

Friday, September 04, 2009

The Singularity and the Fixed Point

The importance of engineering motivation into intelligence.

By Edward Boyden

Some futurists such as Ray Kurzweil have hypothesized that we will someday soon pass through a singularity--that is, a time period of rapid technological change beyond which we cannot envision the future of society. Most visions of this singularity focus on the creation of machines intelligent enough to devise machines even more intelligent than themselves, and so forth recursively, thus launching a positive feedback loop of intelligence amplification. It's an intriguing thought. (One of the first things I wanted to do when I got to MIT as an undergraduate was to build a robot scientist that could make discoveries faster and better than anyone else.) Even the CTO of Intel, Justin Rattner, has publicly speculated recently that we're well on our way to this singularity, and conferences like the Singularity Summit (at which [I'll be speaking in October](http://www.singularitysummit.com/program) (<http://www.singularitysummit.com/program>)) are exploring how such transformations might take place.

As a brain engineer, however, I think that focusing solely on intelligence augmentation as the driver of the future is leaving out a critical part of the analysis--namely, the changes in motivation that might arise as intelligence amplifies. Call it the need for "machine leadership skills" or "machine philosophy"--without it, such a feedback loop might quickly sputter out.

We all know that intelligence, as commonly defined, isn't enough to impact the world all by itself. The ability to pursue a goal doggedly against obstacles, ignoring the grimness of reality (sometimes even to the point of delusion--i.e., against intelligence), is also important. Most science-fiction stories prefer their artificial intelligences to be extremely motivated to do things--for example, enslaving or wiping out humans, if *The Matrix* and *Terminator II* have anything to say on the topic. But I find just as plausible the robot Marvin, the superintelligent machine from Douglas Adams' *The Hitchhiker's Guide to the Galaxy*, who used his enormous intelligence chiefly to sit around and complain, in the absence of any big goal.

Indeed, a really advanced intelligence, improperly motivated, might realize the impermanence of all things, calculate that the sun will burn out in a few billion years, and decide to play video games for the remainder of its existence, concluding that

inventing an even smarter machine is pointless. (A corollary of this thinking might explain why we haven't found extraterrestrial life yet: intelligences on the cusp of achieving interstellar travel might be prone to thinking that with the galaxies boiling away in just 10^{19} years, it might be better just to stay home and watch TV.) Thus, if one is trying to build an intelligent machine capable of devising more intelligent machines, it is important to find a way to build in not only motivation, but motivation amplification--the continued desire to build in self-sustaining motivation, as intelligence amplifies. If such motivation is to be possessed by future generations of intelligence--meta-motivation, as it were--then it's important to discover these principles now.

There's a second issue. An intelligent being may be able to envision many more possibilities than a less intelligent one, but that may not always lead to more effective action, especially if some possibilities distract the intelligence from the original goals (e.g., the goal of building a more intelligent intelligence). The inherent uncertainty of the universe may also overwhelm, or render irrelevant, the decision-making process of this intelligence. Indeed, for a very high-dimensional space of possibilities (with the axes representing different parameters of the action to be taken), it might be very hard to evaluate which path is the best. The mind can make plans in parallel, but actions are ultimately unitary, and given finite accessible resources, effective actions will often be sparse.

The last two paragraphs apply not only to AI and ET, but also describe features of the human mind that affect decision making in many of us at times--lack of motivation and drive, and paralysis of decision making in the face of too many possible choices. But it gets worse: we know that a motivation can be hijacked by options that simulate the satisfaction that the motivation is aimed toward. Substance addictions plague tens of millions of people in the United States alone, and addictions to more subtle things, including certain kinds of information (such as e-mail), are prominent too. And few arts are more challenging than passing on motivation to the next generation, for the pursuit of a big idea. Intelligences that invent more and more interesting and absorbing technologies, that can better grab and hold their attention, while reducing impact on the world, might enter the opposite of a singularity.

What is the opposite of a singularity? The singularity depends on a mathematical recursion: invent a superintelligence, and then it will invent an even more powerful superintelligence. But as any mathematics student knows, there are other outcomes of an iterated process, such as a fixed point. A fixed point is a point that, when a function is applied, gives you the same point again. Applying such a function to points near the fixed point will often send them toward the fixed point.

A "societal fixed point" might therefore be defined as a state that self-reinforces, remaining in the status quo--which could in principle be peaceful and self-sustaining, but could also be extremely boring--say, involving lots of people plugged into the

Internet watching videos forever. Thus, we as humans might want, sometime soon, to start laying out design rules for technologies so that they will motivate us to some high goal or end--or at least away from dead-end societal fixed points. This process will involve thinking about how technology could help confront an old question of philosophy--namely, "What should I do, given all these possible paths?" Perhaps it is time for an empirical answer to this question, derived from the properties of our brains and the universe we live in.

Copyright Technology Review 2009.

Upcoming Events

[Lab to Market Workshop \(http://www.technologyreview.com/emtech/09/workshop.aspx\)](http://www.technologyreview.com/emtech/09/workshop.aspx)

Cambridge, MA

Tuesday, September 22, 2009

<http://www.technologyreview.com/emtech/09/workshop.aspx>

<http://www.technologyreview.com/emtech/09/workshop.aspx>

[EmTech 09 \(http://www.technologyreview.com/emtech\)](http://www.technologyreview.com/emtech)

Cambridge, MA

Tuesday, September 22, 2009 - Thursday, September 24, 2009

<http://www.technologyreview.com/emtech> (<http://www.technologyreview.com/emtech>)

[Nanotech Europe 2009 \(http://www.nanotech.net\)](http://www.nanotech.net)

Berlin, Germany

Monday, September 28, 2009 - Wednesday, September 30, 2009

<http://www.nanotech.net> (<http://www.nanotech.net>)

[2009 Medical Innovation Summit \(http://www.ClevelandClinic.org/innovations/summit\)](http://www.ClevelandClinic.org/innovations/summit)

Cleveland, OH

Monday, October 05, 2009 - Wednesday, October 07, 2009

<http://www.ClevelandClinic.org/innovations/summit> (<http://www.ClevelandClinic.org/innovations/summit>)

[Optimizing Innovation 2009 \(http://www.connecting-group.com/Web/EventOverview.aspx?Identificador=6\)](http://www.connecting-group.com/Web/EventOverview.aspx?Identificador=6)

New York, NY

Wednesday, October 21, 2009 - Thursday, October 22, 2009

<http://www.connecting-group.com/Web/EventOverview.aspx?Identificador=6>

<http://www.connecting-group.com/Web/EventOverview.aspx?Identificador=6>