
NE
E
SE
S
SW
W
NW
N

Modeling Transformative AI Risk (MTAIR)

1. Modelling Transformative AI Risks (MTAIR) Project: Introduction
2. Analogies and General Priors on Intelligence
3. Paths To High-Level Machine Intelligence
4. Takeoff Speeds and Discontinuities
5. Modeling Risks From Learned Optimization
6. Modeling the impact of safety agendas
7. Modeling Failure Modes of High-Level Machine Intelligence
8. Elicitation for Modeling Transformative AI Risks

Modelling Transformative AI Risks (MTAIR) Project: Introduction

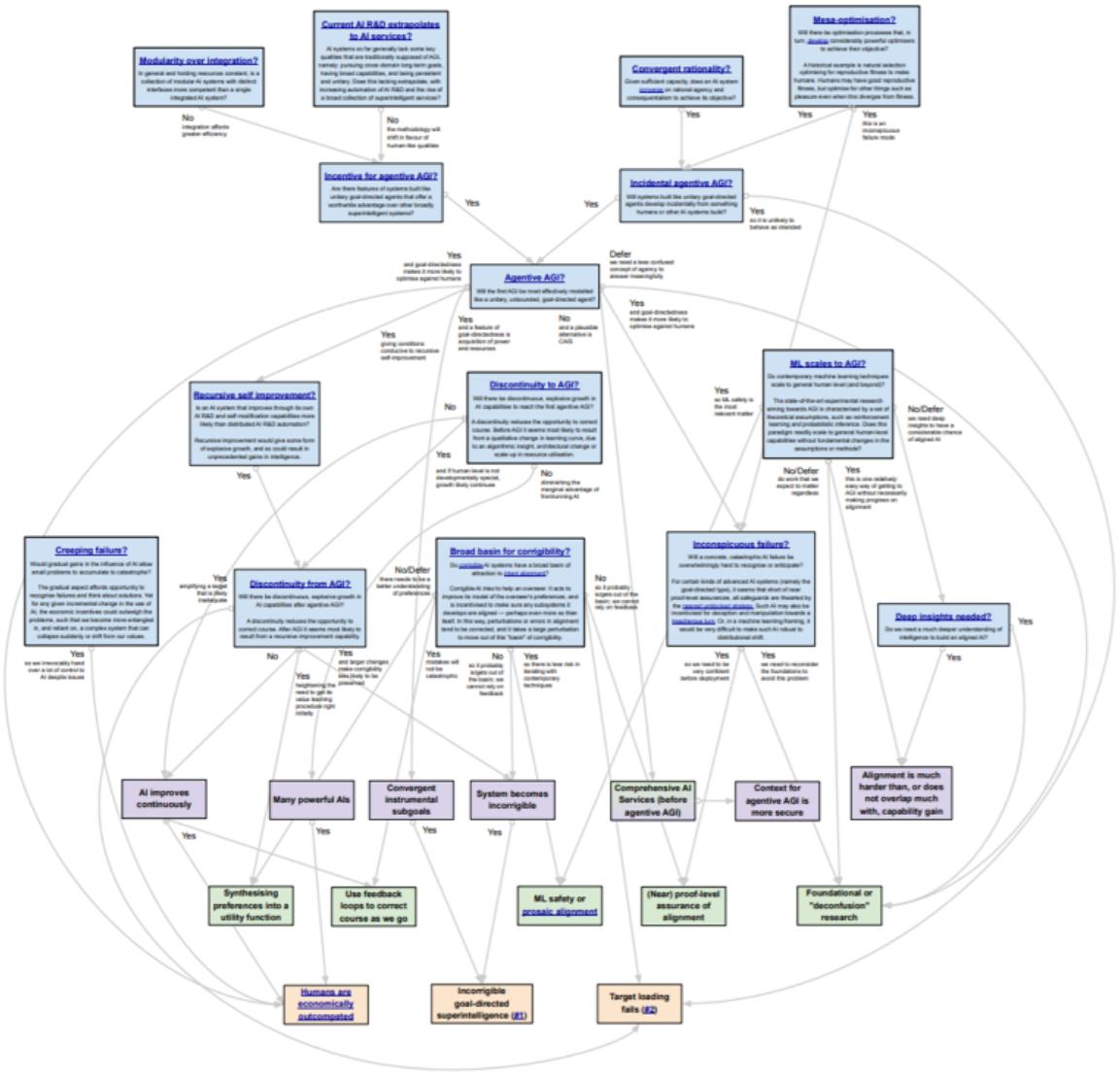
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Numerous books, articles, and blog posts have laid out reasons to think that AI might pose catastrophic or existential risks for the future of humanity. However, these reasons often differ from each other both in details and in main conceptual arguments, and other researchers have questioned or disputed many of the key assumptions and arguments.

The disputes and associated discussions can often become quite long and complex, and they can involve many different arguments, counter-arguments, sub-arguments, implicit assumptions, and references to other discussions or debated positions. Many of the relevant debates and hypotheses are also subtly related to each other.

Two years ago, Ben Cottier and Rohin Shah created a hypothesis map, shown below, which provided a useful starting point for untangling and clarifying some of these interrelated hypotheses and disputes.

The MTAIR project is an attempt to build on this earlier work by including additional hypotheses, debates, and uncertainties, and by including more recent research. We are also attempting to convert Cottier and Shah's informal diagram style into a quantitative model that can incorporate explicit probability estimates, measures of uncertainty, relevant data, and other quantitative factors or analysis, in a way that might be useful for planning or decision-making purposes.



Cottier and Shah's 2019 Hypothesis Map for AI Alignment

This post is the first in a series which presents our preliminary outputs from this project, along with some of our plans going forward. Although the project is still a work in progress, we believe that we are now at a stage where we can productively engage the community, both to contribute to the relevant discourse and to solicit feedback, critiques, and suggestions.

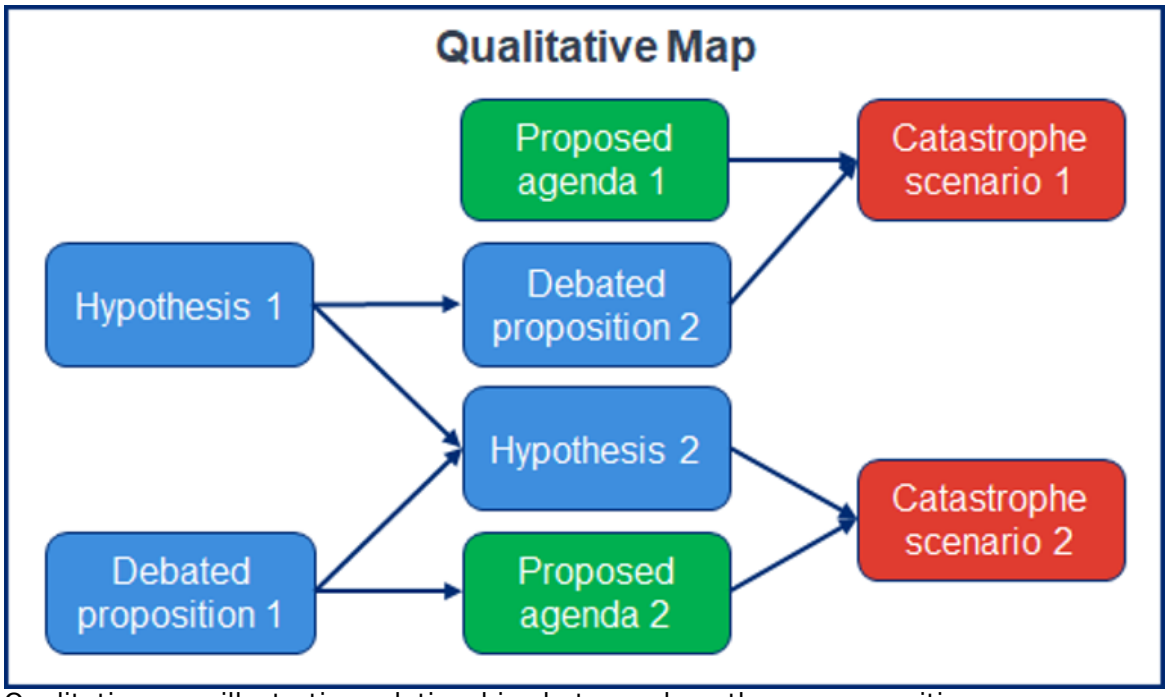
This introductory post gives a brief conceptual overview of our approach and a high-level walkthrough of the hypothesis map that we have developed. Subsequent posts will go into much more detail on different parts of this model. We are primarily interested in feedback on the portions of the model that we are presenting in detail. In the final posts of this sequence we will describe some of our plans going forward.

Conceptual Approach

There are two primary parts to the MTAIR project. The first part, which is still ongoing, involves creating a qualitative map (“model”) of key hypotheses, cruxes, and relationships, as described earlier. The second part, which is still largely in the planning phase, is to

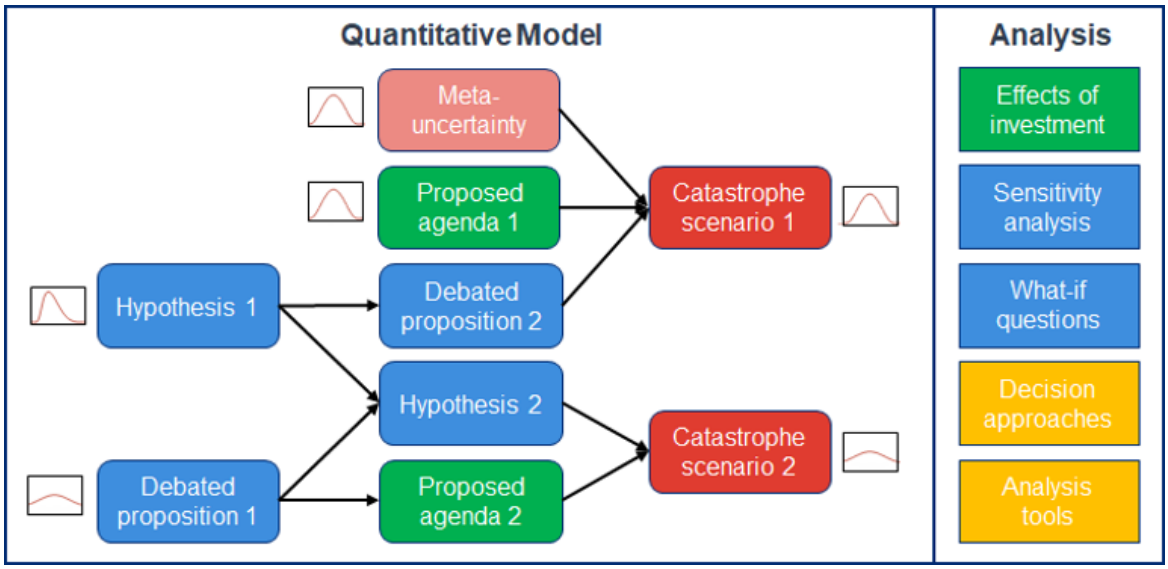
convert our qualitative map into a quantitative model with elicited values from experts, in a way that can be useful for decision-making purposes.

