Conscious Machines? Trajectories, Possibilities, and Neuroethical Considerations

James Giordano

Neuroethics Studies Program, Edmund Pellegrino Center for Clinical Bioethics; and Interdisciplinary Program in Neuroscience, Georgetown University Medical Center, Washington DC, USA and Human Science Center, Ludwig-Maximilians Universität, München, GER

jg353@georgetown.edu

Abstract

Research in neurally-based machine (i.e. computational) systems is expanding. “Reverse-engineered” models of brain-like structures are viable candidates for developing increasing complexification (via generatively encoded “intelligence”) that could instantiate some form of consciousness – albeit not identical to human consciousness. This essay posits how such trajectories could lead to the iterative development of “machine sentence” and addresses issues of what “machine consciousness” might mean for: 1) the ways that humans regard such machine entities as “beings” and/or “persons”, and 2) philosophical, ethical and socio-legal positions which might need to be adapted to guide and govern human treatment of, and interactions with such entities. Herein, I argue that neuroethics contributes crucial insights and viable tools to any meaningful approach to this topic (in synergy with extant discourse in “robo-ethics”). As the fields of neuro- and cognitive science, and computational engineering become increasingly convergent, so too must the philosophical and ethical approaches that can – and should – be employed to direct what convergent science may create. The speed and breadth of such technological development are such that neuroethical address and engagement of these issues and questions must be equivalently paced and iterative, so as to retain preparatory value.

Interactive Computational Engineering

Research in neurally-based machine (i.e. computational) systems is expanding. Efforts to create machines that incorporate neural-like sensory acquisition, information integration and synthesis, and motor output systems, are both an intuitive and predictable trajectory in the fusion of neural, computational, and robotic engineering (Kasabov 2003; Ganz 2008). The approach relies upon and also provides a very useful set of heuristics. First, it is based upon the “tools-to-theory” heuristics of neuroscience, which has allowed significant progress in understanding the structure and basic functions of neural systems (Gigerenzer 1991; Giordano 2012). Second, this has enabled “theory-to-tools” heuristics (Giordano 2012) that have been actualized in the development and use of a variety of technologies that could be created for these “next generation” computational systems and robotics.

I believe that such “reverse-engineered” neural models of brain-like structures and functions will ultimately be a key to unraveling the enigma of consciousness. This will close the heuristic loop through the re-engagement of a “tools-to-theory” approach. Thus, while humans would create the general template for the neural system and device, it is the tool itself (i.e. the neural system “embodied” in the machine) that would develop techniques and implements to identify the features of its physical system that need to be fortified, modified, or discarded, based upon acquired information about the environment in which it exists, and the tasks necessary to act under changing conditions within such environments. Simply put, the system could acquire a form of “physical intelligence”, and then iteratively adapt its functions and physical features to optimize inputs and outputs to “learn” to maximize performance in the environments in which they operate.

This process, generative encoding (Cheney et al. 2013), in many ways represents a form of autopoietic – or self constructing – system, and operates both developmentally and somewhat “evolutionarily”; first to modify itself (i.e. developmentally), and then to affect others of its type to progressively adapt (i.e. evolutionarily) to ever-more complex levels of information acquisition and use, and activity. On a number of levels, such autofunctional systems might be seen as desirable, because they have
“build them and leave them” qualities. Hence, humans would assume the role of the proverbial “blind watchmaker”: These kinds of systems would learn what would be required to know and do to achieve a set of tasks and goals that humans define for them – at least initially (Fogel 2006).

But, the neural networks and vehicles (viz. embodiments) that humans create for the system may not necessarily be those that the system would (“choose to”) develop for itself (i.e. think about that “special gift” someone gave you on your last birthday – the tie with the purple flying pigs on it, or the lime green blazer with red stitching; often, what others think you need and want tend to reflect their taste and wants more than your own). Recall that what’s being proposed is the creation of real cybernetic entities (in the strict sense of the word – a system of progressively adaptive feed-forward/feedback mechanisms, as defined by Norbert Weiner), with the complex, dynamical features that such systems obtain and entail (Wiener 1965; Wiener and Schade 1965). These systems are very sensitive to initial and changing conditions, and are responsive, and adapt to various attractors and constraints (Corning 1995) – which might not be readily apparent to humans from outside the system. So, the neural system could, and likely would establish its own heuristics for what works and what doesn’t, and employ parts-to-whole (i.e. “bottom-up”) and whole-to-parts (i.e. “top-down”) self-assessments, to provide sort of an “inside out” perspective to “teach” its builders what it structurally requires to optimize its functions.

