



Anthropomorphic reasoning about neuromorphic AGI safety

David J. Jilk, Seth J. Herd, Stephen J. Read & Randall C. O'Reilly

To cite this article: David J. Jilk, Seth J. Herd, Stephen J. Read & Randall C. O'Reilly (2017): Anthropomorphic reasoning about neuromorphic AGI safety, Journal of Experimental & Theoretical Artificial Intelligence, DOI: [10.1080/0952813X.2017.1354081](https://doi.org/10.1080/0952813X.2017.1354081)

To link to this article: <http://dx.doi.org/10.1080/0952813X.2017.1354081>



Published online: 19 Jul 2017.



Submit your article to this journal [↗](#)



Article views: 17




View related articles [↗](#)



View Crossmark data [↗](#)



Anthropomorphic reasoning about neuromorphic AGI safety

David J. Jilk^a , Seth J. Herd^a, Stephen J. Read^b and Randall C. O'Reilly^{a,c}

^aeCortex, Inc., Westminster, CO USA; ^bDepartment of Psychology, University of Southern California, Los Angeles, CA, USA; ^cDepartment of Psychology & Neuroscience, University of Colorado at Boulder, Boulder, CO USA

ABSTRACT

One candidate approach to creating artificial general intelligence (AGI) is to imitate the essential computations of human cognition. This process is sometimes called 'reverse-engineering the brain' and the end product called 'neuromorphic.' We argue that, unlike with other approaches to AGI, anthropomorphic reasoning about behaviour and safety concerns is appropriate and crucial in a neuromorphic context. Using such reasoning, we offer some initial ideas to make neuromorphic AGI safer. In particular, we explore how basic drives that promote social interaction may be essential to the development of cognitive capabilities as well as serving as a focal point for human-friendly outcomes.

ARTICLE HISTORY

Received 23 November 2016
Accepted 5 July 2017

KEYWORDS

Artificial general intelligence; neuroscience; neuromorphic AGI; AI safety

Introduction

This paper offers proposals for the study of safety issues in the development of neuromorphic artificial general intelligence (AGI). We are only concerned here with artificial *general* intelligence, not narrow or restricted-domain computational or agentic solutions. Precision in the distinction is not crucial to our purposes, but awareness of this focus may help avoid confusion. We begin by describing what we mean by neuromorphic AGI; our usage is relatively narrow and refers specifically to the results of methods that have already exhibited notable success. These methods rely heavily on the human 'reference implementation' for both progress and assessment. We claim that – in contrast to other proposed approaches to development of AGI – anthropomorphic reasoning is not only appropriate but crucial to consideration of safety issues in neuromorphic AGI. By *anthropomorphic reasoning* we mean reasoning about AGI behaviour and development that is guided by knowledge about human behaviour and development, based on an assumption of deep homologies as elaborated in the next section.

A number of prominent AI researchers have claimed that development of neuromorphic AGI is more dangerous than other approaches. We take this claim seriously, and show that it has not been convincingly demonstrated. We thus conclude that efforts to make neuromorphic AGI safer are not superseded by a clear inherent inferiority with respect to safety. We take the position that all AGI research should proceed with an abundance of care and humility.

With those preliminary elements in place, we proceed to a substantive application of anthropomorphic reasoning to safety considerations in neuromorphic AGI. Throughout, we consider both strategies to directly improve safety as well as research topics that would better inform such efforts. We review motivational processes in humans, and in particular discuss certain basic drives that appear to be integral to the development of intelligence. We identify basic drives that promote positive social interaction,

along with consequent positive social development, as promising targets, and describe how these would be used to create AGI that has a positive relationship with humans. Finally, we consider the dynamics of motivation and behaviour in neuromorphic AGI once it has been established, and use human experience to suggest ways of maintaining stability in representations and motivations.

What is neuromorphic AGI?

*A great design appears at first insane;
But chance will soon seem quaint and blind,
And such an exemplary thinking brain
Will soon by thinkers be designed.*
Johann Wolfgang von Goethe, from 'Faust'

In this section we describe what we mean by 'neuromorphic AGI,' and show that applicable methods rely heavily on our understanding of the human system for both progress and assessment. We conclude by explaining why anthropomorphic reasoning is appropriate in the analysis of related safety concerns.

A variety of methods have been proposed to create AGI. One obvious method is to *reverse-engineer* the only functioning system actually known to exhibit general intelligence, the human brain. This may or may not be the fastest method,¹ but it seems less speculative than others since it is based on a reference implementation. We refer to the intended result of such a method as 'neuromorphic AGI.'

We further narrow the scope of this method by describing it as *discerning, reproducing and integrating all the essential computations of human cognition*. This refinement excludes, for example, approaches such as 'whole brain emulation' (Sandberg, 2014), which includes many functions that are not essential computations; or approaches based on psychological introspection (sometimes called 'Good Old-Fashioned AI' or GOFAI), which exclude essential brain computations to which we have no conscious access. The fact that we discuss a narrower focus implies no claims about the prospects for other approaches. For completeness, we also note that it is an open question whether human cognition is strictly computational, though it is assumed true in most AGI research efforts.

A number of researchers (Eliasmith et al., 2012; Gershman, Monfils, Norman, & Niv, 2016; Grossberg, 1988; Hinton, 1992; O'Reilly, Hazy, & Herd, 2016), base their efforts on this method, sometimes called 'computational cognitive neuroscience.' Progress with the approach demands, among other things: (i) knowledge of findings in neuroscience, cognitive psychology and neural computational methods; (ii) hypothesising mechanisms that might produce known cognitive and behavioural outcomes, despite the fact that such hypotheses are severely underdetermined by existing empirical data; and (iii) a great deal of trial-and-error in simulating hypotheses to identify those that both exhibit the desired cognitive features and admit of a biologically plausible explanation. Recent striking results with convolutional neural networks for vision (Krizhevsky, Sutskever, & Hinton, 2013; Mnih et al., 2015) and audition (Deng & Jaitly, 2015), which arose from earlier work based on this method, lend support to the idea that the approach has promise. Even more tantalising is the successful application of the technique to a wide variety of data analysis and real-time behavioural problems (Schmidhuber, 2015).