Mapping key hypotheses: As mentioned above, this part of the project involves an ongoing effort to map out the key hypotheses and debate cruxes relevant to risks from Transformative AI, in a manner comparable to and building upon the earlier diagram by Ben Cottier and Rohin Shah. As shown in the conceptual diagram below, the idea is to create a qualitative map showing how the various disagreements and hypotheses (blue nodes) are related to each other, how different proposed technical or governance agendas (green nodes) relate to different disagreements and hypotheses, and how all of those factors feed into the likelihood that different catastrophe scenarios (red nodes) might materialize.



Qualitative map illustrating relationships between hypotheses, propositions, safety agendas, and outcomes

Quantification and decision analysis: Our longer-term plan is to convert our hypothesis map into a quantitative model that can be used to calculate decision-relevant probability estimates. For example, a completed model could output a roughly estimated probability of transformative AI arriving by a given date, a given catastrophe scenario materializing, or a given approach successfully preventing a catastrophe.



Notional version of how the above qualitative map can be used for quantification and analysis

The basic idea is to take any available data, along with probability estimates or structural beliefs elicited from relevant experts (which users can modify or replace with their own estimates as desired). Once this model is fully implemented, we can then calculate probability estimates for downstream nodes of interest via Monte Carlo, based either on a subset or a weighted average of expert opinions, or using specific claims about the structure or quantities of interest, or a combination of the above. Finally, even if the outputs are not accepted, we can use the indicative values as inputs for a variety of analysis tools or formal decision-making techniques. For example, we might consider the choice to pursue a given alignment strategy, and use the model as an aid to think about how the payoff of investments changes if we believe hardware progress will accelerate or if we presume that there is relatively more existential risk from nearer-term failures.

Most of the posts in this series will focus on the qualitative mapping part of the project, since that has been our primary focus to date. In our last post we will discuss our plans related to the second, quantitative, part of the project.

Model Overview

The next several posts in this sequence will dive into the details of our current qualitative model. Each post will be written by team members involved in crafting that particular part of the model, as different team members or groups of team members worked on different parts of the model.

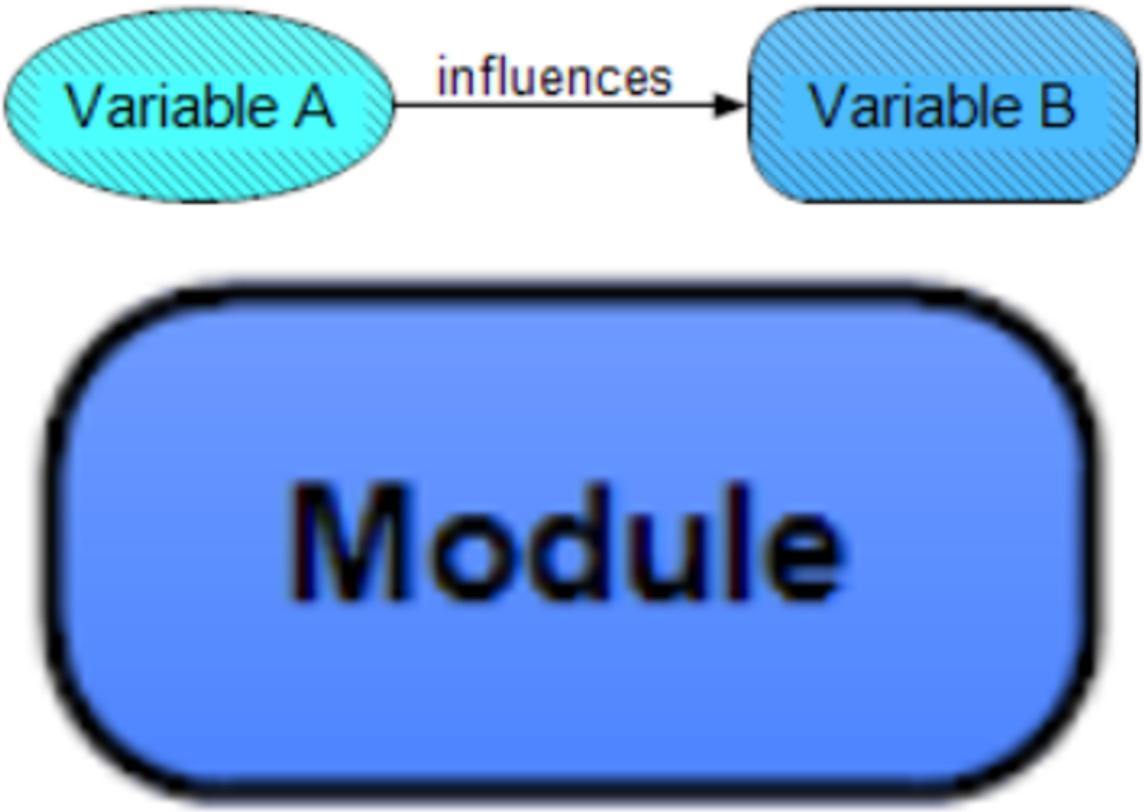
The structure of each part of the model is primarily based on a literature review and the understanding of the team members, along with considerable feedback and input from researchers outside the team. As noted above, this series of posts will hopefully continue to gather input from the community and lead to further discussions. At the same time, the various parts of the model are interrelated. Daniel Eth is leading the ongoing work of integrating the individual parts of the model, as we continue developing a better understanding of how the issues addressed in each component relate to each other.

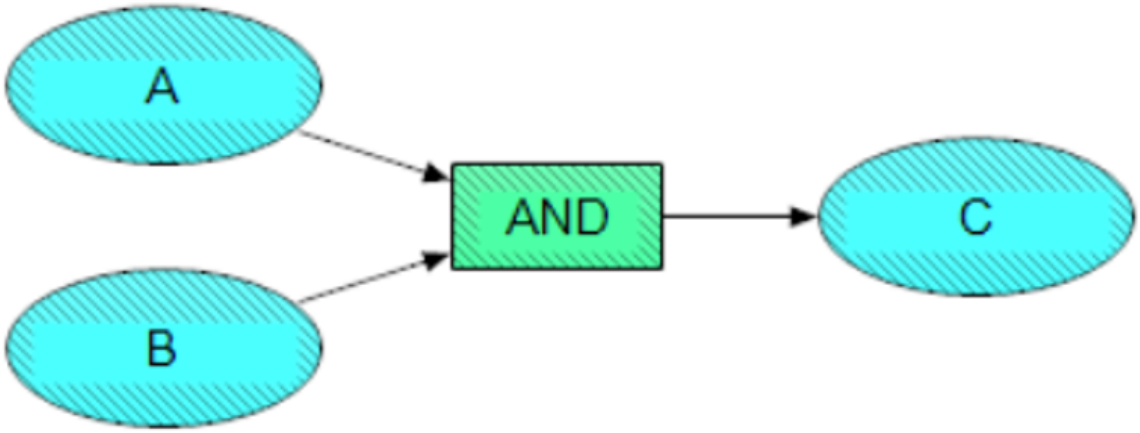
Note on Implementation and Software: At present, we are using Analytica, a “visual software environment for building, exploring, and sharing quantitative decision models that generate prescriptive results.” The models that will be displayed in the rest of this sequence were created using this software program. Note: If you have Windows you can download the

free version of Analytica and once the full sequence of posts is available, we hope to make the model files available, if not publicly, at least on request. To edit the full model you unfortunately need the expensive licensed version of Analytica, since the free version is limited to editing small models and viewing models created by others. There are some ways around this restriction if you only want to edit individual parts of the model - once the sequence has been posted, please message Daniel Eth, David Manheim, or Aryeh Englander for more information.

How to read Analytica models

Before presenting an overview of the model, and as a reference for later posts, we present a brief explanation of how these models work, and how they should be read. Analytica models are composed of different types of nodes, with the relationships between nodes represented as directed edges (arrows). The two primary types of nodes in our model are variable nodes and modules. Variable nodes are usually oval or rounded rectangles without bolded outlines, and correspond to key hypotheses, cruxes of disagreement, or other parameters of interest. Modules, represented by rounded rectangles with bolded outlines, are “sub-models” that contain their own sets of nodes and relationships. In our model we also sometimes use small square nodes to visually represent AND, OR, or NOT relationships. In the software, a far wider set of ways to combine outputs from nodes are available, and will be used in our model - but they are difficult to represent visually.



