Stuart Russell, Professor of Computer Science, UC Berkeley; Author (with Peter Norvig) of *Artificial Intelligence: A Modern Approach*

**Of myths and moonshine**

“*We switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief*.”

So wrote Leo Szilard, describing the events of March 3, 1939, when he demonstrated a neutron-induced uranium fission reaction. According to the historian Richard Rhodes, Szilard had the idea for a neutron-induced chain reaction on September 12, 1933, while crossing the road next to Russell Square in London. The previous day, Ernest Rutherford, a world authority on radioactivity, had given a “warning…to those who seek a source of power in the transmutation of atoms – such expectations are the merest moonshine.”

Thus, the gap between authoritative statements of technological impossibility and the “miracle of understanding” (to borrow a phrase from Nathan Myhrvold) that renders the impossible possible may sometimes be measured not in centuries, as Rod Brooks suggests, but in hours.

None of this *proves* that AI, or gray goo, or strangelets, will be the end of the world. But there is no need for a proof, just a convincing argument pointing to a more-than-infinitesimal possibility. There have been many unconvincing arguments – especially those involving blunt applications of Moore’s law or the spontaneous emergence of consciousness and evil intent. Many of the contributors to this conversation seem to be responding to those arguments and ignoring the more substantial arguments proposed by Omohundro, Bostrom, and others.

The primary concern is not spooky emergent consciousness but simply the ability to make *high-quality decisions*. Here, quality refers to the expected outcome utility of actions taken, where the utility function is, presumably, specified by the human designer. Now we have a problem:

• The utility function may not be perfectly aligned with the values of the human race, which are (at best) very difficult to pin down.

• Any sufficiently capable intelligent system will prefer to ensure its own continued existence and to acquire physical and computational resources – not for their own sake, but to succeed in its assigned task.

A system that is optimizing a function of *n* variables, where the objective depends on a subset of size *k*<*n*, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable. This is essentially the old story of the genie in the lamp, or the sorcerer’s apprentice, or King Midas: you get exactly what you ask for, not what you want. A highly capable decision maker – especially one connected through the Internet to all the world’s information and billions of screens and most of our infrastructure – can have an irreversible impact on humanity.

This is not a minor difficulty. Improving decision quality, irrespective of the utility function chosen, has been the *goal* of AI research – the mainstream goal on which we now spend billions per year, not the secret plot of some lone evil genius. AI research has been accelerating rapidly as pieces of the conceptual framework fall into place, the building blocks gain in size and strength, and commercial investment outstrips academic research activity. Senior AI researchers express noticeably more optimism about the field’s prospects than was the case even a few years ago, and correspondingly greater concern about the potential risks.

No one in the field is calling for regulation of basic research; given the potential benefits of AI for humanity, that seems both infeasible and misdirected. The right response seems to be to change the goals of the field itself; instead of pure intelligence, we need to build intelligence that is *provably* aligned with human values. For practical reasons, we will need to solve the value alignment problem even for relatively unintelligent AI systems that operate in the human environment. There is cause for optimism, if we understand that this issue is an intrinsic part of AI, much as containment is an intrinsic part of modern nuclear fusion research. The world need not be headed for grief.