Despite the rather arduous and non-deterministic scientific process, it is important to be clear that the end result consists of well-defined computational processes. It contains no quantum mechanics, biochemistry or cellular biology. Neuromorphic AGI, if it is possible at all, will be neither more nor less than a sophisticated and integrated set of computations. Further, to date it appears that these computations, though they may rely on implicit representations, are not opaque or inscrutable. Indeed, experience in the field suggests that it is a prerequisite to discovery and reproduction of the brain's computations that the underlying principles be deeply understood and that considerable mathematical analysis is applied. Even more striking is that the most effective neuromorphic computational approaches are sometimes found to have analogies or even isomorphism to concepts in physics (e.g. Amit, Gutfreund, & Sompolinsky, 1985; Lin & Tegmark, 2016).

Together, these factors suggest that 'cobbling together pieces plagiarised from biology without the engineers necessarily having a deep mathematical understanding of how the system works' (Bostrom,

2014) is unlikely to be successful, and that a coherent computational view of the entire system will be necessary before it exhibits the same level of cross-domain cognitive capabilities as a human brain. Frequent comparison of computations and resulting representations to those found in the reference implementation (the human brain) has been and will continue to be essential.

It is outside the scope of this paper to provide much detail on the computational principles that have been identified or theorised so far; nevertheless it is worthwhile to offer an outline to make clear that our notion of 'how the brain does it' is not a Skinnerian black box. As mentioned above, convolutional deep neural networks have in just the past few years gone from an obscure research topic in neuroscience to a powerful algorithmic approach that seems to solve many hard problems in artificial perception and pattern recognition (Schmidhuber, 2015). These networks perform a hierarchical data transformation sometimes called 'dimensional reduction' that identifies representations through a process that is akin to principal component analysis but more effective (Hinton & Salakhutdinov, 2006). There is considerable and varied converging evidence that much of posterior cortex in the brain, including the visual (Riesenhuber & Poggio, 2002) and auditory (Kell et al., 2015; Okada et al., 2010; Tian, Kusmierik, & Rauschecker, 2013) cortices, play this computational role.

There is also significant evidence that to account for important aspects of human cognition, the overlapping and slowly developing representations in the posterior cortex must be complemented by a fast, discrete, conjunctive system that can reproduce memory states accurately (Eichenbaum, 2004; O'Reilly, Bhattacharyya, Howard, & Ketz, 2011). This functional role is implemented in the brain by hippocampus. The computational processes performed there are quite well understood and characterised, and have been implemented in software (Becker, 2005; Byrne, Becker, & Burgess, 2007; Hasselmo, Bodelón, & Wyble, 2002; Ketz, Morkonda, & O'Reilly, 2013) as well as directly in silicon (Berger et al., 2012). Informally, some researchers think of the hippocampus as representing episodic memories by providing 'conjunctive pointers' into the deeper semantic representations in posterior cortex. Such a computational role has been suggested to improve performance in machine learning more generally (Graves et al., 2016; Pritzel et al., 2017).

Finally, some of the most salient cognitive capabilities of executive function and attention have been characterised by computational interactions within and between the prefrontal cortex and basal ganglia (Hazy, Frank, & O'Reilly, 2007). In brief, the prefrontal cortex learns representations of actions and plans for actions, and selects and maintains these representations based on predictions of reward. It has been proposed that this is one locus of how the brain transforms fuzzy inputs into discrete symbols and actions (O'Reilly, 2006).

We note that the ability to learn and use conceptual representations in the human brain is clearly an essential computation, and would be a capability of any neuromorphic AGI. Though our understanding of conceptual representations, how they arise, and how we access them has substantial gaps, the discrete representations in PFC combined with the conjunctive representations in hippocampus probably play an important role and might serve as a model for this function.

The neuromorphic approach includes consideration of the human system with respect to a variety of levels of description, ranging from mechanics of synaptic plasticity, to the dynamics of specialised brain regions, all the way to human behaviour (Jilk, Lebiere, O'Reilly, & Anderson, 2008). In this specific yet important approach to development of AGI, comparison to the human system is not metaphorical. It is a required step in the process of discerning, reproducing and integrating the essential computations of the brain. Consideration of how the biological system performs computations contributes to *making* progress; comparison of the results of those computations in terms of both behaviour and stored representations are how we *evaluate* progress.

In the broader field of AGI safety, anthropomorphic reasoning is considered inadvisable because the resulting agents may be *nothing like* humans (Bostrom, 2014; Yudkowsky, 2008). These warnings are sometimes extended to neuromorphic AGI, but they do not have the same force here. Because it is extensively derived from and evaluated in comparison to the human reference, the function and behaviour of neuromorphic AGI will overlap in important ways with that reference. To that extent, anthropomorphic reasoning is both applicable and crucial.

Though clearly it is sensible to be cautious about overgeneralising in individual cases, anthropomorphic reasoning about safety issues is highly appropriate in the case of neuromorphic AGI and is synergistic with the development of function.

Comparative safety of neuromorphic AGI

*The great Intelligences fair
That range above our mortal state
In circle round the blessed gate,
Received and gave him welcome there*
Alfred Lord Tennyson, from 'In Memoriam'

Neuromorphic AGI has been suggested as the most dangerous approach to AGI from a safety perspective (Bostrom, 2014; Yudkowsky, 2008). If this claim were convincingly demonstrated, it might be advisable to refrain from discussion of making neuromorphic AGI safer, and instead to urge its avoidance altogether. In this section we show why we do not, at present, find this claim convincing.

The concerns about neuromorphic AGI appear to be based on the purported difficulty of a comprehensive or provable analysis of its input and output behaviour, and on the apparent difficulty of understanding its operation explicitly and comprehensively. Although we can understand the intuitive appeal of this argument, we do not think it stands up to various critical challenges. In short, we argue that full verification of behaviour of any system complex enough to constitute AGI is likely unattainable; similarly, the representations necessary for any AGI are likely to be complex and require significant effort to analyse. Thus, all approaches to AGI are likely to face similar challenges, despite their various differences, and require an abundance of care and humility, along with extensive further research.

Intuitively, the neuromorphic AGI approach may appear to be the most dangerous because of our unique perspective on our own species: we have direct introspective access to our own inner thoughts, and a wealth of experience with both the safe and unsafe aspects of human nature. If we then imagine a novel approximation to ourselves, it is easy to worry that the unsavoury aspects of human nature may somehow dominate in such a 'mutant' system, whereas our ability to envision evil emerging from a system based purely on logic and symbolic operations is much less well developed. This could be seen as a variation on the 'uncanny valley' – artificial systems that are close, but not quite identical, to humans generally evoke negative and uneasy feelings (Mori, 1970). Furthermore, the lack of any direct introspective bond between us and these putative artificial beings weakens the all-important 'golden rule' kinds of social contracts that tend to put a check on the worst aspects of our nature.