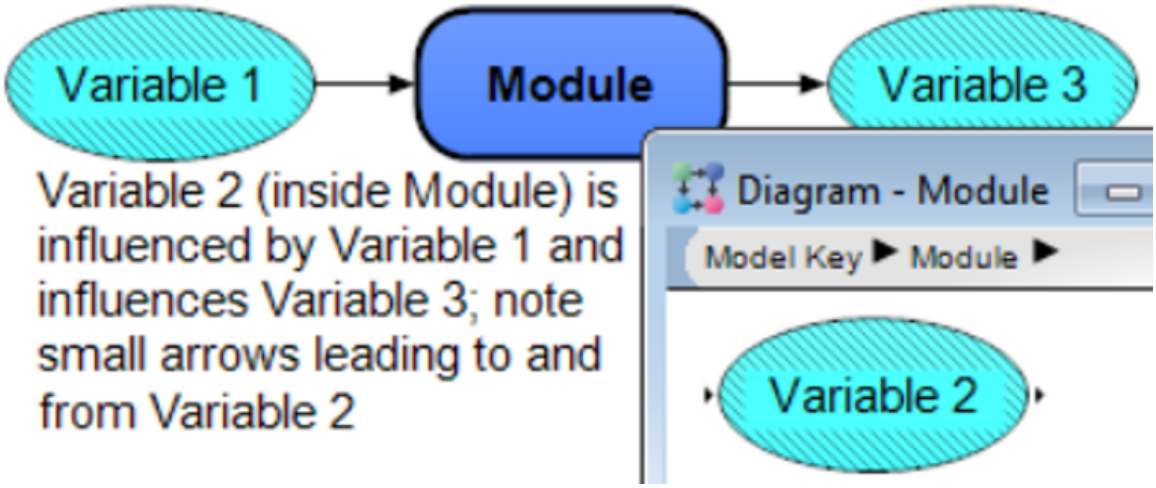


Arrows represent directions of probabilistic influence, in the sense that information about the first node influences the probability estimate for the second node. For example, an arrow from Variable A to Variable B indicates that the probability of B depends at least in part on the probability of A. It is important to note that the model is not a causal model *per se*. An edge from one node to another does not necessarily imply that the first *causes* the second, but rather that there is some relationship between them such that information about the first *informs* the probability estimate for the second. Some edges do represent causal relationships, but only insofar as that relationship is important for informing probability estimates.

Different parts of the model use various color schemes to group nodes that share certain characteristics, but color does not have any formal meaning in Analytica and is not necessary to make sense of the model. The color schemes for individual parts of the model will be explained as needed, but color differences can be safely ignored if they become confusing.

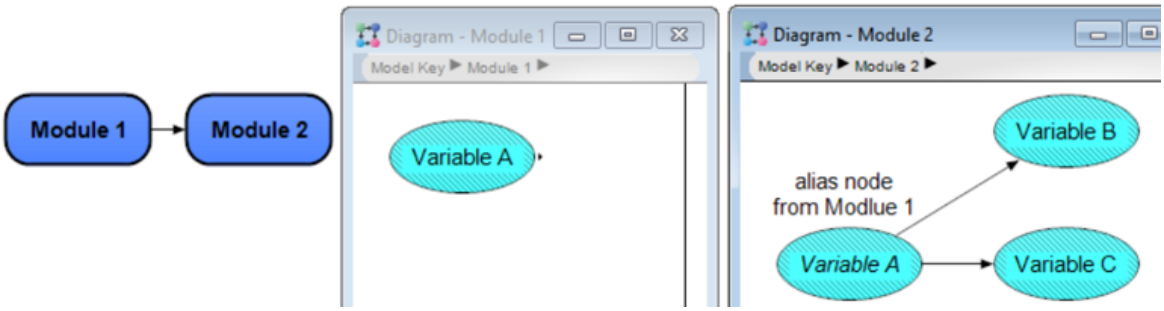
Other things to note:

- In some of the diagrams there are small arrowheads leading into or out of certain nodes, but which do not point to any other node in the diagram. These arrowheads indicate that there are nodes elsewhere in the model that depend on this node or that this node depends on.

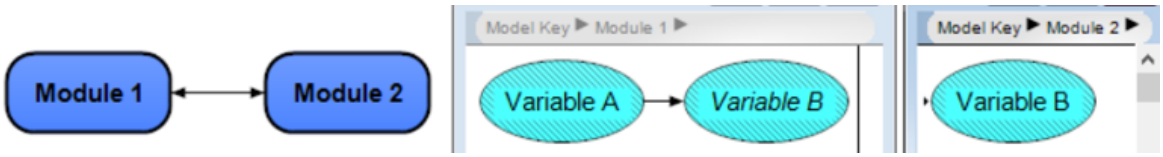


- "Alias nodes" are copies of nodes that link back to the original "real" node, and are mainly useful for display or readability purposes. We use alias nodes in many parts of

our diagrams, especially when a node from one module influences or is influenced by some important node(s) elsewhere in the model. Analytica indicates that a node is an alias by displaying the node name in italics.



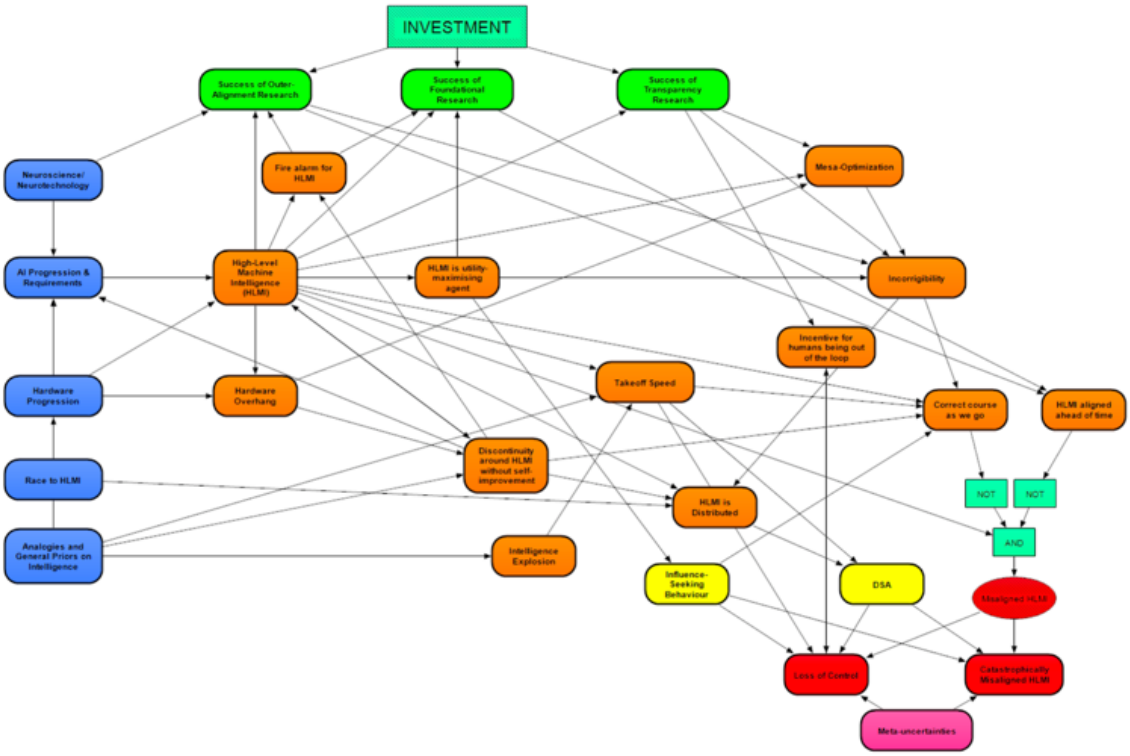
- Our model is technically a directed acyclic graph. However, there are a few places in the model diagrams where Analytica confusingly displays bidirectional arrows between modules even though the direction of influence only goes in one direction. This is because Analytica uses arrows not just to indicate direction of influence, but also to indicate that one module contains an alias node from a different model. For example, the direction of influence in the image below is from Variable A in Module 1 to Variable B in Module 2, but Analytica displays a bidirectional arrow between the modules because Module 1 also contains an alias node from Module 2.



Top-level model walkthrough

The image below represents the top-level diagram of our current model. Most of the nodes in this diagram are their own separate modules, each with their own set of nodes and relationships. Most of these modules will be discussed in much more detail in later posts.

In this overview, we **highlight** key potential nodes and the related questions, and discuss how they are interrelated at a high level. This overview, which in part explains the diagram below, hopes to provide a basic outline of what later posts will discuss in much more detail. (Note that the arrows represent the direction of inference in the model, rather than the underlying causal relationships. Also note that the relationship between the modules reflect dependencies between the individual nodes in the modules, rather than just notional suggestions about the relationships between the concepts represented by the modules themselves.)



High-level model overview

The blue nodes on the left represent technical or other developments or future progress areas that are potentially relevant inputs to the rest of the model. They are: **Neuroscience / Neurotechnology**, **AI Progression & Requirements**, **Hardware Progression**, and **Race to HLMI**. Finally, **Analogies and General Priors on Intelligence**, which address many assumptions and arguments by analogy from domains like human evolution, are used to ground debates about AI takeoff or timelines. These are the key inputs for understanding progress towards HLMI(1).

The main internal portions of the model (largely in orange), represent the relationships between different hypotheses and potential medium-term outcomes. Several key parts of this, which will be discussed in future posts, include paths to **High-Level Machine Intelligence (HLMI)** (and the inputs to it, in the blue modules), **Takeoff/discontinuities**, and **Mesa-optimization**. Impacting these are different safety agendas (along the top in green), which will be reviewed in another post.

Finally, the nodes on the bottom right represent conditions leading to failure (yellow) and failure modes (red). For instance, the possibility of **Misaligned HLMI** (bottom right in red) motivates the critical question of how the misalignment can be prevented. Two possibilities are modelled (orange nodes, right): The first possibility is that HLMI is **aligned ahead of time** (using **Outer Alignment, Inner Alignment** and, if necessary, **Foundational Research**). The second possibility is that we can **'correct course as we go'**, for instance, by using an alignment method that ensures the HLMI is **corrigible.**

While our model has intermediate outputs (which when complete will include estimates of HLMI timelines and takeoff speed), its principal outputs are the predictions for the modules marked in red. **Catastrophically Misaligned HLMI** covers scenarios involving a single HLMI or a coalition achieving a **Decisive Strategic Advantage (DSA)** over the rest of the world and causing an existential catastrophe. **Loss of Control** covers 'creeping failure' scenarios, including those that don't require a coalition or individual to seize a DSA.

