Chapter 9A
Colin Allen and Wendell Wallach on Omohundro's "Rationally-Shaped Artificial Intelligence"

Natural Born Cooperators

Omohundro, citing Kurzweil, opens with the singularitarian credo that "[s]ystems with the computational power of the human brain are likely to be cheap and ubiquitous within the next few decades." What is the computational power of the human brain? The only honest answer, in our view, is that we don't know. Neuroscientists have provided rough total neuron counts and equally rough estimates of neural connectivity, although the numbers are by no means certain (Herculano-Houzel 2009). But we haven't even begun to scratch the surface of neural diversity in receptor expression. Even the genome for the 302 neurons belonging to the "simple" flatworm *C. elegans* encodes "at least 80 potassium channels, 90 neurotransmitter-gated ion channels, 50 peptide receptors, and up to 1000 orphan receptors that may be chemoreceptors" (Bargmann 1998). As Koch (1999) put it in 1999, the combinatoric possibilities for the *C. elegans* nervous system are "staggering", and in the subsequent years things have not come to seem any simpler. We don't know what all these receptors do. Consequently, we don't know how to calculate the number of computations per second in the *C. elegans* "brain"—let alone the human brain.

Kurzweil, meanwhile, has argued that even if he is off by many orders of magnitude in his estimate of the number of computations per second carried out by the human brain, the exponential growth of raw computing power in our machines means that the coming singularity will be only moderately delayed. Irrespective of this, any conjecture about what the exponential growth of computing power means for artificial intelligence and machine-human relations remains unfalsifiable if there is no direct relationship between "raw computing power" and intelligent behavior. There is no such direct relationship. Intelligence does not magically emerge by aggregating neurons; it depends essentially on the the way the parts are arranged. Impressive as it is, IBM's Watson with all its raw computational power lacks the basic adaptive intelligence of a squirrel, even though it can do many things that no squirrel can do. So can a tractor.

Without offering a theory of how intelligence emerges, Kurzweil blithely argues that the organization of all this raw capacity for information flow will only lag modestly behind Moore's law. He also believes that the inevitable acceleration toward the singularity is unlikely to be significantly slowed by the additional complexities that accompany each order of magnitude increase in the number of components on a circuit board. But already Urs Hölzle, Google's first vice president of operations, reports inherent problems maintaining the stability of systems that are dramatically smaller in scale than those imagined by singularitarians (Clark 2011).

Omohundro offers a progressivist story to explain the inevitable evolution of intelligence, from stimulus-response systems, through learning systems, reasoning systems, and self-improving systems, to fully rational systems. Perhaps the squirrel is stuck somewhere at the stage of learning systems, but *C. elegans* can learn too, leaving much to be explained about the evolutionary pathways between. Or perhaps the squirrel is a reasoner. Omohundro maintains that, "[a]dvanced animals with nervous systems do deliberative reasoning." He provides no criteria for testing this claim, however. And if there are self-improving squirrels, how would we know?

We take, it, however, that Omohundro thinks squirrels are not fully rational. He writes that, "In most environments, full rationality is too computationally expensive." The viable alternative is to be "as rational as possible." How rational is it possible to be? Omohundro imagines that within computational constraints it is possible for a "rational shaper" to adjust the system's state transition and action functions so as to maximize the system's expected utility in that environment. If there are environments in which squirrels count as rational utility maximizers, they don't include roads. Rational shapers have blind spots, as is evident even in human behavior.

Omohundro explains that very limited systems can only have a fixed stimulus-response architecture, but as computation gets cheaper, there is a niche for learners to exploit stimulus-response systems. And as computational power increases, less rational agents can be exploited by more rational ones. The "natural progression" towards full rationality is thus an inevitable consequence of the evolutionary arms race, as he sees it. He writes, "If a biological system behaves irrationally, it creates a niche for others to exploit. Natural selection will cause successive generations to become more and more rational." If this is true, it's an exceedingly slow process. Even if today's squirrels are more rational than their forebears, it seems to be the epitome of an untestable hypothesis.

Humans are taken by Omohundro to be at the pinnacle so far of this progression. But he foresees the day when machines will be able to exploit human irrationality. The natural progression is thus towards machines that "will behave in anti-social ways if they aren't designed very carefully." Those who follow the behavioral and cognitive sciences will find it a little surprising to see that *Homo economicus,* the selfish utility-maximizer of twentieth century economic theory, is among the undead. It's about what one would expect for someone whose economics textbook is dated 1995. But it is no longer credible to think that rational models of expected utility maximization are the best way to understand either evolution or economic behavior. Even bacteria cooperate via quorum sensing, and there exist both kin selection and group selection models to explain the evolution of cooperative behavior in many different species. Non-cooperative defection is always a possibility, but it is by no means inevitable even between the species.

Part of Omohundro's thesis should be acknowledged: careful design is necessary if we are to have machines we can live with. But the dangers are unlikely to come in the way he imagines. He proposes a "simple thought experiment" which, in his words, "shows that a rational chess robot with a simple utility function would behave like a human sociopath fixated on chess." In this, Omohundro exemplifies the "Giant Cheesecake Fallacy" described by Yudkowsky (this volume)—i.e., he imagines that just because machines can do something, they will. But it is far from clear that the kind of behavior he imagines would maximize the machine's expected utility, or that we should go along with his Nietzschean view that a "cooperative drive" will be felt only by those at a competitive disadvantage. Man and supermachine.

A more science-based approach is needed. Formal models developed from a priori theories of rationality have proven to be of limited use for understanding the complex details of evolution and intelligence. So-called "simple heuristics", discovered empirically, may make organisms smart in ways that cannot be easily exploited in typical environments by more cumbersome rational optimization procedures. If this is all that Omohundro means by the phrase "as rational as possible" then his thesis has no teeth, predicting nothing but allowing everything. Careful design must proceed from detailed study and understanding of actual processes of evolution and the real, embodied forms of moral agency that evolution has provided.

The article by Omohundro exemplifies a broader problem with the singularity hypothesis. Gaps in the hypothesis are rationalized away or filled in with additional theories that are just as vague or just as difficult to verify as the initial conclusion that a technological singularity is inevitable. While the singularity may appear plausible to its proponents, the speculation, ad hoc theorizing, and inductive reasoning used in its support fall far short of scientific rigor.

References

Bargmann, C. I. (1998). Neurobiology of the Caenorhabditis elegans genome. *Science, 282*(5396), 2028–2032.

Clark, J. (2001). Google: 'At scale, everything breaks' ZDNet UK, June 22, 2011. Available online at http://www.zdnet.co.uk/news/cloud/2011/06/22/google-at-scale-everything-breaks-40093061/. Accessed March 23, 2011

Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience, 3*, 1–11.

Koch, C.,& Laurent, G. (1999). Complexity and the nervous system. *Science, 284*(5411), 96–98.