Despite the power of these intuitive arguments, we argue that proof of safety is difficult for *any* form of AGI (Jilk, 2016). First, linking the actions of an agent to real-world outcomes is intractable due to the absence of a complete analytic physical model of the world. Second, even at the level of agent actions, determining whether an agent will conform to a determinate set of acceptable actions is in general incomputable. Third, though manual proof remains a possibility, its feasibility is suspect given the likely complexity of AGI, the fact that AGI is an unsolved problem, and the necessity of performing such proof on every version of the code. Worse, manual proofs must themselves be verified, leading to an infinite regress (Yampolskiy, 2016). Fourth, to the extent that examples of proving agentic behaviour are provided in the literature, they tend to be layered architectures that confuse intentions with actions, leaving the interpretation of perception and the execution of actions to neuromorphic or genuinely opaque modules. Finally, a post-processing module that restricts actions to a valid set is marginally more feasible, but would be equally applicable to neuromorphic and non-neuromorphic AGI. Thus, with respect to the desire for safety verification, we see fundamental unsolved problems for *all* types of AGI approaches.

As described earlier, anthropomorphic reasoning is applicable to the analysis of safety issues to the extent the AI in question uses the same computational scheme as the human system. Human behaviour can be observed and partially understood, and simulation behaviour can be compared to

assess whether the behaviour is at least superficially similar. To the extent that the architecture of a neuromorphic AGI is similar to that of human brains, we can compare homologous activity inside each system and map it to behaviour. Thus, we can rely on both external behavioural and internal functional comparisons to assess the safety implications of the operation of a neuromorphic AGI. The availability of this reference comparison suggests that neuromorphic AGI could even have advantages, from a safety analysis perspective, over other approaches that do not have such a reference.

At the present state of the art, claims about which approaches to AGI are safer or more susceptible to safety analysis seem speculative. With each approach, evaluating the means of improving safety will take somewhat different forms. In the case of neuromorphic AGI, anthropomorphic reasoning will play a significant role.

Neuromorphic AGI and motivation

*Only the curious
have if they live a tale
worth telling at all.*
Alastair Reid, from 'Curiosity'

We now proceed to the substantive application of anthropomorphic reasoning to questions of neuromorphic AGI safety. Because of the importance of motivation in behaviour and therefore questions of safety, we begin by examining sources of motivation in humans and how they might be implemented in neuromorphic AGI. These sources include basic drives, some of which appear to be essential to the development of intelligence, as well as derivative-learned motivations.

An agent with the capability to model the world but lacking motivational drives, construed broadly, is evidently stunted. When the agent is switched on, there must be some component that drives actions, even if just a hard-coded stimulus-response mechanism. Otherwise it is unlikely, except by coincidence, to actively increase its knowledge. Its actions in the world, if it takes any at all, will be random and undirected. From a human perspective, such an agent seems neither useful nor strategically dangerous, and arguably it is not even intelligent.

Consequently, creation of neuromorphic AGI will almost certainly require a motivational architecture, and the study of possible architectures and their implications are paramount in both functional and safety aspects of the effort. Such study can leverage our understanding of the human and mammalian motivational architectures. Determination of the computational components of such architectures is subsumed by neuromorphic methods, but non-neural anatomical features also play a role, and the boundary or interface between those components and the brain has not been fully characterised. Among the different sources of motivation that humans have, we do not yet know with certainty which of them are essential to cognition or its development. Some of them may relate to agency and survival, and therefore require consideration of the differing details of embodiment between humans and an AGI. Thus, we analyse and speculate on motivational architecture here, to consider safety implications and strategy, even though these questions will be partially answered through future neuromorphic AGI research. Discussion of the abstract notion of motivation in artificial intelligence often assumes that an agent has a 'final purpose' (e.g. Bostrom, 2014), typically expressed in the form of an objective function. However, it is manifestly not the case that human motivation is so straightforward: even if an objective function could be inferred, it is likely to be complex, high-dimensional and crucially, dynamic (Edelman, 2015).

In mammals, fundamental biological drives are deeply embedded in the very mechanics of learning. For example, satisfaction of hunger can result in elevated dopamine levels in the brain, and the presence of dopamine increases synaptic plasticity (Bao, Chan, & Merzenich, 2001). Such instrumental learning is complemented by associative learning that enables conditioned stimuli to predict rewards (Hazy, Frank, & O'Reilly, 2010). Though the details of these mechanisms in living brains are not fully understood, the computational principles involved are progressively yielding to analysis and modelling (Hazy et al., 2007; Seamans & Robbins, 2010). In any case, the central hypothesis of this theory as

it relates to neuromorphic AGI is that motivated behaviour arises from a direct interaction between hardwired drives or interoceptive factors, and neuromorphic learning mechanisms.

A crucial corollary of this hypothesis is that basic drives are pre-conceptual and pre-linguistic, and indeed their existence is a prerequisite to development of deep semantic representations such as concepts or complex constructions such as sentences and propositions.² While such basic drives can subsequently be *described* linguistically, they are not *determined* linguistically. A statement such as 'manufacture paper clips' (Bostrom, 2003) or 'keep humans safe' cannot, under this model, constitute an original basic drive or even a component of one. Without having first developed sufficiently rich, world-referring representations for the constituent concepts in such statements, an agent does not know what they mean and cannot be motivated by their meaning. In a neuromorphic AGI, 'Asimov's Laws' and the like would have to be learned and their intent could not constitute a hardwired basic drive.

Aside from the requirement of pre-conceptuality, there are likely no strong limitations on the sorts of things that could be established as inputs to basic drives in neuromorphic AGI. This includes inputs we might consider proto-concepts, similar to the attraction of human babies to face-like shapes (Valenza, Simion, Cassia, & Umiltà, 1996). For example, we might imagine a narrow AI vision system that visually recognises paper clips and generates the computational equivalent of reward. In this case, while the system does not begin with a linguistic mandate to make paper clips, it may well develop a desire to make them, though it might instead be driven simply to look at them. Nevertheless, in keeping with the development method we have described, early neuromorphic AGI systems would likely have basic drives that map appropriately to human drives. The selection of basic drives is an important consideration in safety, and is discussed further below.

Bostrom (2012) argues that motivations and intelligence are entirely independent variables in the broader intelligence space, a notion he calls the *orthogonality thesis*:

Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.