The Model is (Already) Wrong

We expect that readers will disagree with us, and with one another, about various points - we hope you flag these issues. At the same time, the above is only a high level overview, and we already know that many items in the above overview are contentious or unclear - which is exactly why we are trying to map it more clearly.

Throughout this work, we attempt to model disagreements and how they relate to each other, as shown in the earlier notional outline for mapping key hypotheses. As a concrete example, whether HLMI will be agentive, itself a debate, influences whether it is plausible that the HLMI will attempt to self-modify or design successors. The feasibility of either modification or successor design is another debate, and this partly determines the potential for very fast takeoff, influencing the probability of a catastrophic outcome. As the example illustrates, the values and the connections between the nodes are all therefore subject to potential disagreement, which must be represented in order to model the risk. Further and more detailed examples are provided in upcoming posts.

Further Posts and Feedback

The further posts in this sequence will cover the internals of these modules, which are only outlined at a very high level here. This is intended to be a sequence that will be posted over the coming weeks, starting with the post on **Analogies and General Priors on Intelligence** later this week, followed by **Paths to HLMI**.

If you think any of this is potentially useful, or if you already disagree with some of our claims, we are very interested in feedback and disagreements and hope to have a productive discussion in the comments. We are especially interested in places where the model does not capture your views or fails to include an uncertainty that you think could be an important crux. Similarly, if the explanation seems confused or confusing, flagging this is useful - both to help us clarify, and to ensure it doesn't reflect an actual disagreement. It may also be useful to flag things that you think are *not* cruxes, or are obvious, since others may disagree.

Also, if this seems interesting or related to any other work you are doing to map or predict the risks, please be in touch - we would be happy to have more people to consult with or who wish to participate directly.

Footnotes

1. Note that HLMI is viewed as a precursor to, and a likely cause of, transformative AI. For this reason, in the model, we discuss HLMI, which is defined more precisely in later posts.

Acknowledgements

The MTAIR project (formerly titled, "AI Forecasting: What Could Possibly Go Wrong?") was originally funded through the Johns Hopkins University Applied Physics Laboratory (APL), with team members outside of APL working as volunteers. While APL funding was only for one year, the non-APL members of the team have continued work on the project, with additional support from the EA Long-Term Future Fund (except for Daniel Eth, whose funding comes from FHI). Aryeh Englander has also continued working with the project under a grant from the Johns Hopkins Institute for Assured Autonomy (IAA).

The project is led by Daniel Eth (FHI), David Manheim, and Aryeh Englander (APL). The original APL team included Aryeh Englander, Randy Saunders, Joe Bernstein, Lauren Ice, Sam

Barham, Julie Marble, and Seth Weiner. Non-APL team members include Daniel Eth (FHI), David Manheim, Ben Cottier, Sammy Martin, Jérémy Perret, Issa Rice, Ross Gruetzemacher (Wichita State University), Alexis Carlier (FHI), and Jaime Sevilla.

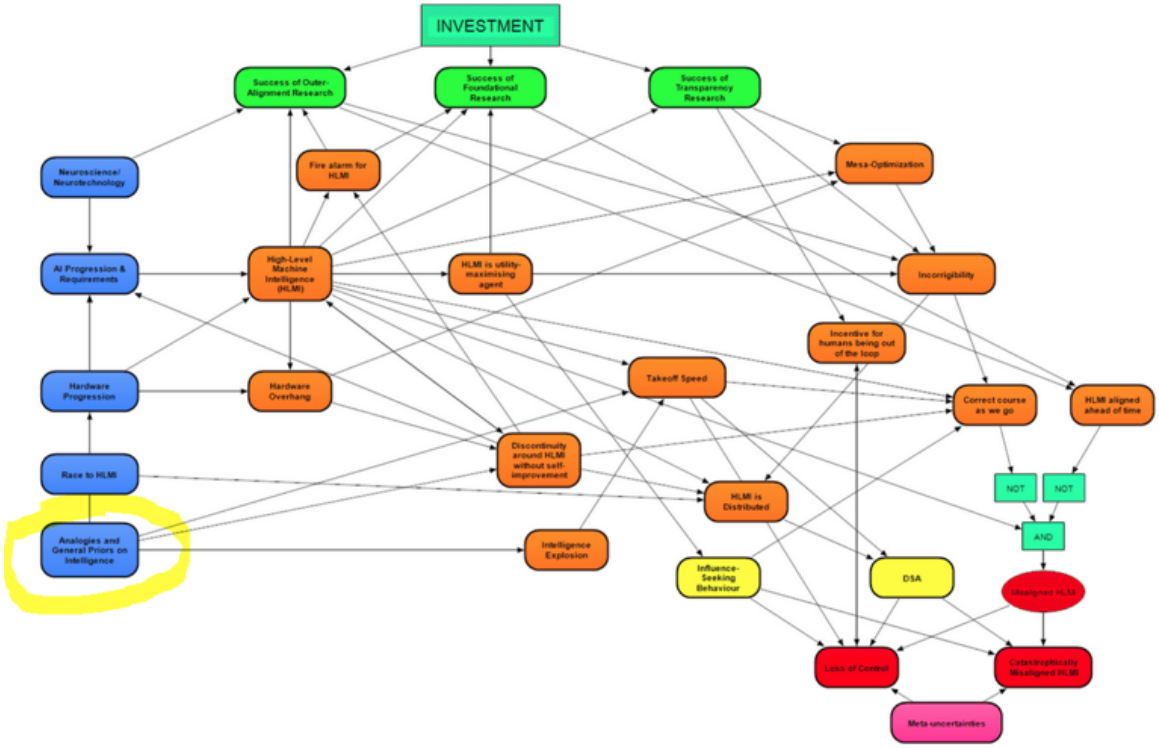
We would like to thank a number of people who have graciously provided feedback and discussion on the project. These include (apologies to anybody who may have accidentally been left off this list): Ashley Llorens (formerly APL, currently at Microsoft), I-Jeng Wang (APL), Jim Scouras (APL), Helen Toner, Rohin Shah, Ben Garfinkel, Daniel Kokotajlo, and Danny Hernandez, as well as several others who prefer not to be mentioned. We are also indebted to several people who have provided feedback on this series of posts, including Rohin Shah, Neel Nanda, Adam Shimi, Edo Arad, and Ozzie Gooen.

Analogies and General Priors on Intelligence

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

*This post is part 2 in our sequence on Modeling Transformative AI Risk . We are building a model to understand debates around existential risks from advanced AI. The model is made with Analytica software, and consists of nodes (representing key hypotheses and cruxes) and edges (representing the relationships between these cruxes), with final output corresponding to the likelihood of various potential failure scenarios. You can read more about the motivation for our project and how the model works in the **Introduction post** . Future posts will explain how different considerations, such as AI takeoff or mesa-optimization, are incorporated into the model.*

This post explains our effort to incorporate basic assumptions and arguments by analogy about intelligence, which are used to ground debates about AI takeoff and paths to High-Level Machine Intelligence (HLMI[1]). In the overall model, this module, *Analogies and General Priors on Intelligence*, is one of the main starting points, and influences modules (covered in subsequent posts in this series) addressing the possibilities of a *Discontinuity around HLMI* or an *Intelligence Explosion*, as well as *AI Progression* and HLMI *Takeoff Speed*.

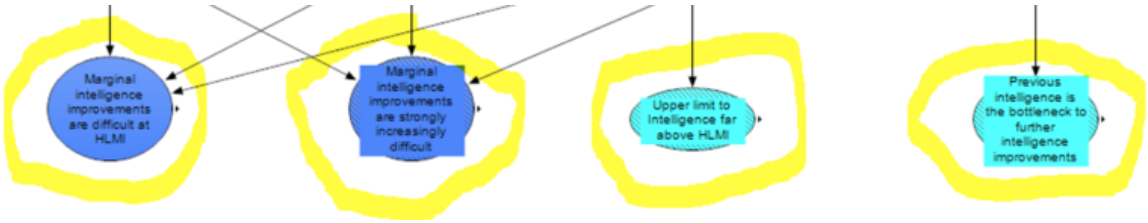


The *Analogies and General Priors on Intelligence* module addresses various claims about AI and the nature of intelligence:

- The difficulty of marginal intelligence improvements at the approximate 'human level' (i.e., around HLMI)
- Whether marginal intelligence improvements become increasingly difficult beyond HLMI at a rapidly growing rate or not

    - ‘Rapidly growing rate’ is operationalized as becoming difficult exponentially or faster-than-exponentially
- Whether there is a fundamental upper limit to intelligence not significantly above the human level
- Whether, in general, further improvements in intelligence tend to be bottlenecked by previous improvements in intelligence rather than some external factor (such as the rate of physics-limited processes)

These final outputs are represented by these four terminal nodes at the bottom of the module.

