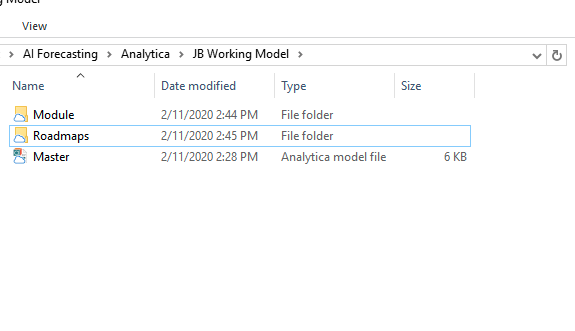
Aryeh and Lauren,

For reference, the “Clarifying some key hypotheses in AI alignment” chart by Rohin Shah and Ben Cottier can be found through the following [link](https://www.lesswrong.com/posts/mJ5oNYnkYrd4sD5uE/clarifying-some-key-hypotheses-in-ai-alignment#ML_scales_to_AGI_) and then clicking on “Diagram.” This chart, of course, is what we are basing our Analytica model on.

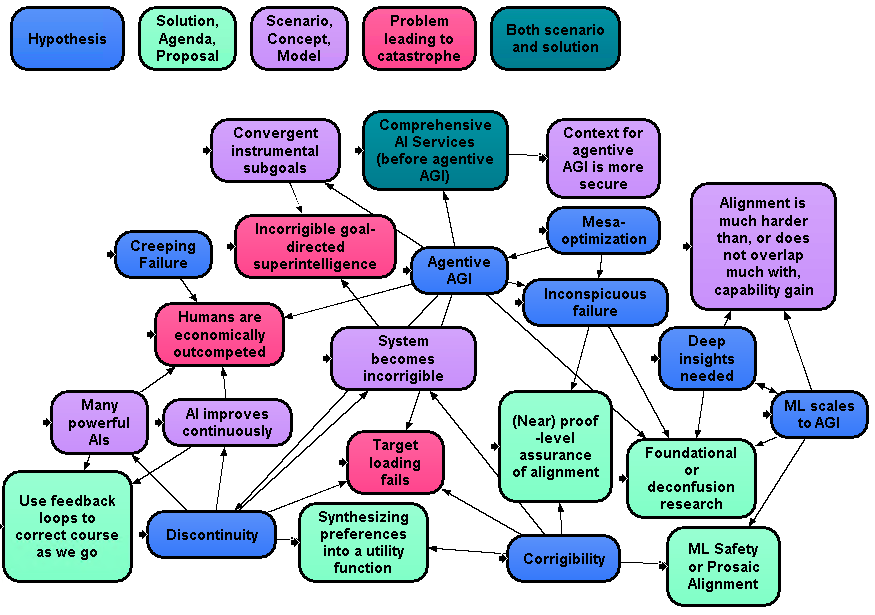
The current form of the Analytica model can be found in the FATE box folder:

*AI Forecasting\Analytica\JB Working*

Inside the folder you will see three things:



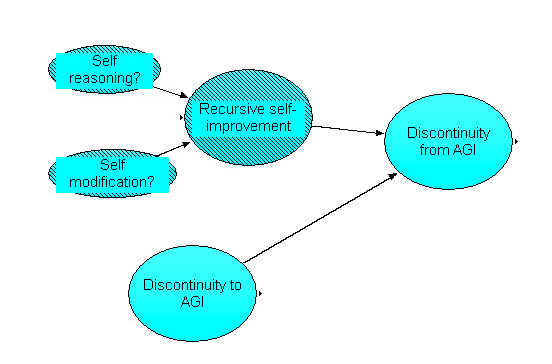
If you open the *Master* file, it will display the main Analytica tree that is most similar to the chart by Rohin Shah and Ben Cottier.



The top row functions as the index (i.e. the blue nodes are hypotheses, green are solutions/agendas/proposals, etc.)

Each node is a module. A module is an externally-created Analytica file embedded into an Analytica object. In other words – modules enable different files to be linked to each other. Therefore, the chart we have is, in effect, a group of Analytica files pointing to one another. Each module will consist of nodes that determine the outcome of the node.

