Chapter 10A
Colin Allen on Yudkowsky's "Friendly Artificial Intelligence"

Friendly Advice?

Yudkowsky begins with a warning to his readers that "By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it". He ends by reminding us that software written to "Do-What-I-Mean is a major, nontrivial technical challenge of Friendly AI". Yudkowsky suggests a history of over-exuberant claims about AI, commenting that early proponents of the idea that artificial neural networks would be intelligent were engaged in "wishful thinking probably more analogous to alchemy than civil engineering". He indicates that anyone who predicts strongly utopian or dystopian outcomes from superhuman AI is committing the "Giant Cheesecake Fallacy"—the mistake of thinking that just because a powerful intelligence could do something it will do that thing. His message seems to be that we should be neither terrified of superhuman AI nor naive about the challenge of building superhuman AI that will be "nice".

So, what is to be the approach to designing Friendly AI? Yudkowsky characterizes the challenge as one of choosing a powerful enough optimization process with an appropriate target. Engineers, he asserts, use a rigorous theory to select a design and then build structures implementing the calculated designs. But, he cautions, we must beware of two kinds of errors: "philosophical failure", i.e. choosing the wrong target, and "technical failure", i.e. wrongly assuming that a system will work optimally in contexts other than those in which it has been tested.

As heuristics, these can hardly be faulted. But like the classic stockbroker's platitude, "buy low, sell high", they give no practical advice. Yudkowsky's repetition of a an apocryphal story about the failure of a neural network program at classifying photographs of tanks—a story that I remember hearing over 25 years ago—hardly enlightens. (If the advice is "Don't rely on backprop!" this is hardly news.) Likewise, to be told that to "build an AI that discovers the orbits of the planets, the programmers need know only the math of Bayesian probability theory" is facile.

Yudkowsky correctly points out that engineering, like evolution, explores a tiny fraction of design space, but the rest of his story is shallow. Both processes are historically-bound. They work by modification of designs that are received from the past. Engineers do not start only with a target specification, but with a choice of platforms from which to try to reach that target. Inspired engineering sometimes involves taking something that was designed for one context and applying it in another, but always it involves cycles of testing and refinement, and it is far from guaranteeing optimization. Where should those who want to program "Friendly AI" begin?

Yudkowsky cites nothing more recent than 2004, but in the interim many new books and articles have been published, some proposing quite specific

architectures or discussing particular programming paradigms for well-behaved autonomous systems. It would have been nice to know whether Yudkowsky thinks any of this work is on the right track, and if not, why not. If Bayesian theory can discover the orbits of planets, is it suitable for discovering "nice" AI? If not, why not? In describing a developmental neural network approach to AI, Yudkowsky shows a tendency, all too common among writers on this topic, when he asks us to "[f]lash forward to a time when the AI is superhumanly intelligent". We jump straight to sci fi without being given any clue how that flash occurs.

I hoped for more in the context of the present volume, with its stated goal to "reformulate the singularity hypothesis as a coherent and falsifiable conjecture and to investigate its most likely consequences, in particular those associated with existential risks". For our assessment of the existential risks, some knowledge of the current engineering pathways is crucial. If the path to Friendly AI with superhuman intelligence goes through explicit, top-down reasoning the existential risks may be rather different than if it goes through implicit, bottom-up processes. Different kinds of philosophical and technical failures are likely to accompany the different approaches. Similarly, if the route to superhuman AI runs through our self-driving automobiles, the risks may be rather different than if they run through our battle-ready military robots. What is clear is that our current understanding of how to build intelligent machines is low, but we have only the vaguest ideas about how to make it high.