010), *Embodiment or Envatment? Reflections on the Bodily Basis of Consciousness*. In *Enaction: Towards a New Paradigm for Cognitive Science,* eds. John Stewart, Olivier Gapenne, and Ezequiel di Paolo, MIT Press
12. Warwick, K., Nasuto, S.J., Becerra, V.M. and Whalley, B.J. (2010) *Experiments with an In-Vitro Robot Brain*. Chapter in, Yang Cai, (Ed), *Instinctive Computing*. Lecture Notes in Artificial Intelligence, Springer, Vol. 5987.
13. Stuber, G.D., Sparta, D.R., Stamatakis, A.M., van Leeuwen, W.A., Hardjoprajitno, J.E., Cho, S., Tye, K.M., Kempadoo, K.A., Zhang, F., Deisseroth, K. and Bonci, A., (2011), *Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking*. Nature: 475, pp. 377-380.