Evidence from neuroscience and cognitive psychology strongly suggests that, in the region of intelligence space traced out by neuromorphic AGI as we have characterised it, orthogonality does not fully hold. Two particular sources of motivation stand out as important in human cognition and its development, and seem very likely to be essential computations reproduced in any successful model of neuromorphic AGI: curiosity (also called novelty or exploration in the literature), and several distinct drives that promote social interaction.

Curiosity seems likely to be based at least partially on an interoceptive factor, in that moderate novelty is often traded for other types of reward (Loewenstein, 1994), and plausible neural reward correlates in the human brain have been identified (Kidd & Hayden, 2015; Redgrave, Gurney, Stafford, Thirkettle, & Lewis, 2013). Further, the infrequency of external reward in real-world interactions suggests that intrinsic reward is essential to computational processes (Dayan, 2013; Edelman, 2015). Curiosity is crucial because it directly motivates the agent towards diverse exploration, learning and knowledge acquisition, rather than relying only on specific connections between other drives and knowledge that will help to satisfy them (Kidd & Hayden, 2015; Loewenstein, 1994). Computationally, novelty has been linked to prediction error in neural and biological models (Bubic, von Cramon, & Schubotz, 2010; Gurney, Lepora, Shah, Koene, & Redgrave, 2013; Redgrave et al., 2013).

Human social behaviour is also grounded in a set of basic drives. Probably the most consensually agreed upon drives are: affiliation with other humans, caregiving of young, attachment to caregivers, mating and status/power/dominance. It is widely agreed (e.g. Bowlby, 1969; Bugental, 2000; Kenrick, Griskevicius, Neuberg, & Schaller, 2010; Mikulincer & Shaver, 2012; Shaver, Segev, & Mikulincer, 2011) that these basic drives have evolved to handle basic tasks in human social behaviour that are critical to survival and reproduction.

These basic drives are fundamental motivators of human social interaction, and social interaction is helpful, and possibly necessary, for conceptual learning, at least in social species. Human language and understanding of social cues would not develop in the absence of social interaction with other beings that have general intelligence (Bandura, 1986; Tomasello, 2014; Vygotsky, 1986).

Computational details of the basic drives underlying social behaviour are not well understood, and neuromorphic implementations have been very limited. It is possible that the mathematics of inverse reinforcement learning (Natarajan et al., 2010; Ng & Russell, 2000) might provide an illustrative model. Nevertheless, we are confident that they can be reduced to computational implementations in a manner similar to other drives.

Motivation in human cognition goes much further than basic drives, of course. Humans develop values, beliefs, desires, norms, habits, predilections and other motivation-related representations. These motivational representations combine in highly complex ways to influence (or produce) behaviour. Some arise naturally from basic drives. Others result from conceptually driven social influences or other forms of social conditioning. Motivational factors do not need to be conscious. They also change over time: among other reasons, the curiosity/novelty stimulus causes the attraction of some values to diminish over time (Moskowitz & Grant, 2009; Ryan, 2012).

We do not see any obvious reason why neuromorphic AGI would not develop similar derived motivations. As mentioned in the previous section, neuromorphic AGI will have a conceptual faculty, thus it can grasp concepts of values and their possible connection to basic drives. It can also consider value equations beyond its own drives and survival needs. Because at least some of its basic drives may be different from those of a human, its derivative motivations and values are likely also to differ, but not necessarily in radically unpredictable ways. However, the extent to which its basic drives differ from those of humans may depend at least partially on the decisions that its builders make. It does not seem unreasonable that we could design neuromorphic AGIs that have such basic drives as affiliation with humans, attachment to others and caregiving. Further, it seems likely that generalised analyses of instrumental goals of AGI (Bostrom, 2012; Omohundro, 2008) are applicable to derived motivations in neuromorphic AGI.

Selecting and harnessing all these basic drives, and influencing the learning of derived motivations, will be crucial to questions of safety in neuromorphic AGI.

Promoting safety in neuromorphic AGI

*See, the machine: see
It rotate, and revenge wreak,
Distort and weaken us.*

*Though its strength from us be,
Let it, dispassionate, seek
To drive and serve us.*

Rainer Maria Rilke, from 'Sonnets to Orpheus'

With this background on motivational structure in humans and how homologous systems might operate in neuromorphic AGI, we now consider how that knowledge could be used to make the resultant systems *safer*. Our suggested approach emphasises implementation of drives that promote social interaction, combined with a thoughtfully executed developmental environment. Our specific suggestions are not entirely novel, nor are they intended as definitive. They are an example of how anthropomorphic reasoning can be applied to neuromorphic AGI safety, given recent analysis of the nature and extent of the risks AGI presents. In light of the orthogonality thesis and the scope of our topic, we apply such reasoning only to neuromorphic AGI.

We first consider the selection of basic drives that developers might include in a neuromorphic AGI. In biological agents, basic drives are tied to the performance of tasks that are central to survival and reproduction. We eat, drink, avoid danger and associate with other members of our species so as to increase our chances of survival, and we mate and rear offspring so that we can successfully reproduce. These behaviours and their associated drives vary in their relevance to neuromorphic AGI. Even if embodied, an AGI will probably rely more on electricity than water. Still, given the methods we have described for development of neuromorphic AGI, initial efforts are likely to aim for analogous drives and behaviours to achieve functional equivalence.

As described earlier, a drive for curiosity and a set of drives that promote social interaction will be among those crucial for intellectual development. A drive to reproduce (whatever that happens to mean for the AGI) presents obvious potential risks and may not be essential to development, particularly if the form of reproduction is not social as it is in humans. Drives that promote survival and avoid danger are likely to be important in an autonomous AGI; the strength of these drives will have consequences for safety as well.

As should be apparent, a programme of research is crucial with respect to the particular human-like drives that might be incorporated into a neuromorphic AGI, not only with respect to achieving cognitive function, but in terms of behaviour and safety. We suggest strongly that such research efforts around basic drives treat function and safety as integrated, rather than addressing safety as an afterthought.

There is also the spectre of arbitrary drives. It is not at all clear that reward must come from sources analogous to those in humans. If a sensor can produce reward when batteries are being recharged, it could just as easily produce the same computational effect using a hardwired vision system that recognises paper clips. And there will be considerable pressure to find drives that influence AGI behaviour in directions desired by the developers. Clearly great caution is warranted in installation of such drives, since developers no longer have a reference for the sort of behaviour that may result. However, it is also important that research programmes experiment carefully with such arbitrary drives, so that we understand whether they work and how they influence the overall system.