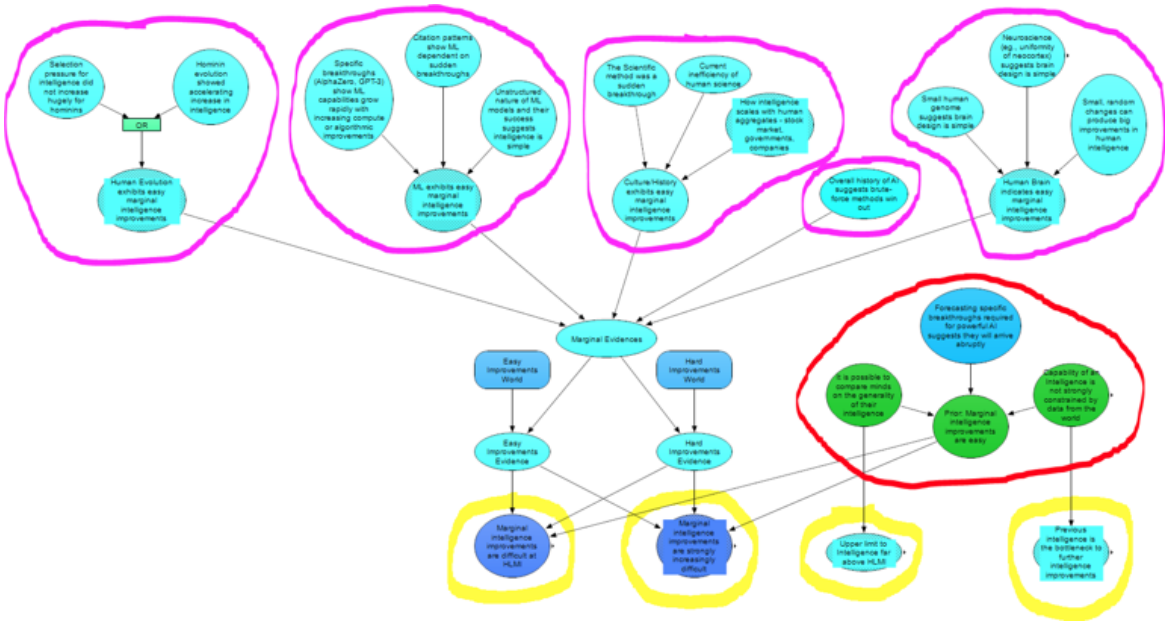


These four claims depend on arguments that analogize the development of HLMI to other domains. Each of these arguments is given a specific submodule in our model, and each argument area is drawn from our review of the existing literature. In the ‘outputs’ section at the end of this article, we explain why we chose these four as cruxes for the overall debate around HLMI development.

Each argument area is shown in the image below, a zoomed out view of the entire *Analogies and General Priors on Intelligence* module. The argument areas are: human biological evolution, machine learning, human cultural evolution, the overall history of AI, and the human brain. Each of these areas is an example of either the development of intelligence or of intelligence itself, and might be a useful analogy for HLMI.

We also incorporate broad philosophical claims about the nature of intelligence as informative in this module. These claims act as priors before the argument areas are investigated, and cover broad issues like whether the concept of general intelligence is coherent and whether the capabilities of agents are strongly constrained by things other than intelligence.

Module Overview

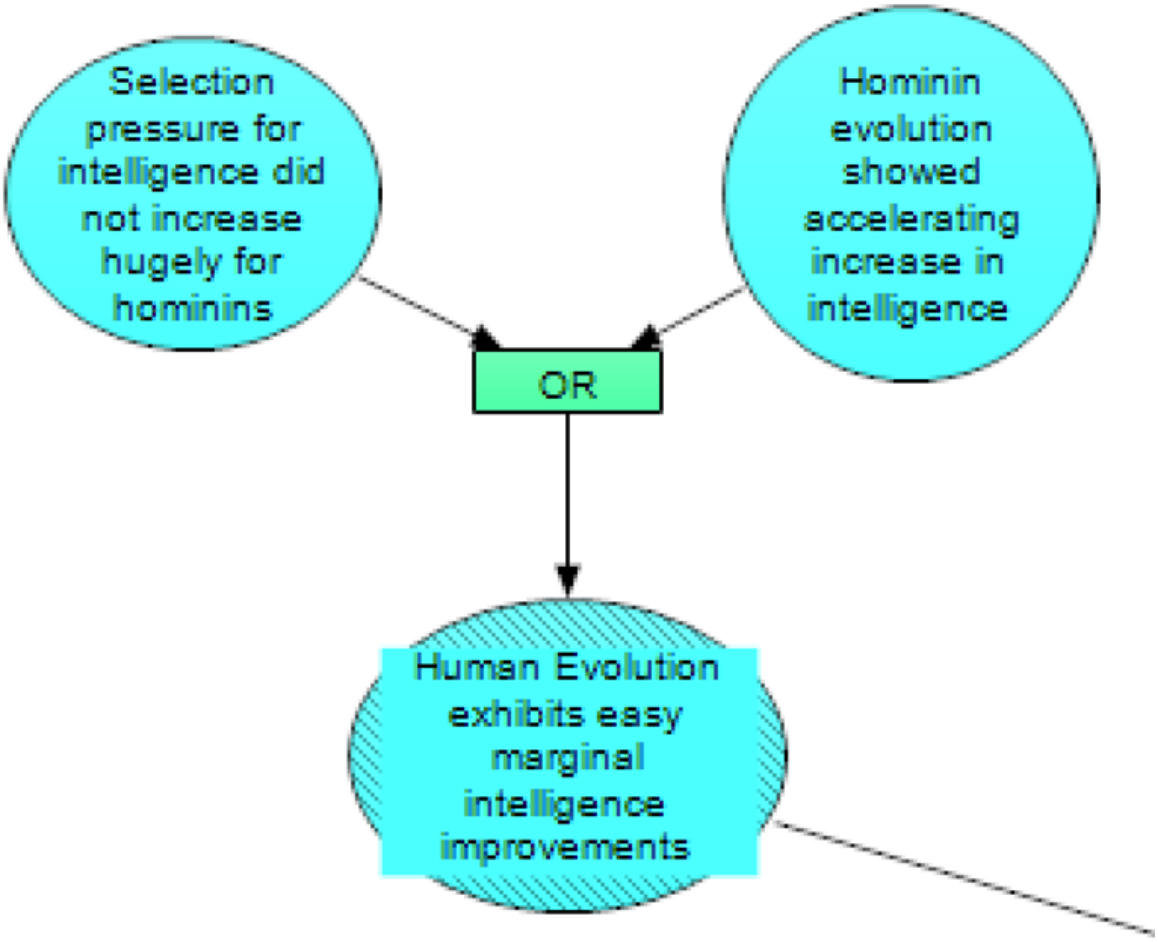


In this diagram of the overall module, the groups of nodes circled in purple are (left to right), Human Evolution, Machine Learning, Human Culture and History, History of AI and the Human Brain, which feed forward to a classifier that estimates the ease of marginal intelligence improvements. Red shows General Priors, and the yellow nodes are the final model outputs as discussed above. We discuss each of these sections in this order in the rest of this post.

The Cruxes

Each of the argument area submodules, as well as the General Priors submodule, contains cruxes (represented by nodes) that ultimately influence the terminal nodes in this module.

Human Evolution



Human evolution is one of the areas most commonly analogised to HLMI development. [‘Intelligence Explosion Microeconomics’] argued that human evolution demonstrated accelerating returns on cognitive investment, which suggests marginal intelligence improvements were (at least) not rapidly increasingly difficult during this process.

There are two sources of potential evidence which could support these claims about evolution. The first (**left-hand node in the module above**) is if hominin evolution did not involve hugely more selection pressure for intelligence than did the evolution of other, ancestral primates. It is generally assumed that hominins (all species more closely related to *Homo sapiens* than to chimpanzees – e.g., all species from *Australopithecus* and *Homo*) saw much faster rises in intelligence than what had happened previously to primates. If this unprecedented rise in intelligence was not due to selection pressures for intelligence changing drastically, then that would provide evidence that this fast rise in intelligence did not involve evolution “solving” increasingly more difficult problems than what came before. On the other hand, if selection pressures for intelligence had been marginal (or less) up until this point and were suddenly turned way up, then the fast rise in intelligence could be squared with evolution solving more difficult problems (as the increase in intelligence could then be less than proportional to the increase in selection pressures for intelligence).

Even if selection pressures for intelligence were marginal before hominins, we could still obtain evidence that human evolution exhibited easy marginal intelligence improvements – if we observe a rapid acceleration of intelligence *during* hominin evolution, up to *Homo sapiens* (**right-hand node in the module above**). Such an acceleration of intelligence would be

less likely if intelligence improvements became rapidly more difficult as the human level was approached.

We must note, however, that the relevance of these evolutionary considerations for artificial intelligence progress is still a matter of debate. Language, for instance, was evolved very late along the human lineage, while AI systems have been trained to deal with language from much earlier in their relative development. It is unknown how differences such as this would affect the difficulty-landscape of developing intelligence. The amount to update based on analogies to evolution, however, is not handled by this specific submodule, but instead by the classifier (mentioned above, described below in the section **The Outputs**).

Sources:

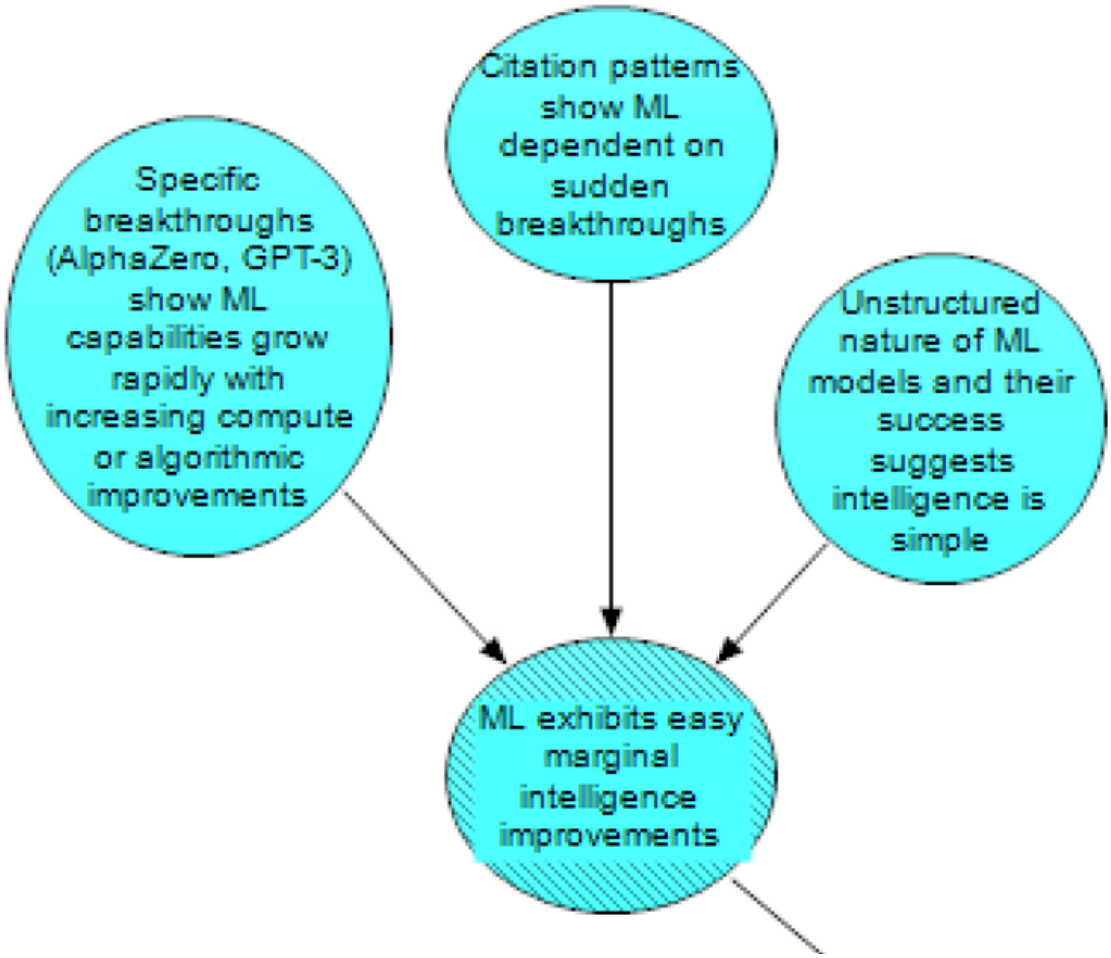
Likelihood of discontinuous progress around the development of AGI