**Putative Embodiment**

Could these types of systems create environments and “bodies” for themselves? To answer these questions, let’s start a simple (and most probable) system, and then open the discussion to include a more provocative, futuristic vision. In so doing, let’s also establish some basic presumptions about how a paradigm for such physically intelligent systems/machines would be initiated and sustained. A neurally-modeled, physically intelligent system capable of generative encoding would need to possess mechanisms for acquiring data, information, and therefore some type of “knowledge” about both itself (i.e. interoceptive knowledge), and the environments in which it is/would be embedded and engaged (i.e. exteroceptive knowledge). Access to “real-world” environments could provide a medium for acquiring these data, which could be augmented by the richness of information available on the internet. The system could “plug into” 1) the real world via multi-channel sensors for light, sound, tactile, and perhaps even olfactory inputs – at levels that may have very different thresholds than humans (for example, infrared and ultraviolet light, and ultra and subsonic frequencies), and 2) the informational resource of the internet. Taken together, this would provide the system with direct-access data/information, and indirect-access, interpretive and analytic data/information that would greatly augment the amount and type(s) of “knowledge” the system could and would acquire, and be able to use.

Once these basic functions were established, the system could interpret its “real world” environments, and parse massive reams of data on the internet to create a “mosaic” of the information needed to “auto-develop” approaches to optimize its function in the environments it encounters, and those it seeks to engage. This could then be relayed to humans in order to create “systems’ desiderata” – a “needs’ and wish list” – to inform how to structure and/or modify the components of the system to achieve certain tasks, goals and even progressively more expanding ends. The system could bring together a variety of other system components – both of the neural network and the physical structures that provide its inputs and outputs – and “present” these to its human builders as aspects of what would be needed to iteratively fine tune its functions and capabilities. A potential advantage of this approach would be the ability of the machine system to side-step the limitations of human “builders’ biases”, to instead emphasize and exploit the dispositions, requirement, and needs of the system to self-assess and support its own functions.

Taking this a few steps further; there is the possibility that the system could develop and/or evolve the capacity to “re-tool” itself, and in this way attempt to “remove the middle man”, so to speak. The system could “propose” a robotic component that could enable and/or sustain the physical expression(s) of the encoding process. In other words, it could “request” the parts needed for a “building device” that would allow the system to execute physical autopoiesis – more simply put, the ability to build new parts of itself, or construct new systems (without humans necessarily being part of the process). These new parts could be sub-components that synergize the activities of the major (or alpha) system, and in this way establish a multi-tasking “support network”. This is not far-fetched; the capacity for self-regulation is inherent to most, if not all, cybernetic and complex dynamical systems (Corning 1995), and the achievement of certain goals would then feed back to the system and provide an ever-expanding palette of new niches, requirements and tasks – and through successive self re-modeling – the generation of new capabilities. Moreover, this could occur fairly rapidly, as the processes employed by the system for performance optimization might not be bounded by the restrictions of “outside-in” (i.e. human) perspectives.
Limitations

Now before we conjure dystopian visions of “intelligent machines” taking over the world, let’s ground these possibilities to some practical realities. Certainly, there are a number of mitigating factors present. First is that these systems need power, so they’d be dependent upon existing power supplies for their resources. But, it is also possible that such systems, if and when working in synergy, could establish a “divert on demand” mechanism/pathway to provide access to necessary power supplies to sustain function across a range of environmental deprivations. There is current discussion of the likelihood of such mechanisms being generated by cognitive systems as some form of “survival strategy” that almost any self-referential, cognitive system would be likely to develop (Johnson and Omland 2004).