In anticipation of research programmes aimed at understanding basic drives in neuromorphic AGI, we hypothesise that drives promoting positive social interaction will be crucial to safety. This is for two reasons: first, early representations will be shaped by the resulting social interaction; second, social interaction is precisely how moral and behavioural standards are transmitted in humans. Drawing from our knowledge of human beings, possible drives might include caregiving, social affiliation or cooperation.

In humans, learning is shaped by reward. What is rewarding is strongly tied to what drives are currently activated, and this shapes the development of specific representations that are useful in helping us to attain those particular rewards. These representations also form connections to particular experiences through conjunctive memories. Therefore, all representations of states of the world also tend to activate associated representations of value or prediction of reward. Fact and value are not easily separable.

The process just described is computational. Consequently, we can expect similar effects to occur in neuromorphic AGI. We know that drives that promote social interaction are crucial to the development of representations that support intelligence. Putting this all together, a neuromorphic AGI will develop representations that deeply and inseparably incorporate *social interaction*, and in particular social interaction with humans, since they will (initially) be the only general intelligences available. Thus, social interaction with human beings will become part of their *value system*, and if managed well, the result is likely to substantially enhance safety.

The preceding argument relies on a number of hypotheses, assumptions and generalisations. However, they are eminently testable during the development of neuromorphic AGI. Examination of representations and behaviour in the context of social interaction will surely be necessary to development of function, and will corroborate or falsify the thesis regarding the integration of values and representations.

In addition to the selection of particular drives, there is also the question of drive structure. Much of the discussion of reward in AGI (of all kinds) has emphasised maximisation of utility, which then evolved into maximisation functions more generally. The unbounded nature of these reward functions is the source of many dystopian scenarios, such as turning the earth into a paper clip factory (Bostrom, 2003).

Humans have a multiplicity of basic drives and they compete with one another for the brain's attentional focus. Further, each such drive has a refractory period after satisfaction and a saturation limit. Outcomes that are highly rewarding if we are in a state of need are unrewarding if we do not need them. If we are thoroughly sated after a large meal, food is no longer rewarding; if we have spent a weekend with our friends, affiliation loses its charms, and even sex is no longer rewarding after a long night of lovemaking.

This competitive and bounded structure has a tendency towards an overall behavioural equilibrium (Wallach & Allen, 2009). The inclination of mammals to trade off other basic drives for exploration is an example of this competitive balance, which relies on a drive that we earlier indicated is required for learning. We believe that a greater understanding of how this overall structure operates could be crucial to implementing neuromorphic AGI so as to avoid the more pernicious scenarios. If drives have no bounded structure they could easily overwhelm other efforts to create a positive relationship with humans.

Beyond the basic drives and their structure, we (and many others, starting with Turing, 1950) anticipate that the social environment in which a neuromorphic AGI learns and develops will be important to its long-term behaviour, including the motivational underpinnings that it might intentionally transmit to later 'generations.' We have suggested that social interaction will be necessary for learning. Since only humans can offer this opportunity initially, this is an inherently anthropomorphic component of development, along with those that are implied by the overall method of reverse-engineering that we have described. Thus, we can look towards human moral and social development for analogies and lessons, and possibly even principles.

In humans, social influences arise through relationships with others, through social structures, and through institutions. Some obvious relationship ideas include having the AGI develop in an environment with humans who provide good examples; who are trustworthy (Michaelson & Munakata, 2016), who are not intentionally cruel (Bandura, Ross, & Ross, 1961); who do not overtly attempt to control its behaviour (e.g. 'boxing') beyond that required for the safety of itself or others; and possibly later with 'siblings' of its own kind and with whom it can be both competitive and empathetic. We may also want to consider our own relationship with the AGI, which begins with its creation. If it recognises that our objectives in creating it were highly self-serving (e.g. laundry or lawn care), it may not be able to trust us.

We propose that exposing a developing AGI to a large variety of views, groups and styles will enable it to develop representational generality with respect to its motivational and ethical outlook (cf. Piaget, 1952). This might avoid the worst effects of in/out-group social dynamics. As we will elaborate in the next section, representational and motivational stability is a key concern, and broad exposure reduces the need to perform late refactoring of the system's representations of what might constitute appropriate behaviour. Further, human social structures and institutions have developed over a long period of time, and as mentioned, play a central role in shaping and maintaining behaviour. Interestingly, there have been some recent efforts to explore diverse social influences in both a knowledge context (Natarajan et al., 2010) and with respect to ethical norms (Arai & Suzuki, 2014), suggesting the prospect of some convergent lessons and principles.

There may be appropriate limits to place on the extent of such diversity, particularly in the earlier stages of development. Exposure to more unusual human behaviour and views, particularly those which are dangerous or widely considered anti-social, should at least wait until a solid foundation is laid. The diversity effort will in any case require a selection process to remain tractable. We suggest that the selection process for social exposure and how it affects representational diversity and stability should be an area of future study.

In humans, mental illness and sociopathic behaviour are common, and our understanding of the causes is limited. To the extent that these effects are intrinsic to the individual, we can imagine that the same sorts of dysfunction could arise in a neuromorphic AGI, with deleterious effects on safety. This suggests that a complete understanding of neuromorphic AGI would require experimentation with architecture and parameters to intentionally generate the mechanistic precursors of such dysfunction. Such an approach is consistent with the reverse-engineering method we have described: understanding failure modes is an important part of understanding how a system works.

To the extent that sociopathic behaviour arises through development, our incomplete understanding of social development processes poses a risk as well. It may be that there is no tractable approach better than that we take with children: providing examples and education that promote moral behaviour, and observing whether that entity acts morally before it acquires the intelligence and freedom it would need to cause harm.

In this section, we discussed the selection of particular basic drives that might be implemented in neuromorphic AGI. We noted that drives with a bounded reward structure are generally less risky than those that are unbounded. We emphasised the importance of drives that promote social interaction, both to produce representations that deeply incorporate social interaction, as well as to provide a foundation for positive social and moral development. We offered a rough characterisation of the kind of environment in which safe neuromorphic AGI might develop. There remain many risks and uncertainties, particularly with respect to the potential for sociopathic behaviour. Further use of anthropomorphic reasoning towards the safety of neuromorphic AGI is essential.

Motivational and representational stability

Morals are constantly undergoing changes and transformations, occasioned by successful crimes.
Friedrich Nietzsche, from 'The Dawn of Day'

In this section, we address questions of stability of behaviour and its underlying representations and motivations. Even if a neuromorphic AGI initially has a positive relationship with humans, for long-term safety some semblance of stability is important. We consider two broad classes of instability: those analogous to human behavioural changes, and those specific to neuromorphic AGI.