Takeoff speeds – The sideways view

Hanson-Yudkowsky AI Foom Debate

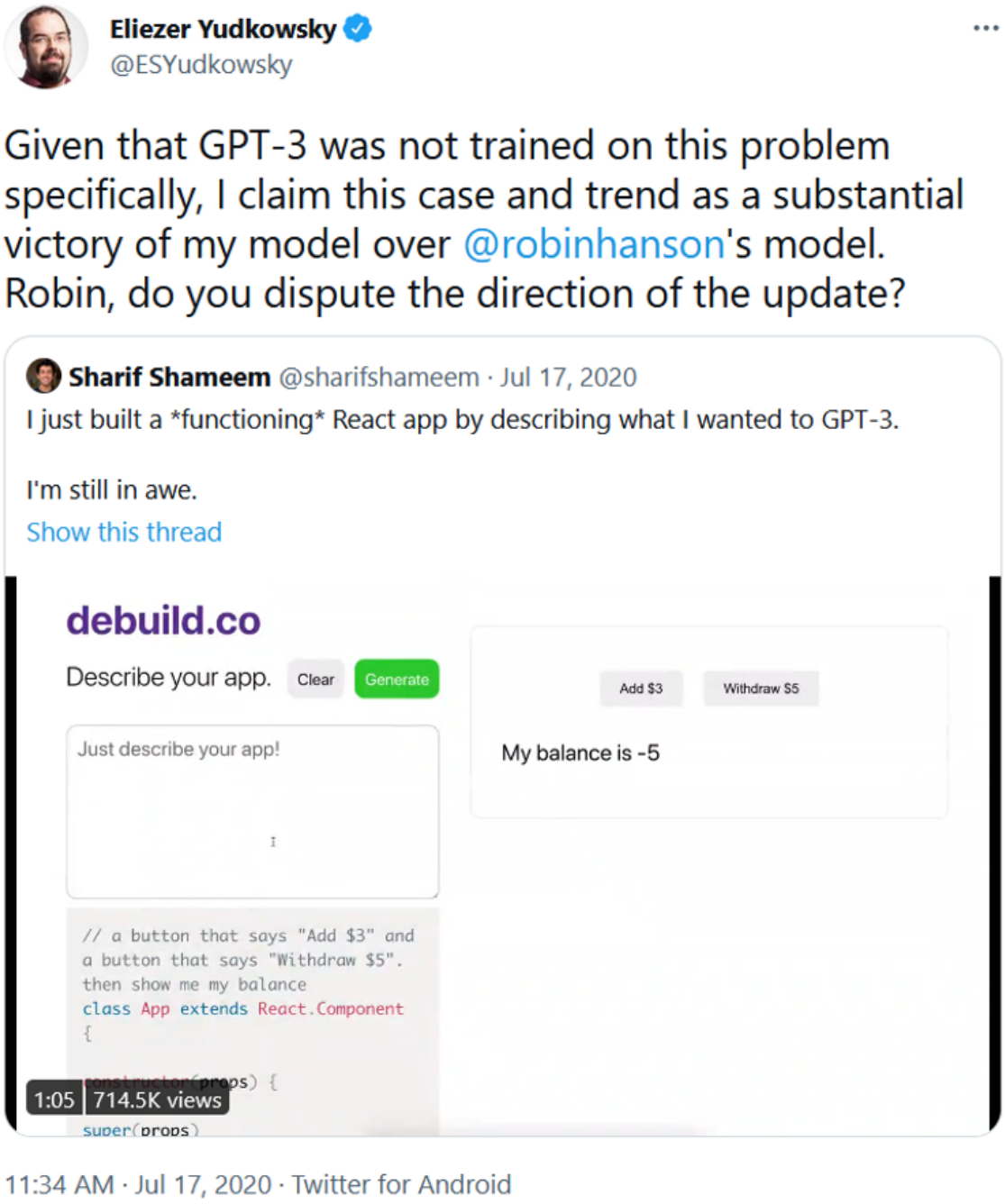
Thoughts on Takeoff Speeds

Machine Learning



Much of the debate around HLMI has focused on what, if any, are the implications of current progress in machine learning for how easy improvements will be at the level of HLMI. Some see developments like AlphaGo or GPT-3 as examples of (and evidence for) the claim that marginal intelligence improvements are not increasingly difficult, though others disagree with those conclusions.

For the first of the ways that we could conclude that current ML exhibits easy marginal intelligence improvements, see this argument:



Which we understand to mean,

> GPT-3 is general enough that it can write a functioning app given a short prompt, despite the fact that it is a relatively unstructured transformer model with no explicitly coded representations for app-writing. The fact that GPT-3 is this capable suggests that ML models scale in capability and generality very rapidly with increases in computing power or minor algorithm improvements...

In order to understand whether claims like Yudkowsky's are true, we must understand the specific nature of the breakthroughs made by cutting-edge ML systems like GPT-3 or Alphazero and the limits of what these systems can do (**left node in the image above**).

Claims about current ML systems are related to a broader, more qualitative claim that the general success of ML models indicates that the fundamental algorithms needed for general intelligence are less complex than we might think (**right node in the image above**). Specific examples of 'humanlike' thinking or reasoning in neural networks, for example OpenAI's discovery of Multimodal Neurons, lend some support to this claim.

Alternatively, Robin Hanson claims that if machine learning was developing in sudden leaps, we would expect to see a pattern of citations in ML research where a few breakthrough papers received a very disproportionately large amount of citations. If Hanson is right about this, and in reality citations aren't distributed in an unusually concentrated pattern in ML compared to other fields, then we have reason to expect marginal intelligence improvements from ML are hard (**middle node in the image above**).

Sources:

Hanson-Yudkowsky AI Foom Debate

Searching for Bayes-Structure - LessWrong 2.0 viewer

Will AI undergo discontinuous progress?