Another limiting factor is that the materials for auto-fabrication would need to be available and acquired if such systems were to attempt to generate physical structures to expand their own functions. So, if humans do not provide these substrates, then arguably, any steps toward autopoiesis would be restrained. Perhaps, but what remains to be seen is if and how such systems could/would learn to fabricate or innovate their own components, so as to create physical adaptations necessary to execute evermore advanced/complex functions. This too is not as much of a stretch as it sounds. A machine system that is modeled after or upon a human neuro-cognitive template could in fact, manifest something of a Bayesian bias toward “tool use”, and in light of this, could learn to use the resources at hand to alter its structure in such ways as to adapt to new environmental challenges and “get the job done”, if not insure its own survival (Johnson and Omland 2004).

Another possible constraint is that the digital nature of machine computation may impose limits upon the freedom with which expanded capability could be realized. Here I assert that a bit of caution is warranted; a neuro-identically modeled system (if not an idealized neurally-modeled system) could rapidly achieve vast degrees of functional freedom by changing the sensitivities to functional thresholds. If the system were, for example, to develop a broad range of discriminations between a no-response value and response value (say, a Planck-based scale between 0 and 1 that involves almost infinitesimally small distinctions between each point value), it could develop very finely-grained capacity and patterns of parsing inputs and outputs and in this way, greatly refine – and expand – its functional repertoire.

Neural systems actually operate in this way: while the net effect in a given neural network may be activity or non-activity, and that of a nerve cell may be “fire/do not fire”, these overall characteristics reflect very finely-tuned, small-scale inputs and outputs (e.g. at various regions of cell membranes, and at large numbers of points of inter-neuronal connections within the neural network) that are graded, and whose spatio-temporal pattern of activity cumulatively summate to produce a “go/no-go” effect (Giordano, Rossi, and Benedikter 2013). So, a system modeled after or upon such neural activity could, or more probably would, function in much the same way, and this might provide the basis for its ongoing complexification.

Machine Consciousness, Ethical Implications

The issue that lurks right over the horizon of such possibilities is whether increasing complexification in generatively encoded “intelligent machines” could instantiate some form of consciousness. I argue that the most probable answer is “yes”; although it is likely that it will not be identical to human consciousness. Prima facie, this would not matter; the question is not what type of consciousness, but whether consciousness exists. (This opens discussion – if not a Pandora’s Box – about whether the occurrence of consciousness in various organisms represents a distinction of “kind” or “degree”; a full treatment of which is beyond the scope of the present paper). At minimum, the system would become auto-referential, and in this way, acquire a “sense of self”. Leaving aside philosophical views of the concept of “self”, at the most basic level this means that the system would develop awareness of its internal state and of external conditions and be able to discriminate between itself and things that are not itself. This is an important step, as it would lead to relationality – a set of functions that provide resonance or dissonance with particular types of (internal and/or external) environmental conditions, reinforcements and rewards for achieving certain goals or states. In this way the system would acquire and obtain a sense of what neuroscientist Antonio Damasio has called “a feeling of what happens”; in other words, a form of consciousness (and self-consciousness) (Damasio 1999).

The question is: what then? Not only as regards the trajectories of such machine systems, but as regards human response and responsibility. What will we do with – and about - this? This prompts derivative questions of if, and in what ways we can be prepared for such new developments (Giordano and Benedikter 2012). Granted, the likelihood of conscious machines – as exciting as it may be – is still years away, even given the most fruitful predictions. Any prediction involves relative probabilities of if, how, and what trajectories such as those described here will occur. There is always an element of chance, and as such, there is also the whiff of the fictional. Arguably, fictional accounts can contribute to – and have a place in – the discourse
(Wurzman and Giordano 2013). Yet, while (well-conceived) fiction may be useful to illustrate social anticipations and anxieties, I offer that rigorously-modeled and pragmatic “what if” speculation about mindful machines are important to 1) developing realistic scenarios – and an appreciation for the lability of the timelines involved in their occurrence, 2) the ‘what about’ questions raised by the nature and implications of scientific and technological development(s), and 3) the ways we form and formulate philosophical beliefs, ethics, policies and laws.