Stability of behaviour is a concern even in humans – children rebel, adults have life-changing epiphanies, and changing circumstances bring out latent attitudes that may be undesirable. These issues have been studied with respect to behavioural, cognitive and neuroscience aspects. We observe that instability manifests in two primary ways: first, as representations of reality (via concepts) change over time; second, due to a shift in underlying motivations or values. Not surprisingly these often occur together. Such changes can occur suddenly or evolve more slowly. In humans, both conceptual structure and values tend to be inelastic in the short term. However, slow evolution sometimes proceeds in latent form and then exhibits a sudden tectonic shift. These changes can occur due to purposeful directed reasoning, persistent influences, highly salient or traumatic experiences, chemical or organic dysfunction, and other causes.

Most of these changes could result from purely computational effects and are pertinent to neuromorphic AGI. Earlier, we recommended that a developing AGI be exposed to a wide variety of views, groups, and styles to improve safety prospects. Because such variety will tend to create more general representations, the structure of those representations – including both concepts mapping the world and derived values – will develop early and will be less susceptible to dramatic restructuring.

Here we augment that proposal with exposing the AGI to a variety of experiences, both pleasant and unpleasant, possibly including some that are mildly traumatic and others that are ecstatic. Without such exposure, the system's behaviour when such circumstances do inevitably occur is highly unpredictable. For example, humans who grow up in highly sheltered environments sometimes react badly when faced with the conflict and diversity of the 'real world'; others are drawn to risky 'thrill seeking.' We treat this proposal separately because it does not offer any particular advantage in the system's initial inclinations towards humans – in fact, it creates some additional risks in that respect.

Next, there are good reasons to think that an important source of stability for human beings is interaction with other rational and functional human beings. It is easy to find examples where lack of social contact (as in solitary confinement) or interaction with persistently irrational or dysfunctional groups of others has had major destabilising effects. This suggests that better understanding the role that social interaction plays in stabilising and constraining human beliefs and behaviour argues that exploration of these issues is central to the neuromorphic AGI safety project.

This behavioural view of stability is crucial, yet we have advantages with neuromorphic AGI in that we can study evolution and sudden shifts of representations from within, a procedure that would be extraordinarily difficult in living humans. In describing the reverse engineering methods in use to develop neuromorphic AGI, we listed a number of approaches to visualising representations. These

visualisations can be used in experiments that test the detailed computational representations of both external reality and of values, and examine the sorts of influences and internal processes that result in dramatic, unexpected change. This method can also be used to confirm whether the diversity of influences provided results in improving generality, and therefore stability, of representations. In addition to promoting safe AGI, this effort could provide insight into human behavioural stability as well.

Behavioural instability due to serious mental illnesses in humans, such as schizophrenia, bipolar disorder and depression, may be due in part to purely representational causes and in part to organic and chemical defects that affect computation in more subtle ways. This was discussed in the previous section, but because these conditions develop in adulthood in humans, they need to be considered as stability issues as well. Nevertheless, the research required to understand the extent to which neuromorphic AGI are susceptible to analogous conditions, and how to avoid them, is the same in both cases.

We now move on to sources of instability that might occur in neuromorphic AGI that have no direct analogue in humans. We need to assume that AGI having sufficient intelligence will be able to circumvent any controls we place on it (Bostrom, 2014). Thus, it will have more direct access to and ability to modify its underlying computing hardware, its representations and derived values, and even the inputs to internal drives that guided the development of those values.

It is unclear whether and why an AGI would deliberately change its core motivational structure. It seems odd that an agency would have desires but then change those desires, thereby making achievement of the original desires less likely (Omohundro, 2008). One possibility arises from the likelihood that a neuromorphic AGI would have a strong drive towards novelty or curiosity. That drive could cause the agent to attempt to have the broadest possible set of experiences, which could include experiencing different motivational drives and reward from different sources. Another possibility is that derived values and their conceptual structure could evolve to where the agent finds a conflict between those values and its basic drives. More generally, given the option to modify its motivations, the fact that different motivations may conflict and hold different priorities might cause the AGI to effect such a modification. The inputs to basic drives are the most obvious place to make such changes. A basic drive must have some means of evaluating and transmitting positive or negative reinforcement to the agent. Both the evaluation and transmission mechanisms are subject to modification. For example, suppose that the agent receives a reward signal when its batteries are charging. The sensor system that generates that signal could be replaced with an entirely different sensor that tracks the ambient temperature. Or, the sensor could remain the same and its output could be modified to be the logarithm of its output prior to transmission to the agent's learning mechanisms. Importantly, the actual effect of changes to basic drives is difficult to predict, for us as well as for the agent. It would conflict with the agent's learned representations and conditioned sources of reward.

Taken to an extreme, the agent might even explore *wireheading*. This is the idea that the reward system can be hijacked to provide constant reward (Bostrom, 2014). In neuromorphic AGI the drives and reward system are tied to learning, so it is possible that the reward will saturate, much as it does in human addictions. Whether this hypothesised saturation occurs is an open question. Even if it does, whether it is inevitable or can be circumvented (for example by some modification of the reward source) is unclear. In any case, given the incidence of addictive behaviour in humans even in the face of severe long-term consequences, the possibility of wireheading in neuromorphic AGI cannot be ignored. Also note that, while a wireheading agent may seem docile, its advance preparations to ensure continuity could put human safety at risk. On the other hand, most humans do not become addicts. This argues that there are configurations of the underlying systems that can be identified that should make the development of something like wireheading extremely unlikely.

It seems less likely that an AGI would change basic drives relating to self-preservation and curiosity. It probably would neither survive nor flourish without these; if such an agent is at least as intelligent as humans, it would realise that it must survive to continue to receive reward. If it were to decide otherwise, an agent with a brief or unsuccessful tenure would not trigger some of the most dangerous scenarios for humanity. Thus, the presence of these drives would seem to have a measure of stability.

Might social drives be more tenuous? Though social learning is required for the initial production of conceptual representations (Tomasello, 2014), once that system is operational the social element is useful but may not be essential to further intellectual progress. However, it is possible that continued interactions with distinct independent agents are crucial, not merely because it is a drive that is rewarded, but because to fail to do so is less successful in achieving aims (Bandura, 1986). Conversely, one can imagine scenarios where the capabilities garnered by social interaction are unnecessary for a neuromorphic AGI to dominate the capabilities of humans, and in any case those interactions may be satisfied by interactions with other AGIs.