For example, if we double click on the “Discontinuity” module, the following will pop up:



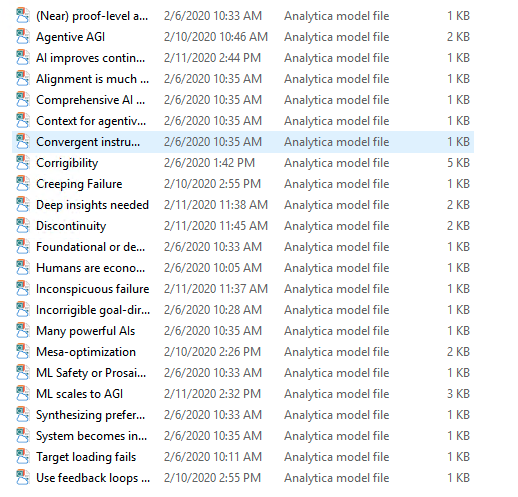
Not only does this contain the three nodes from the original chart by Shah and Cottier (Discontinuity to AGI, Discontinuity from AGI, and Recursive self-improvement), it also contains the self-reasoning and self-modification nodes that I added. Thus, everything found in the Discontinuity module will be related to whether or not there is a discontinuous jump in AI capabilities.

All of the deviations between the original chart and the current manifestation (such as the self-improvement/modification nodes) will be relayed at the end of this document.

For now, close out of Analytica and reopen back to the cloud folder:

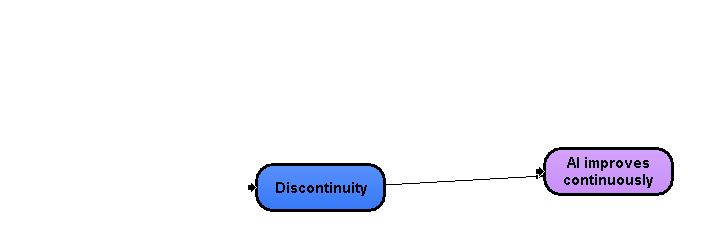
*AI Forecasting\Analytica\JB Working Model*

Now, if we go into the *Module* folder, we will see a long list of files.



Each file is a separate module from the master file that contains the various nodes related to the topic. For example, if we open up the file *Discontinuity*, we will find the same exact chart as we found in the master. Therefore, if you edit this file, then the displays would be shown on *Master*, too. Thus, we are all able to work on separate modules concurrently (so long as we edit them directly from their module file, not the from the master file.)

Last but not least, we have the *Roadmaps* folder. Go inside, and open the file titled *AI improves continuously roadmap*.



Eventually, there will be a roadmap for each scenario, solution, and problem (the green, purple, and red nodes). This provides a way to look at the scenarios and how they arise in a cleaner fashion than the master chart. Once again, if you double click the nodes, you will find they link back to the same modules that are on the master.

Note the arrow on the left side of discontinuity. This simply means that it is a module from another file – pay it no heed.

The organization scheme is to be used as a jumping off point, and as we gain more clarity with Analytica, I expect it to change.

In conclusion, we have the *Master* file, which is linked to the many modules in the *Module* folder, as well as the fledgling *Roadmaps* folder which will contain the simplified versions of how each outcome occurs. All that is left to do is go over the changes I have made to the chart from the differences with Shah and Cottier.

If you have any questions about the above or below, feel free to send an email.

Changes to the chart, by Module:

(Note: These are not exhaustive tweaks to the chart – even corrigibility, which is the most delved into, needs to be further expanded.)

All sources come from links in the “Clarifying some key hypotheses in AI alignment” – however, the articles have not been read to the depth/exhaustion necessary. This is just a beginning effort.

The top level bullets are the original nodes from Shah/Cottier, and the sub-bullets are the appended nodes.

One final thought: As I added nodes to the chart, I did not cite my sources. Going forward, it would be a good idea to post the link to each node in a central place, as to be able to cross-reference/look-up as necessary. As I go forward and see the ideas recur, I will link them to the nodes in a separate document. For now, know that they all come from Clarifying some key hypotheses in AI alignment’s sublinks.