Conceptual issues in AI safety: the paradigmatic gap

The Scaling Hypothesis · Gwern.net

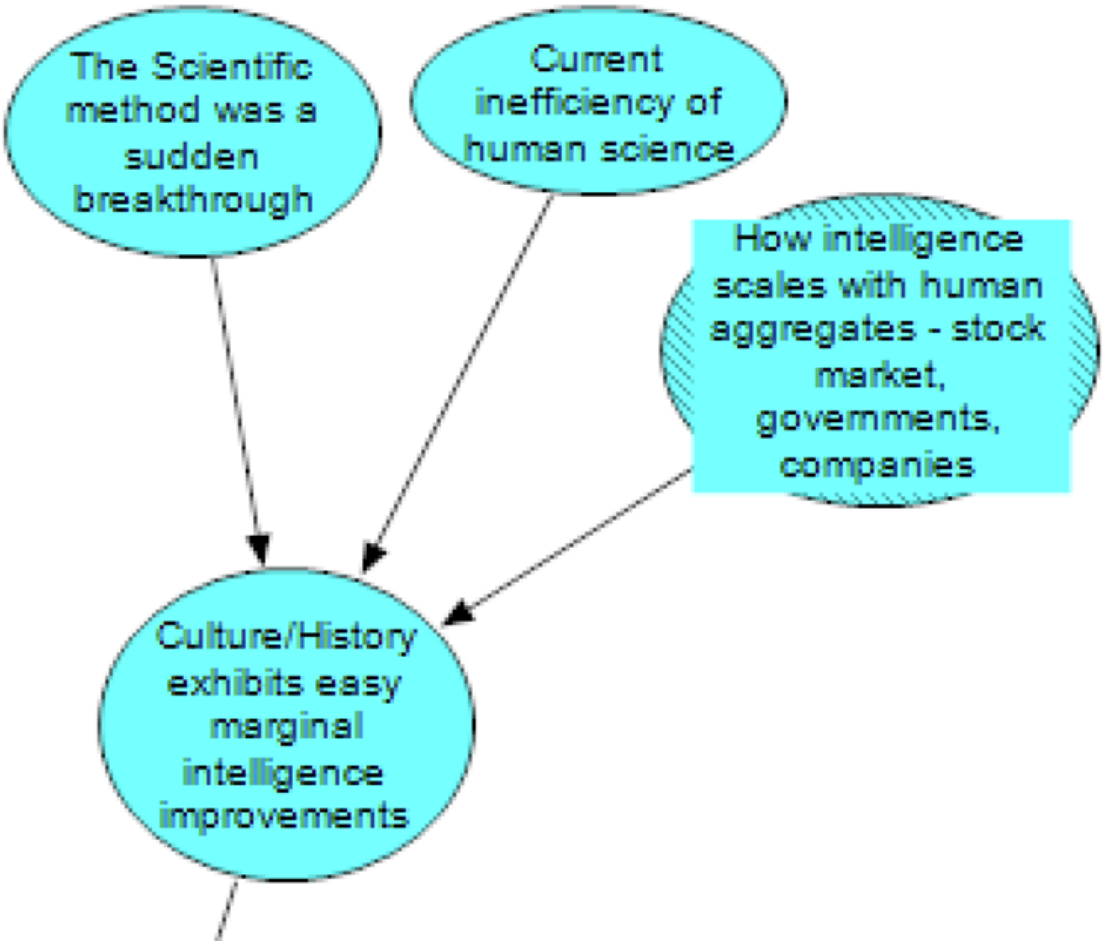
Eliezer Yudkowsky on AlphaGo's Wins

GPT-2 As Step Toward General Intelligence

GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

GPT-3: a disappointing paper

Human Culture and History



Another source of evidence is human history and cultural evolution. During the Hanson-Yudkowsky debate, Eliezer Yudkowsky argued that the scientific method is an organisational and methodological insight that suddenly allowed humans to have much greater control over nature, and that this is evidence that marginal intelligence improvements are easy and that AI systems will similarly have breakthroughs where their capabilities suddenly increase (**left-hand node)**.

In "Intelligence Explosion Microeconomics", Eliezer Yudkowsky also identified specific limitations of human science that wouldn't limit AI systems (**middle node**). For example, human researchers need to communicate using slow and imprecise human language. Wei Dai has similarly argued that AI systems have greater economies of scale than human organizations because they do not hit a Coasean ceiling. We might expect that a lot of human intelligence is 'wasted', as organisations containing humans are not proportionately more intelligent than their members, due to communication limits that won't exist in HLMI (**right-hand node**). If these claims are right, simple organisational insights radically improved humanity's practical abilities, but we still face many organisational limitations an AI would not have to deal with. This suggests marginal improvements in practical abilities could be easy for an AI. On the other hand, even early, relatively unintelligent AIs don't face human limitations such as the inefficiency of interpersonal communication. This means AI might have already "baked in" whatever gains can be achieved from methodological improvements before reaching HLMI.

Sources:

Hanson-Yudkowsky AI Foom Debate (search “you look at human civilization and there's this core trick called science”)

Debating Yudkowsky (point 5 responds to Eliezer)

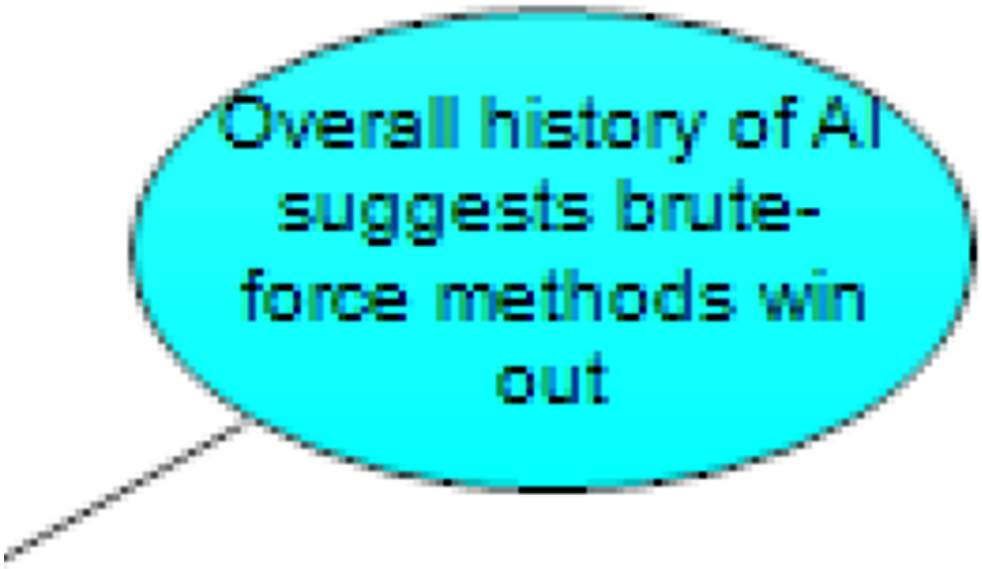
(above links from Science argument)

Intelligence Explosion Microeconomics (3.5)

AGI and Economies of Scale

Continuing the takeoffs debate - LessWrong 2.0 viewer

History of AI



This section covers reference-class forecasting based on the history of AI, going back before the current machine learning paradigm. Principally, the ‘bitter lesson’ (2019):

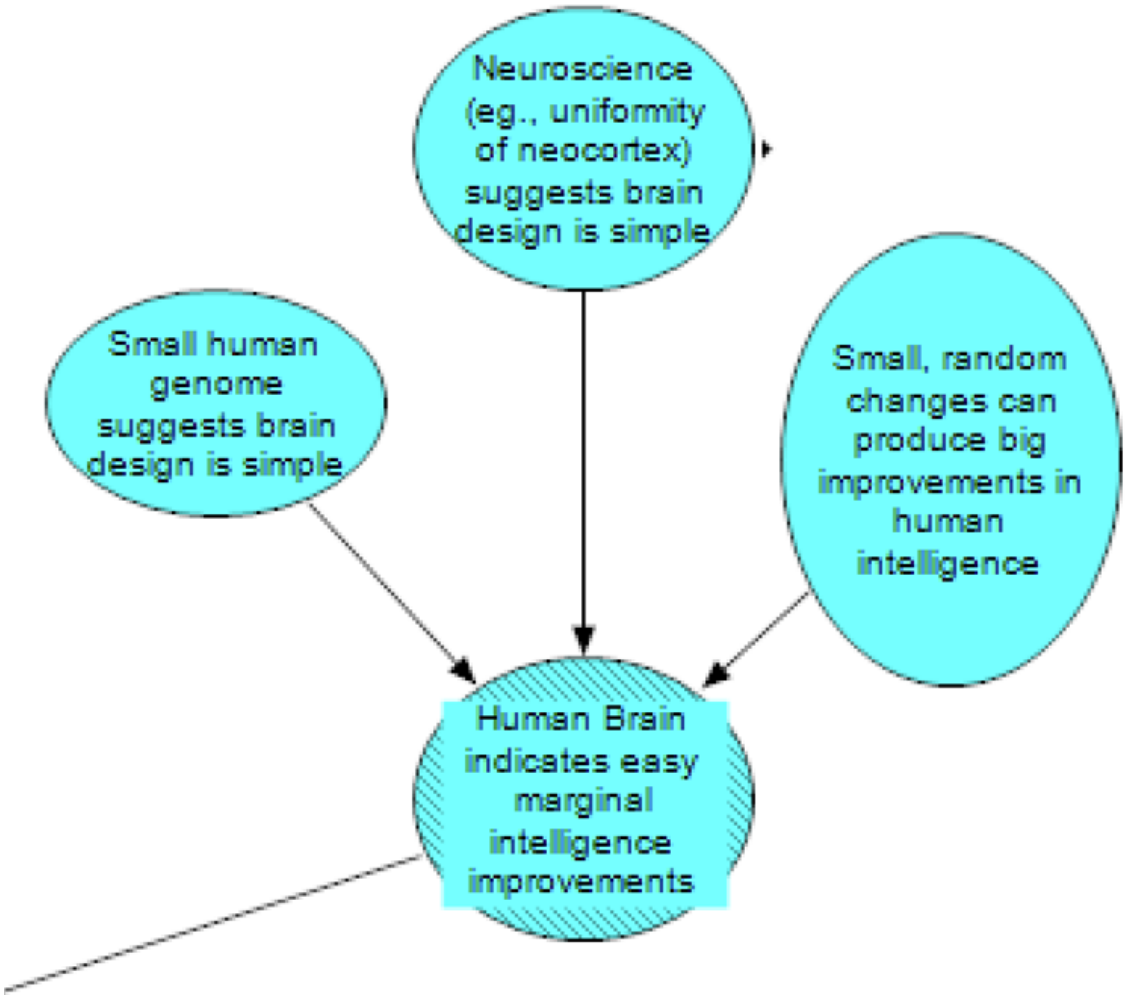
> The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin... Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation.

The bitter lesson is much broader than machine learning and also includes, for example, the success of deep (relatively brute-force) search methods in computer chess. If the bitter lesson is true in general, it would suggest that we can get significant capability gains from scaling up models. That in turn should lead us to update towards marginal intelligence improvements being easier. The conjecture is that we can continually scale up compute and data to get smooth improvements in search and learning. If this is true, then plausibly scaling up compute and data will also produce smooth increases in general intelligence.

Sources:

Bitter lesson

The Human Brain



Biological details of the human brain provide another source of intuitions about the difficulty of marginal improvements in intelligence. Eliezer Yudkowsky has argued that (search for "750 megabytes of DNA" on that page) the small size of the human genome suggests that brain design is simple (this is similar to the Genome Anchor hypothesis in Ajeya Cotra's AI timelines report, where the number of parameters in the machine learning model is anchored to the number of bytes in the human genome) (**left-hand node**).

If the human neocortex is uniform in architecture, or if cortical neuron count strongly predicts general intelligence, this also suggests there is a simple "basic algorithm" for intelligence (possibly analogous to a common algorithm already used in ML). The fact that the neocortex can be divided into different brain regions with different functions, and that the locations of these different regions are conserved across individuals, is evidence against such a simple, uniform algorithm, but on the other hand, the ability of neurons in certain regions to be recruited by other regions (e.g., in ferrets that have had retinal projections redirected to the auditory thalamus, or in blind humans that can learn to echolocate via mouth clicks and apparently using brain regions typically devoted to vision) is an argument in favour. If the human brain employs a simple algorithm, then we should think it more likely that there are other algorithms that can be rapidly scaled to produce highly intelligent behaviour. These all fall under the evidence from neuroscience node (**middle node**).

