Clarifying some key hypotheses in AI alignment

Suggested usage

First, note this is not exactly a flowchart, nor a tree. Not every node has "yes" and "no", lest it grow and branch excessively, and there are multiple starting points. The intention is to look at different sub-diagrams or paths that are interesting or important to you at any given time.

1. Take a zoomed-out overview.
2. Choose a box that particularly interests you.
3. Follow the arrows up or down from the box. To avoid getting overwhelmed, focus on one connection at a time.
4. If you are interested in learning more or reading author comments about a particular box, look it up in the written version. Use the link on the title of a box to take you straight to the corresponding heading.

Interpretation

Arrows:

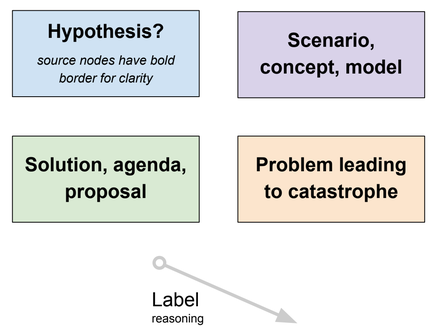
- Question to X: The **closer** your belief is to answering the connected question with the arrow label, the **more** it supports X. For example, the more you believed in the incentive for agentive AGI, the more you would believe agentive AGI will arise, all else equal.
- Question to question: The **closer** your belief is to answering the tail question with the tail label, the **more** it supports "yes" to the head question.
- Scenario to X: Given yes/no to the scenario, X is more likely.

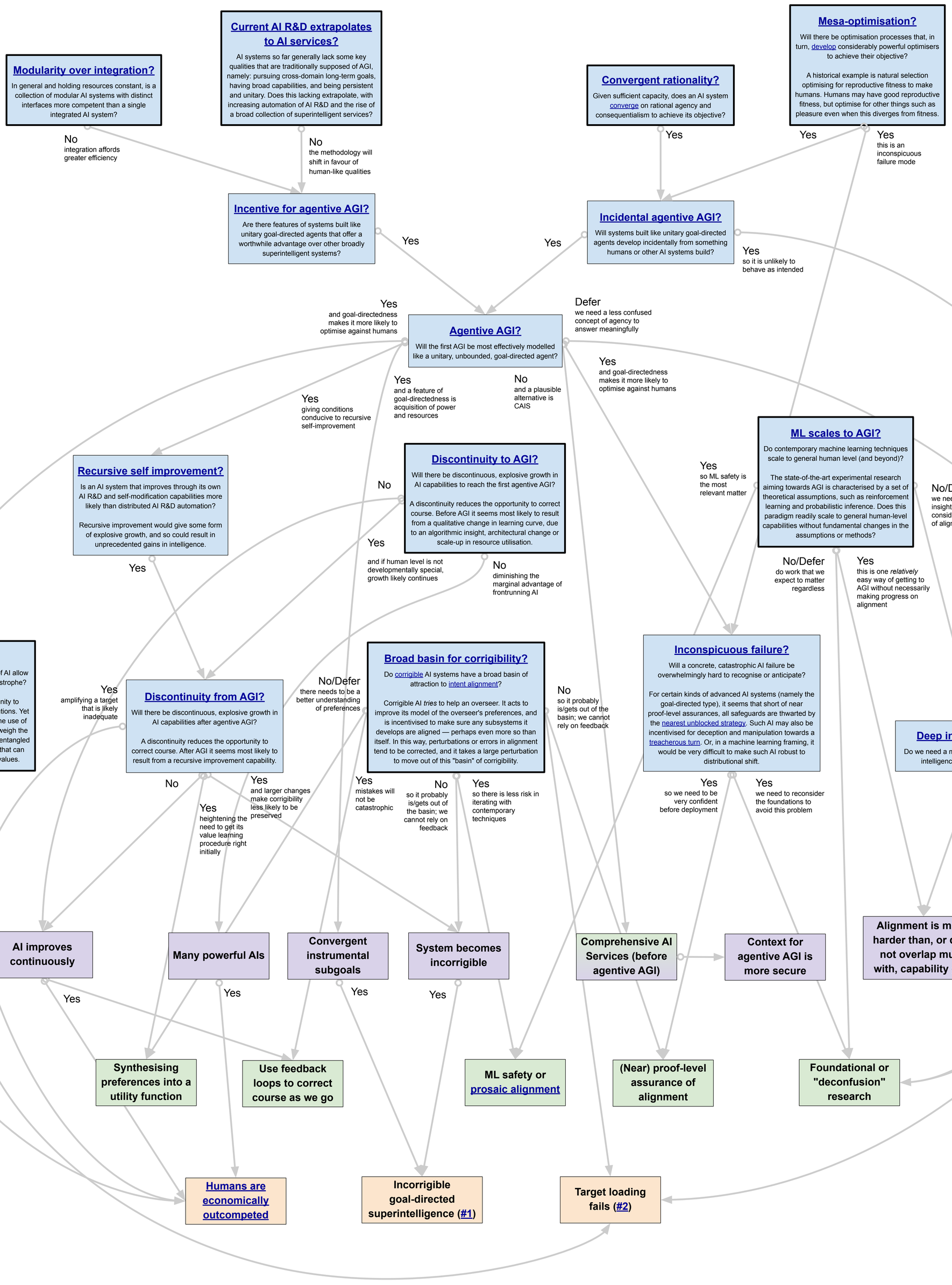
This diagram highlights **key** hypotheses within some areas of AI alignment. Hypotheses that do not seem debated **and** important are omitted.

Definitions

- AGI: a system (not necessarily agentive) that, for almost all economically relevant cognitive tasks, *at least* matches any human's ability at the task. Here, "agentive AGI" is essentially what people in the AI safety community usually mean when they say AGI. References to before and after AGI are to be interpreted as fuzzy, since this definition is fuzzy.
- CAIS: comprehensive AI services. See Reframing Superintelligence.
- Goal-directed: describes a type of behaviour, currently not formalised, but characterised by generalisation to novel circumstances and the acquisition of power and resources. See Intuitions about goal-directed behaviour.

Key





by Ben Cottier and Rohin Shah

Thanks to Stuart Armstrong, Wei Dai, Daniel Dewey, Eric Drexler, Scott Emmons, Ben Garfinkel, Richard Ngo and Cody Wild for helpful feedback on drafts of this work. Ben especially thanks Rohin for his generous feedback and assistance throughout its development.