Aaron Turner — Personal Research Agenda

aaron.turner@bigmother.ai • BigMother Labs (BML) • Cambridge, UK

The ASI Endgame Problem

Armed with a 1983 BSc in Computer Science from Queen Mary College, London, I have been addicted to AGI — as we now call it — since 1985, and have spent the vast majority of the subsequent 40+ years as an "independent AGI researcher"[1], striving to understand AGI — and its inevitable successor, ASI — a little more deeply with each year that has passed. Prior to 2018, I was primarily focused on how it might be possible to construct a computer-based system (a cognitive architecture for AGI) capable of genuine machine cognition. In 2018, I founded a 1-person non-profit AGI lab (BigMother Labs) as the vehicle for my efforts, and my focus then shifted from mere **capability** (i.e. "raw" cognition) towards **alignment** and **validation**. After four decades, I am now committed to spending the entirety of my remaining life addressing the **ASI Endgame Problem** as defined below:

- for my purposes, the **ASI Endgame** (AE) is the point in time at which humanity develops **fully autonomous systems** (able to sense and affect the physical world) that are **more intelligent than any human**; the net effect of the ASI Endgame is that humanity will have introduced a **new "species"** — **ASI** — such that humans will no longer be the most intelligent species on Earth; the prudent assumption is that this will be an **irreversible trapdoor** moment in human history
- the **ASI Endgame Problem** (AEP) is as follows: "How can we ensure that the ASI Endgame that **actually transpires** is as close as possible (and ideally identical) to the **ideal** ASI Endgame (from the perspective of the human species as a whole)?"; assuming the AE is an irreversible trapdoor, we have one chance to get it right, after which the fate of all humanity will be irrevocably sealed
- this is a **civilisation-level** problem, requiring not just **academic theories**, but **actionable plans**.

Please note that Professor Emeritus Leslie Smith (University of Stirling; associate editor of JAGI) has been a priceless source of advice, encouragement, mentorship, and adult supervision since ≈ 2021.

Personal Research Agenda

My **personal research agenda** is to (A) **formulate** a detailed actionable plan which, if executed, will lead to an **actual** AE that is as close as possible to the **ideal** AE (from the perspective of the human species as a whole), and to then (B) **execute that plan** (e.g. via BML, ideally expanded with additional personnel). I do not expect to live long enough to see the plan completed in my lifetime; the best I can hope to achieve is to make sufficient progress that others may complete it after I'm gone.

I view the AEP as an **engineering problem** (or, more exactly, a multi-disciplinary socio-technical-scientific-engineering problem) to be addressed via top-down, breadth-first, recursive decomposition.

As described on the next page, Phase A will generate ≥ 3 papers, which form the input to Phase B.

In an ideal world, I would be able to pursue Phase A as a **PhD candidate**[2], perhaps with the thesis submitted as a series of papers[3] written/published before/during candidature[4]. Strictly speaking, I don't *need* to do a PhD in order to complete Phase A. I *could* produce the planned papers independently, and then publish them on (say) Zenodo, for input to Phase B. Nevertheless, given the arguably disproportionate weight that the AI field's various gatekeepers[5] tend to give to certain proxy cues, the entire project would most likely be greatly facilitated if I were able to do Phase A as a PhD.

[1] that is, self-funded, self-directed, and largely self-taught
[2] probably part-time, and possibly either fully or partially remote, depending on the distance from Cambridge
[3] plus a connecting synthesis document ("kappa") that ties the papers together and makes the overall doctoral argument
[4] variously called "alternative format", "journal format", or "thesis by published papers", depending on the institution
[5] for example, at AI funding sources, and peer-reviewed AI publications

Phase A — Preliminary work (output: ① -④):

The central concept (unifying **ASI R&D** and **AGI Governance**) is **Gold-Standard (GS) AGI**, defined as:

- **practically-maximally-aligned (PMA)** — impossible[a] to be more strongly aligned with humans, and
- **practically-maximally-validated (PMV)** — impossible[a] to be more convincingly proved correct.

If every deployed ASI is a GS AGI then the actual AE will be as close as practically possible to the ideal AE.

ASI R&D:

- ① — **address Normative ASI Superalignment** ("How do we define a final goal $\mathbf{FG}_S$ (for superintelligent agent S) which, if pursued as intended, ensures practical-maximal-alignment?") [1][b]
- ② — **address Technical ASI Superalignment** ("How do we build a practically-maximally-validated agent S that pursues $\mathbf{FG}_S$ as intended?") using the requirements spec (accompanying $\mathbf{FG}_S$) developed in [1] — STATUS: this is what I mostly spent 1985-2018 designing, and is largely already in my head
- ③ — **initial partial implementation** (in Isabelle) of a small but critically-important component of the GS ASI architecture developed in ② (namely, a conservative extension of first-order logic)

[a] in actual practice

[b] "The TTQ paper is certainly a tour de force. Aaron sets out a carefully argued process for producing an AGI in as safe a manner as possible. I hope that people read it and at minimum use it as a check list of things to consider." — Professor Emeritus Steve Young CBE FRS (University of Cambridge)

AGI Governance:

- ④ **Global AGI Governance**, including:
  - 4a draft **International Standard for Supervalidated Existentially-Critical Systems**
  - 4b draft **Global Treaty on Supervalidated Existentially-Critical Systems**

Phase B — Ongoing work (input: ① -④):

Together, the following **AGI Governance** and **ASI R&D** tracks ensure that $\forall x : ASI(x) \Rightarrow GS(x)$

AGI Governance Track[a]:

The **AGI Governance track** will continue (indefinitely) until $\forall x : AGI(x) \Rightarrow GS(x)$, as follows:

- build a global network of **competent Supervalidated AGI certification authorities** based on 4a
- **lobby for adoption of the Supervalidated AGI Treaty** 4b by ultimately all UN Member States.

[a] focused on AGI (rather than ASI) in order to provide a safety-buffer around ASI

ASI R&D Track:

The **ASI R&D track** will continue (indefinitely) until $\exists x : ASI(x) \wedge GS(x)$, as follows:

- **complete the PMV implementation** of the GS ASI architecture partially implemented in ③
- **refine said implementation** until formally certified as GS ASI satisfying 4a as required by 4b

These two tracks — persisting over many decades — will require **highly-specialised project management**.

References

[1] Turner, A. *An Implementation-Neutral Solution to the Normative ASI Superalignment Problem.* 2026. Preprint available at https://doi.org/10.5281/zenodo.24094915.