These issues – and the questions they spawn – get particularly complicated given the capacity of neurally modeled machines to self-assess, manifest awareness, and self-develop and/or replicate. Yet, the very fact that there is realistic discussion about moral consideration of and for machines represents a shift in our epistemology and ethical paradigm. The creation of, and discovery that an entity is sentient is not esoteric, but instead means something both about that organism, and the ways that it should be considered. Neurocentric criteria, namely, whether a being manifests the ability for pain/suffering, some form of sentience, and the type and extent of these properties, are arguably important for the way we morally regard – and ethically, legally and socially treat – other beings (Loveless and Giordano 2014a; Ryder 1970; Giordano 2014). Could such criteria sustain a revision in the ontological status afforded to sentient machines (or perhaps even ontological notions generated and maintained by sentient machines themselves)? Might such entities be considered to be “persons”, given recent re-examination of the basis and meaning of the term and construct (White 2007)?

What ethical approach would be best suited to address these possibilities and the issues, questions and problems that they may generate? Fictional accounts (e.g. - Asimov’s Laws) of mindful machines and human-machine engagement(s) have been offered as possible starting points for ethical deliberation (Asimov 1964). Asimov’s fiction is engaging, but as multiple critiques have noted, should not be taken out of context [for overview, see Clarke (1994)]. Instead, any ethical approach must begin from and be based upon fact, and the discipline of ‘robio-ethics’ has been proposed and developed to appropriate the realities of potentially sentient machines (Sawyer 2007; Asaro 2006). There is merit to this; however, it may be that a more fundamental perspective is required that enables a deeper and broader view of the issue(s) and its implications.

I posit that what we are really grappling with is if, and to what extent, machines are conscious (and self-conscious), and what that means for the ways we as humans 1) regard them as “beings”, and 2) direct our treatment of, and interactions with them. In this light, I argue that neuroethics contributes crucial dimensions and insights to any meaningful approach to this topic and its articulation in practice (Loveless and Giordano 2014a; Giordano 2011a; Giordano 2011b; Loveless and Giordano 2014b). Without doubt, some of the questions and problems fall squarely into the domain of basic ethics; yet, others center squarely upon the unique neuroscientific questions – and answers – generated by the creation and advancement of sentient machine entities (e.g.- about the nature and limits of consciousness, morality, the future vision of humanity, etc (Giordano and Benedikter 2012; Benedikter, Giordano, and FitzGerald 2010). This begs questions of what it means to create a sentient entity, what responsibilities arise therefrom, and whether the presence of certain neurocognitive capabilities and characteristics represent not simply consciousness, or even a “person”, but life. Ongoing neuroethical debates regarding the use of neurological criteria to ascertain death and the status of the human embryo reveal and reflect the contemporaneity – and gravitas – of these issues (Gazzaniga 2006). Of course, realistic discussions of machine consciousness are still incipient. But, the speed and breadth of technological development are such that neuroethical address of these issues and questions must be equivalently paced and iterative, so as to retain preparatory value.

Science, technology, and the knowledge and artifacts they produce are human creations; therefore, responsibility to deal with the outcomes of our creations rests in human hands. As the fields of neuro- and cognitive science, and computational and bioengineering become increasingly convergent, so too must the philosophical, ethical and socio-legal approaches that can – and should – be employed to guide and govern what convergent science may create (Giordano 2012). And this harkens back to my musings about conscious machines. It’s provocative, arguably a useful exercise – and perhaps fun – to speculate on what neuroscience and neurotechnology hold for the future (Wurzman and Giordano 2013). But it’s folly not to critically assess what this science holds for the present, foolhardy not to recognize the promise and perils that such science and technology may incur, and frighteningly dangerous not to devote time, effort and resources to studying, and developing ways to be prepared for, and prudently guide each and every step ahead.

Acknowledgments

This work was funded in part by the J. W. Fulbright Foundation, Clark Foundation, and William H. and Ruth Crane Schaefer Endowment.
References