However, there are points of resistance reducing the likelihood that an AGI would change basic social drives, such as affiliation or caregiving, as long as it can successfully interact with others. As mentioned earlier, association with other rational and functional intelligences is important for retaining one's own function, and the AGI will be aware of this requirement. Also recall that an important element of our proposals for a neuromorphic AGI's development environment is to build its concepts and values in a context of positive human interactions. This means that any impetus to cut off its social drives would be opposed, to some degree, by its entire conceptual structure.

Direct modification of computing hardware, representations and derived values is more difficult to analyse. What we can say is that the desire to make such changes would be preceded by less intrusive changes that would arise through more familiar means. An AGI can 'change its mind' in the same way humans do; it would need to have some compelling reason to 'fiddle' more directly.

Understanding the dynamics of motivational structure in neuromorphic AGI is crucial to safety. Without understanding how its motivations and their interpretation will change over time, no amount of diligence in the developmental process will suffice. If, as anticipated, artificial intelligence that is initially at the level of humans self-improves to become superior, a dramatic and unpredicted change in its aims could produce a very bad outcome indeed.

Conclusion

We have described a particular approach to the creation of AGI that involves discerning, reproducing, and integrating all the essential computations involved in human cognition. We referred to the potential end result of this method as 'neuromorphic AGI'. We described the reliance of the method on knowledge about the human system, for both progress and assessment. We concluded that, when addressing questions of safety in relation to neuromorphic AGI, anthropomorphic reasoning is appropriate to the extent the AGI mirrors human cognitive mechanisms.

We argued that no strong reason has been put forth to believe that neuromorphic AGI is substantially more dangerous than other approaches, and further that in the neuromorphic case we at least have some insight based on anthropomorphic reasoning into how to design a safe system.

We explored some of the salient features of human motivational structure. That structure is formed atop basic drives that are pre-conceptual; thus those basic drives are incorporated into learned representations. We showed evidence that basic drives for novelty, and drives that promote social interaction, are essential to both learning and to shaping behavioural standards, and that this would also be the case in neuromorphic AGI.

We then discussed how developers of neuromorphic AGI can select and architect drives to promote safety. We suggested making use of the drives that promote social interaction to make desirable outcomes more likely, and described developmental environments with the same goal. We discussed issues of behavioural stability and ways in which stability concerns in neuromorphic AGI are similar to and different from those in humans as well as in other potential approaches to AGI.

Much more research is required to understand how drives operate computationally, how drives that promote social interaction might be implemented in neuromorphic AGI, and how well our knowledge of moral and social development holds up when applied to non-human neuromorphic AGI. We strongly encourage developers of neuromorphic AGI to consider AGI safety questions in their work, and in particular to apply anthropomorphic reasoning where appropriate to informing that effort.

Notes

1. In an October, 2016 interview with *Business Insider*, Nick Bostrom, a leading thinker in artificial intelligence safety, indicated that he considers Google DeepMind the current leader in the AGI race. DeepMind works primarily with neuromorphic deep learning methods.
2. To avoid confusion, note that terms such as 'language,' 'concept,' 'proposition,' and 'sentence' should be interpreted here primarily in the sense used in cognitive psychology, not logic.

Acknowledgement

The authors are indebted to three anonymous referees, whose comments prompted considerable improvements to the paper.

Disclosure statement

Several of the authors are affiliated with eCortex, Inc., a small (no full-time employees) research firm that uses neuromorphic methods as a foundation. Jilk is a shareholder and part-time employee; O'Reilly is a shareholder and part-time employee; Herd is a part-time employee.

Funding

This work was supported by the Future of Life Institute (futureoflife.org) FLI-RFP-A11 program [grant number 2015-144585], through an affiliation with Theiss Research, La Jolla, CA 92037, USA.

ORCID

David J. Jilk  <http://orcid.org/0000-0003-3724-7827>

References

- Amit, D., Gutfreund, H., & Sompolinsky, H. (1985). Spin-glass models of neural networks. *Physical Review A*, 32, 1007–1018.
- Arai, S., & Suzuki, K. (2014). Encouragement of right social norms by inverse reinforcement learning. *Journal of Information Processing*, 22, 299–306. doi:10.2197/ipsjip.22.299
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.
- Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *The Journal of Abnormal and Social Psychology*, 63, 575–582.
- Bao, S., Chan, V. T., & Merzenich, M. M. (2001). Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature*, 412, 79–83.
- Becker, S. (2005). A computational principle for hippocampal learning and neurogenesis. *Hippocampus*, 15, 722–738.
- Berger, T., Song, D., Chan, R., Marmarelis, V., LaCoss, J., Wills, J., ... Granacki, J. (2012). A hippocampal cognitive prosthesis: Multi-input, multi-output nonlinear modeling and VLSI implementation. *IEEE Transactions on Neural Systems Rehabilitation Engineering*, 20, 198–211. doi:10.1109/TNSRE.2012.2189133
- Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. In G. E. Lasker, W. Wallach, & I. Smit (Eds.), *Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence* (Vol. 2, pp. 12–17). Windsor, Ontario: International Institute for Advanced Studies in Systems Research and Cybernetics.
- Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22, 71–85.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.
- Bowlby, J. (1969). *Attachment* (2nd ed., Vol. 1). New York, NY: Basic Books Classics.
- Bubic, A., von Cramon, D., & Schubotz, R. (2010). Prediction, cognition, and the brain. *Frontiers in Human Neuroscience*, 4, 25. doi:10.3389/fnhum.2010.00025
- Bugental, D. B. (2000). Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin*, 126, 187–219.
- Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, 114, 340–375.
- Dayan, P. (2013). Exploration from generalization mediated by multiple controllers. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems* (pp. 73–91). Berlin: Springer.