* Mesa Optimization – The information comes from the first two and final post of: <https://www.lesswrong.com/s/r9tYkB2a8Fp4DN8yB?_ga=2.142046961.297008699.1581452092-898759574.1578940291>   
  Therefore, the other posts can be looked into for probable addition to the chart.
  + “Robust to Scale” – will Mesa-optimization be robust to scale, as in, as the system gains resources, will it and all of its components be able to properly scale up, and if it loses resources, will the components properly scale down.
    - <https://www.alignmentforum.org/s/Rm6oQRJJmhGCcLvxh/p/ChierESmenTtCQqZy>
  + “Formalize Optimization Definition” – because mesa-optimization is tied in with optimizing, a formalization of optimization would be useful.
  + “Formalize conditions required for mesa-optimization” – formalization clarifies feasibility.
  + “Subcomponents enable self-modification” – The extent to which an agent can modify itself/its components can modify themselves/each other and to what extent, all plays into the viability of mesa-optimization
  + “Subcomponents enable self-reasoning” – limits of self-reasoning will play a bearing on mesa-optimizations.
  + “Can system enforce subsystem to be aligned?” – will the overarching system, on its own, develop mechanisms to make sure the subsystems are aligned?
* Inconspicuous failure – All sources are from main article in the “Clarifying some key hypotheses in AI alignment” (I have not read into the sub-bulleted articles)
  + “Interpretability” -- interpretability of results will play into how clear or not the failure of a system.
  + “Verification” – whether or not the AI can be verified provides clarity in event of failure.
* Deep insights needed
  + “Meta-learning”
  + “Currently developed technical solutions may be irrelevant”
  + “AI development paradigm gap/shift” – On further thought, this may be reversed, i.e. deep insights affect ai development paradigm gap/shift
  + “Ability for machine counterfactual reasoning” – can machines properly harness counterfactual reasoning?
  + “Goal ambiguity” – how do we precisely define goals?
* ML scales to AGI
  + “Growing computation power affects ability to reach AGI” – is AGI an extension of ML by way of compute?
  + “RL scales to AGI” – will RL scale to AGI
  + “MDP can sufficiently describe the world” – can Markov Decision Processes (MDPs) describe the world enough for AGI?
  + “Can discrete states effectively characterize continuous spaces?” – Ties into the above, but more general than just MDPs.
  + “Inverse Reinforcement Learning effectiveness” – does this tie into building AGI? Note: not knowledgeable about IRL.
  + “Effective online learning?” – Can online learning be made effective and put to use in the route to AGI.
  + “One-shot learning feasible” – does one shot learning (learning to recognize something from a single instance) seem to have promise?
* Corrigibility – again, these nodes stem from the “Clarifying some key hypotheses in AI alignment” links, as well as links off shooting those as well. I wish to stress it as unexhaustive.
  + “Broad basin for corrigibility” – this is the original node from Shah/Cottier, and I did not wish to put it into the other nodes created.
  + “Combine utility functions in a safe manner” – corrigible agents correcting their utility functions safely.
  + “AI reason internally from programmer’s external view” – can AI’s reason about themselves as if they were in the programmer’s pants? (Self awareness?)
  + “Self-repairing safety mechanisms” – can this idea be developed?
  + “Safely communicate AI’s predicted result without risk to operator/humanity” – can an AI only communicate without acting on result in an honest way?
  + “Test mechanisms for corrigibility” – is corrigibility testable?
  + “Humans can share their preferences in an intelligible way to agent” – Similar to online learning, creating a feedback loop that the agent values and uses. This puts the agent on the road to corrigibility
  + “AI can determine side effects of action” – this would enable agents to then reason better about their actions
  + “Agent designed in a way with no preference/pressure to interfere with programmer attempts to modify” – pushes the agent towards honesty/transparency
  + “Corrigibility formalizable?” – if formalizable, then illuminates the ease or difficulty of achieving.
  + “Build agents that prefer to build other agents with overseer’s preferences?” – touches on inner/outer alignment, can agents figure out how to ensure the overseer (operator/programmer/user) preferences are respected, as opposed to the agent itself? I.e. Agent A spawns agent B – will agent B optimize for A’s want, or for the end-user’s want?
  + “Corrigibility Achievable” – is corrigibility achievable?
  + “Agent not deceptive” – Agent is transparent.
  + “Corrigibility affects AI efficiency” – What is the cost of corrigibility? Makes it less worthwhile/used?
  + “Instrumental corrigibility?” – Instrumental corrigibility is an ill-defined term in the literature as far as I can tell – but it nonetheless will impact outcomes.
  + “Corrigibility affects AI cooperating with other AIs” – does corrigibility hamper communication cross AIs?
  + “Agent seeks to preserve corrigibility” – does the agent seek to ensure its corrigible behavior/paradigm is maintained?
  + “Act-based corrigibility” – consider our short term preferences when making its actions.
  + “Robustness of corrigibility” – is it robust, i.e., will being corrigible maintain itself, or will it “slip” out of being corrigible?
* Discontinuity
  + “Self-reasoning?” – can the agent be self-reasoning?
  + “Self-modification?” – can the agent self-modify?