Can't We Just Talk?
Commentary on Arel's "Threat"

William J. Rapaport

**Department of Computer Science and Engineering,
Department of Philosophy, Department of Linguistics,
and Center for Cognitive Science
University at Buffalo, The State University of New York, Buffalo, NY 14260-2500**

rapaport@buffalo.edu
http://www.cse.buffalo.edu/∼rapaport/

Itamar Arel (2012) argues that:

1. artificial general intelligence (AGI) "is inevitable" (§1.1),
2. techniques including a "fusion between deep learning, . . . a scalable situation inference engine, and reinforcement learning [RL] as a decision-making system may hold the key to place us on the path to AGI" (§2), and
3. "a potentially devastating conflict between a reward-driven AGI system and the human race. . . is inescapable, given the assumption that an RL-based AGI will be allowed to evolve" (§2).

Why "inescapable"? If I understand Arel correctly, it is a mathematical certainty:

> [F]rom equations (2) and (4) [Arel 2012, §§4.1, 6.1, the details of which are irrelevant to my argument], it follows that the agent continuously attempts to maximize its "positive" surprises [i.e., "its wellbeing"]. . . while minimizing "negative" surprises. This process. . . is unbounded. . . . [O]nce such a bonus is received on a regular basis, it becomes the new norm and no longer yields the same level of satisfaction. This is the core danger in designing systems that are driven by rewards and have large cognitive capacity; by continuously striving to gain positive (relative) reinforcement, they will inevitably pose a danger to humanity.

Let's suppose so. But why should it be "inevitable"? Despite Arel's faith in the inevitability of *AGI* (which I share), he seems to be committing the fallacy of thinking that AGIs must differ in crucial respects from humans.

This is the fallacy that John Searle commits when claiming that the inhabitant of his Chinese Room (Searle 1980) doesn't "understand a word of Chinese and neither

does any other digital computer because all the computer has is what [the inhabitant] ha[s]: a formal program that attaches no meaning, interpretation, or content to any of the symbols" (Searle 1982: 5). As I have pointed out elsewhere, this assumes "that external links are needed for the program to 'attach' meaning to its symbols" (Rapaport 2000, §3.2.2). The fallacy can be seen by realizing that "*if* external links *are* needed, then surely a computer could have them as well as—and presumably in the same way that—humans have them" (Rapaport 2000, §3.2.2).

Why do I think that Arel is committing this fallacy? Because, presumably, *humans also* "attempt to maximize [their] wellbeing". Now, I can agree that humans themselves have been known, from time to time, to "pose a danger to humanity" (for a discussion of this, see Dietrich 2001, 2007). *But we have also devised methods for alleviating such dangers.* Clearly, then, rather than wringing our hands over the "inevitability" of AGIs wreaking havoc on their creators, we should give them some of those methods.

And, indeed, Arel sketches out some possibilities along these lines: education and "limit[ing] such [a] system's mental capacity" (§6.2). But he seems to neglect one obvious possibility, one that is, in fact, a *necessity* for any AGI: For an AGI to really have GI—general intelligence—it must have cognition: (1) It must be able to use and understand *language*—and, presumably, *our* language, so that *we* can communicate with it, and vice versa (see Winston 1975 and my discussion of "Winston's problem" in Rapaport 2003)—and (2) it must be able to *reason* consciously (e.g., via an explicit knowledge-representation-and-reasoning system, as opposed to tacit reasoning by, say, an artificial neural network). If we can reason with it in natural language, then we can hope to be able to collaborate and negotiate with it, rather than compete with it. Such natural-language and reasoning competence is, in any case, a prerequisite (or at least a product) of education, but it requires no limitation on the AGI's mental capacity.

References

Arel, Itamar (2012), "The Threat of a Reward-Driven Adversarial Artificial General Intelligence", in Amnon Eden, Johnny Søraker, James H. Moor, & Eric Steinhart (eds.), *The Singularity Hypothesis: A Scientific and Philosophical Analysis* (Springer).

Dietrich, Eric (2001), "Homo Sapiens 2.0: Why We Should Build the Better Robots of Our Nature", *Journal of Experimental and Theoretical Artificial Intelligence* 13(4) (October): 323–328.

Dietrich, Eric (2007), "After the Humans Are Gone", *Journal of Experimental and Theoretical Artificial Intelligence* 19(1): 55–67.

Rapaport, William J. (2000), "How to Pass a Turing Test: Syntactic Semantics, Natural-Language Understanding, and First-Person Cognition", *Journal of Logic, Language, and Information*, 9(4): 467–490; reprinted in James H. Moor (ed.), *The Turing Test: The Elusive Standard of Artificial Intelligence* (Dordrecht: Kluwer, 2003): 161–14.

Rapaport, William J. (2003), "What Did You Mean by That? Misunderstanding, Negotiation, and Syntactic Semantics", *Minds and Machines* 13(3): 397–427.

Searle, John R. (1980), "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3: 417–457.

Searle, John R. (1982), "The Myth of the Computer", *New York Review of Books* (29 April 1982):

3–6; cf. correspondence, same journal (24 June 1982): 56–57.

Winston, Patrick Henry (1975), “Learning Structural Descriptions from Examples”, in Patrick Henry Winston (ed.), *The Psychology of Computer Vision* (New York: McGraw-Hill): 157–209; reprinted in Ronald J. Brachman & Hector J. Levesque (eds.), *Readings in Knowledge Representation* (Los Altos, CA: Morgan Kaufmann, 1985): 141–168.