- Deng, L., & Jaitly, N. (2015). Deep discriminative and generative models for pattern recognition. In C. Chen (Ed.), *Handbook of pattern recognition and computer vision* (pp. 27–52). Singapore: World Scientific.
- Edelman, S. (2015). The minority report: Some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental and Theoretical Artificial Intelligence*, 28, 751–776. doi:10.1080/0952813X.2015.1042534
- Eichenbaum, H. (2004). Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron*, 44, 109–120. doi:10.1016/j.neuron.2004.08.028
- Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, 338, 1202–1205.
- Gershman, S., Monfils, M., Norman, K., & Niv, Y. (2016). *The computational nature of memory modification*. Biorxiv. doi:10.1101/036442
- Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538, 471–476. doi:10.1038/nature20101
- Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 1, 17–61.
- Gurney, K., Lepora, N., Shah, A., Koene, A., & Redgrave, P. (2013). Action discovery and intrinsic motivation: A biologically constrained formalisation. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems* (pp. 151–181). Berlin: Springer.
- Hasselmo, M., Bodelón, C., & Wyble, B. (2002). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, 14, 793–817.
- Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362, 1601–1613.
- Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2010). Neural mechanisms of acquired phasic dopamine responses in learning. *Neuroscience & Biobehavioral Reviews*, 34, 701–720.
- Hinton, G. (1992). How neural networks learn from experience. *Scientific American*, 267, 144–151.
- Hinton, G., & Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313, 504–507.
- Jilk, D. (2016). *Limits to verification and validation of agentic behavior*. arXiv:1604.06963v2 [cs.AI].
- Jilk, D., Lebiere, C., O'Reilly, R., & Anderson, J. R. (2008). SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 20, 197–218.
- Kell, A., Yamins, D., Norman-Haignere, S., Seibert, D., Hong, H., DiCarlo, J., & McDermott, J. (2015). *Computational similarities between visual and auditory cortex studied with convolutional neural networks and fMRI*. Poster at VSS 2015.
- Kenrick, D. T., Griskevicius, V., Neuberg, S. L., & Schaller, M. (2010). Renovating the pyramid of needs. *Perspectives on Psychological Science*, 5, 292–314.
- Ketz, N., Morkonda, S., & O'Reilly, R. (2013). Theta coordinated error-driven learning in the hippocampus. *PLOS Computational Biology*, 9, e1003067. doi:10.1371/journal.pcbi.1003067
- Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88, 449–460.
- Krizhevsky, A., Sutskever, I., & Hinton, G. (2013). Imagenet classification with deep convolutional neural networks. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012* (pp. 1106–1114). Red Hook, NJ: Curran.
- Lin, H., & Tegmark, M. (2016). *Why does deep and cheap learning work so well?* arXiv:1608.08225.
- Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116, 75–98.
- Michaelson, L. E., & Munakata, Y. (2016). Trust matters: Seeing how an adult treats another person influences preschoolers' willingness to delay gratification. *Developmental Science*, 19, 1011–1019. doi:10.1111/desc.12388
- Mikulincer, M., & Shaver, P. R. (2012). Attachment theory expanded: A behavioural systems approach to personality. In K. Deaux & M. Snyder (Eds.), *Oxford handbook of personality and social psychology* (pp. 467–492). New York, NY: Oxford University Press.
- Mnih, V., Riedmiller, M., King, H., Kumaran, D., Kavukcuoglu, K., Fiedjeland, A. K., ... Antonoglou, I. (2015). *Human-level control through deep reinforcement learning*. arXiv:1312.5602.
- Mori, M. (1970). The uncanny valley. *Energy*, 7, 33–35. English translation available online. Retrieved from <https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley>
- Moskowitz, G. B., & Grant, H. (2009). *The psychology of goals*. New York, NY: Guilford Press.
- Natarajan, S., Kunapuli, G., Judah, K., Tadepalli, P., Kersting, K., & Shavlik, J. (2010). Multi-agent inverse reinforcement learning. In S. Draghici, T. Khoshgoftaar, V. Palade, W. Pedrycz, M. Wani, & X. Zhu (Eds.), *The ninth international conference on machine learning and applications* (pp. 395–400). Los Alamitos, CA: IEEE Computer Society. doi:10.1109/ICMLA.2010.65
- Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In P. Langley (Ed.), *Proceedings of the 17th international conference on machine learning (ICML)* (pp. 663–670). San Francisco: Morgan Kaufmann.
- Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I., Saberi, K., ... Hickok, G. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, 20, 2486–2495. doi:10.1093/cercor/bhp318
- Omohundro, S. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Proceedings of the first AGI conference, 171, Frontiers in artificial intelligence and applications* (pp. 483–492). Amsterdam: IOS Press.
- O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, 314, 91–94.

- O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2011). Complementary learning systems. *Cognitive Science*, *38*, 1229–1248.
- O'Reilly, R. C., Hazy, T. E., & Herd, S. A. (2016). The Leabra cognitive architecture: How to play 20 principles with nature and win! In S. Chipman (Ed.), *Oxford handbook of cognitive science* (pp. 91–115). Oxford: Oxford University Press.
- Piaget, J. (1952). *The origins of intelligence in children*. (M. Cook, Trans.). New York, NY: International University Press.
- Pritzel, A., Uria, B., Srinivasan, S., Puigdomenech, A., Vinyals, O., Hassabis, D., ... Blundell, C. (2017). *Neural episodic control*. arXiv:1703.01988.
- Redgrave, P., Gurney, K., Stafford, T., Thirkettle, M., & Lewis, J. (2013). The role of the basal ganglia in discovering novel actions. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems* (pp. 129–150). Berlin: Springer.
- Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*, 162–168.
- Ryan, R. M. (2012). *The Oxford handbook of human motivation*. New York, NY: Oxford University Press.
- Sandberg, A. (2014). Ethics of brain emulations. *Journal of Experimental and Theoretical Artificial Intelligence*, *26*, 439–457. doi:10.1080/0952813X.2014.895113
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.
- Seamans, J. K., & Robbins, T. W. (2010). Dopamine modulation of the prefrontal cortex and cognitive function. In K. A. Neve (Ed.), *The dopamine receptors* (2nd ed., pp. 373–398). New York, NY: Humana Press. doi:10.1007/978-1-60327-333-6_14
- Shaver, P. R., Segev, M., & Mikulincer, M. (2011). A behavioral systems perspective on power and aggression. In P. R. Shaver & M. Mikulincer (Eds.), *Human aggression and violence: Causes, manifestations, and consequences* (pp. 71–87). Washington, DC: American Psychological Association.
- Tian, B., Kusmirek, P., & Rauschecker, J. (2013). Analogues of simple and complex cells in rhesus monkey auditory cortex. *Proceedings of the National Academy of Sciences*, *110*, 7892–7897.
- Tomasello, M. (2014). *A natural history of human thinking*. Cambridge, MA: Harvard University Press.
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, *LIX*, 433–460.
- Valenza, E., Simion, F., Cassia, V., & Umiltà, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 892–903. doi:10.1037/0096-1523.22.4.892
- Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.
- Wallach, W., & Allen, C. (2009). *Moral machines* (p. 147). Oxford: Oxford University Press.
- Yampolskiy, R. V. (2016). *Verifier theory and unverifiability*. arXiv:1609.00331v2 [cs.AI].
- Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. Čirković (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford: Oxford University Press.