The Golden Rules of AGI Design

Aaron Turner
aaron.turner@bigmother.ai https://bigmother.ai

This is not a paper. It's a collection of thoughts that may one day find their way into a paper via a blog.

The BigMother project is guided by the following AGI design principles

Rule G – Everything must be general

In AGI, the "G" is paramount. Every component of an AGI must be defined as **generally** as possible.

Rule R – Everything must be rational

For an AGI to be maximally safe, benevolent, and trustworthy, it must first be maximally **rational**. In AGI, therefore, everything (beliefs, computations) must be rational. For the rest, we have humans.

Rule H – We're *not* building a human!

Humans, as a species, are **irrational** (rather than proclaiming what a miracle of evolution the human mind is, it would be much more accurate to describe humans as only slightly less dumb than the rocks that evolution started out with), so a perfect emulation of the human mind will be irrational. As a corollary, any AGI that is modelled too closely on humans risks inheriting that irrationality. Thus **we're not building a human**. To build a human, all you need is another human (of the contradictory gender).

Rule C – Don't conflate intelligence/cognition with human (or biological) intelligence/cognition

Just because human (or biological) cognition works in a particular way, and has particular attributes, does not mean that cognition in general (or AGI) must necessarily work in the same way and have the same attributes (e.g. consciousness, emotion, embodiment, etc). This fallacy is a direct consequence of using human intelligence/cognition (implicitly or otherwise) as the de facto archetype for machine intelligence/cognition ("human-like" intelligence etc, as AGI is most often described), which is an easy trap to fall into when you don't understand intelligence/cognition well enough to be able to define or think about these things in any other way. **There is another way**, and that is to think about intelligence, or, more usefully, cognition, as **concepts in their own right, independent of any biological archetype**.

Rule A – Don't assume the solution

Just because the AI field is obsessed with a particular technology doesn't mean that that technology must necessarily form either the basis of an AGI or any part of it. It's in the toolbox, but that doesn't mean you have to use it; there may be good reasons (e.g. safety) not to. Don't **assume the solution**.

Rule 5 – The last 5%

In many cases, it will be possible to achieve 95% of what you want (a "95% solution") via a quick-and-easy superfast-ROI low-hanging-fruit solution (such as scraping loads of data from the internet and pushing it through loads of compute) that displays a **shallow semantic understanding** of a task rather than a **deep semantic understanding** of it (i.e. **shallow intelligence** vs **deep intelligence**). Generally speaking, (a) a **super-intelligent** (i.e. "100%") solution will require **deep intelligence** (a deep semantic understanding), and (b) the first 95% is relatively easy to achieve, the last 5% is really, really hard to achieve. However, as in evolution (humans vs all other life on Earth), the **smartest AGI** (i.e. the 100% solution possessing a deep semantic understanding) **will always prevail in the long run**. And so, when faced with multiple design alternatives, always prioritise whatever promises the deepest semantic understanding, even if it takes one or two orders of magnitude more effort to implement. In the long run, the quick-and-easy method will hit a semantic brick wall, but the more difficult method will not.

Rule L – Don't be distracted by low-hanging fruit

As a corollary to Rule 5, don't be distracted by **low-hanging fruit** (some specific, non-general problem that is easier to solve, and/or which promises a quick return on investment, either financial or reputational), unless it is on a clear path (well-defined roadmap) to a general solution, no matter how shiny it seems, or how easy it may be to rationalise doing so. The sugar high will only be temporary!

Rule F1 – Stay focussed

There is a deeply insidious but nevertheless very real risk that the project will at some point **collapse into Brownian motion** (attempting to go in every direction at once, without making any significant forward progress), greatly exacerbated by the project's decades long timescales. Without constant vigilance, the likelihood that the project will survive this risk over many decades is vanishingly small.

Rule F2 – Stay fair – TODO - reword

There is a deeply insidious but nevertheless very real risk that the project will at some point **collapse into politics**, motivated (often subconsciously) by the innate human desire to pursue short-term self-interest, particularly in light of the unprecedented potential for competitive advantage promised by AGI (which more and more people will gradually realise as time goes by). Thus sovereign nation states, commercial organisations, special interest groups, political parties, and even project contributors such as donors and team members will increasingly attempt (more or less subtly) to influence the project for their own selfish ends rather than in the best interests of mankind as a whole. Without constant vigilance, the likelihood that the project will survive this risk over many decades is vanishingly small.