Variation in scientific and other intellectual ability among humans (compare Von Neumann to an average human) who share the same basic brain design also suggests improvements are easy at the HLMI level. Similarly, the fact that mood or certain drugs (stimulants and psychedelics) can sometimes improve human cognitive performance implies that humans aren't at a relative maximum, as if we were, simple blunt changes to our cognition should almost always degrade performance (and rare reports of people gaining cognitive abilities after brain damage provide a potentially even more extreme version of this argument). All of these provide evidence for the claim that small, random changes can produce big improvements in human intelligence (**right-hand node**).

Sources:

Human genome

Hanson-Yudkowsky Debate

Source code size vs learned model size in ML and in humans?

(the above links are from Missing gear for intelligence and Secret sauce for intelligence)

Investigation into the relationship between neuron count and intelligence across differing cortical architectures

Neurons And Intelligence: A Birdbrained Perspective

Jeff Hawkins on neuromorphic AGI within 20 years

General Priors

One’s beliefs about the possibility of an intelligence explosion are likely influenced to a large degree by general priors about the nature of intelligence: whether “general intelligence” is a coherent concept, whether nonhuman animal species can be compared by level of general intelligence, and so on. Claims like "intelligence isn't one thing that can just be scaled up or down – it consists of a bunch of different modules that are put together in the right way", imply that it is not useful to compare minds on the basis of general intelligence. For instance, François Chollet denies the possibility of an intelligence explosion in part based on these considerations. As well as affecting the difficulty of marginal intelligence improvements, general priors (including the possibilities that *previous intelligence is a bottleneck to future improvements* and that there exists an *upper limit to intelligence*) also matter because they are potential defeaters of a fast progress/intelligence explosion scenario.

Sources:

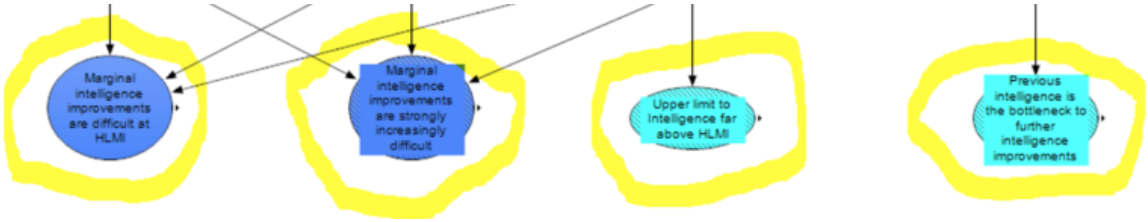
The implausibility of intelligence explosion

A reply to Francois Chollet on intelligence explosion

[General intelligence]

[General Priors]

The Outputs



The empirical cruxes mentioned above influence the outputs of this module, which further influence downstream nodes in other modules. The cruxes this module outputs are: **Marginal Intelligence Improvements are difficult at HLMI**, **Marginal Intelligence Improvements are Strongly Increasingly Difficult**, **the Upper Limit to Intelligence is far above HLMI**, and **Previous intelligence is a bottleneck to future intelligence improvements**. These cruxes influence probabilities the model places on different paths to HLMI and takeoff scenarios.

For the two Difficulty of Marginal Intelligence Improvements nodes, we implement a [naive bayes classifier] in the Analytica model (i.e., a probabilistic classifier that applies Bayes' theorem with strong independence assumptions between the features). Each claim about one of the domains analogous to HLMI (e.g., in the domain of the human brain, the claim that the neocortex is uniform) is more likely to be true in a world where marginal intelligence improvements *in general* are either easy or hard. Therefore, when taken together, the analogy areas provide evidence about the difficulty of marginal intelligence improvements for HLMI.

This use of a naive Bayes classifier enables us to separate out the prior likelihoods of the original propositions, for example that the human neocortex is uniform, and these propositions' relevance to HLMI (likelihood of being true in a world with easy vs hard marginal intelligence improvements).

This extra step of using Bayes classification is useful because the claims about the analogy domains themselves are often much more certain than their degree of relevance to claims about the ease of improvements to HLMI. Whether there was a rapid acceleration in intelligence during hominin evolution is something that can be assessed by domain experts or a review of the relevant literature, but the relevance of this claim to HLMI is a separate question we are much less certain about. Using the naive Bayes classifier allows us to separate these two factors out.

Difficulty of Marginal Intelligence Improvements

How difficult marginal intelligence improvements are at HLMI and how rapidly they become more difficult at higher levels of intelligence are some of the most significant cruxes of disagreement among those trying to predict the future of AI. If marginal intelligence improvements are difficult around HLMI, then HLMI is unlikely to arrive in a sudden burst. If marginal intelligence improvements rapidly become increasingly difficult beyond HLMI, it is unlikely there will be an intelligence explosion post-HLMI.

For an [example] from Yudkowsky of discussion relating to whether marginal intelligence improvements are 'strongly increasingly difficult':

> The Open Problem posed here is the quantitative issue: whether it's possible to get sustained returns on reinvesting cognitive improvements into further improving cognition

('Strongly increasingly difficult' has often been operationalized as 'exponentially increasingly difficult' , as in that way it serves as a defeater for the 'sustained returns' that we might expect from powerful AI accelerating the development of AI)

Paul Christiano has also claimed models of progress which include "key insights [that]... fall into place" are implausible, which we model as a claim that marginal intelligence improvements at HLMI are *difficult* (since if they are difficult, a few key insights will not be enough). For a direct example of a claim that difficulty of marginal intelligence improvements is a key crux, see this quote by Robin Hanson:

> So I suspect this all comes down to, how powerful is architecture in AI, and how many architectural insights can be found how quickly? If there were say a series of twenty deep powerful insights, each of which made a system twice as effective, just enough extra oomph to let the project and system find the next insight, it would add up to a factor of a million. Which would still be nowhere near enough, so imagine a lot more of them, or lots more powerful.

In our research, we have found that what is most important in many models of AI takeoff is their stance on whether marginal intelligence improvements are difficult at HLMI and if they become much more difficult beyond HLMI. Some models, for instance, claim there is a 'secret sauce' for intelligence such that once it is discovered, progress becomes easy. In the **Discontinuities and Takeoff Speeds** section of the model (presented in a subsequent post in this series), we more closely examine the relationships between claims about ease of marginal intelligence improvements and progress around HLMI.

Why did we attempt to identify an underlying key crux in this way? It is clear that beliefs about AI takeoff relate to beliefs from arguments by analogy (in this post) in fundamental ways, for example:

Paul Christiano et al.: Human Evolution **is not** an example of massive capability gain given constant optimization for intelligence + *(other factors)* **➔ [Implicit Belief A about nature of AI Progress] ➔ Continuous** Change model of Intelligence Explosion, likely **not** highly localized

Eliezer Yudkowsky et al.: Human Evolution **is** an example of massive capability gain given constant optimization for intelligence + *(other factors)* **➔ [Implicit Belief B about nature of AI Progress] ➔ Discontinuous** Change model of Intelligence Explosion, likely highly localized

We have treated implicit beliefs A and B as being about the difficulty of marginal intelligence improvements at HLMI and beyond.

We have identified 'Previous intelligence is a bottleneck' and 'There is an upper limit to intelligence' as two other important cruxes. They are commonly cited as defeaters for scenarios that involve any kind of explosive growth due to an acceleration in AI progress. Both of these appear to be cruxes between sceptics and non-sceptics of HLMI and AI takeoff.

Previous Intelligence is a bottleneck

The third major output of this module is whether, in general, further improvements in intelligence tend to be bottlenecked by the current intelligence of our systems rather than some external factor (such as the need to run experiments and wait for real-world data).

This output is later used in the intelligence explosion module (covered in a subsequent post in this series): if such an external bottleneck exists, we are unlikely to see rapid acceleration

of technological progress through powerful AI improving our ability to build yet more powerful AI. There will instead be drag factors preventing successor AIs from being produced without reference to the outside world. This view is summarised by Francois Chollet:

> If intelligence is fundamentally linked to specific sensorimotor modalities, a specific environment, a specific upbringing, and a specific problem to solve, *then you cannot hope to arbitrarily increase the intelligence of an agent merely by tuning its brain — no more than you can increase the throughput of a factory line by speeding up the conveyor belt*. Intelligence expansion can only come from a co-evolution of the mind, its sensorimotor modalities, and its environment.

In short, this position claims that positive feedback loops between successive generations of HLMI could not simply be closed, but instead would require feedback from environmental interaction that can't be arbitrarily sped up. While Chollet appears to be assuming such bottlenecks will occur due to necessary interactions with the physical world, it is possible in principle that such bottlenecks could occur in the digital world – for instance, brain emulations that were sped up by 1Mx would find all computers (and thus communications, access to information, calculators, simulations, etc) slowed down by 1Mx (from their perspective), potentially creating a drag on progress.

A somewhat more tentative version of this claim is that improvements in intelligence (construed here as 'ability to do applied science') require a very diverse range of discrete skills and access to real-world resources. Ben Garfinkel makes this point:

> ...there's really a lot of different tasks that feed into making progress in engineering or areas of applied science. There's not really just this one task of "Do science". Let's take, for example, the production of computing hardware… I assume that there's a huge amount of different works in terms of how you're designing these factories or building them, how you're making decisions about the design of the chips. Lots of things about where you're getting your resources from. Actually physically building the factories.

Upper Limit to Intelligence

Finally, this module has an output for whether there is a practical upper limit to intelligence not significantly above the human level. This could be true if there are physical barriers to the development of greater intelligence. Alternatively, it seems more likely to be effectively true if we cannot even compare the abilities of different minds along a general metric of "intelligence" (hence a "no" answer to "it is possible to compare minds on the generality of their intelligence" leads to a "yes" to this crux). For the "not possible to compare minds on the generality of their intelligence" claim, from François Chollet:

> The first issue I see with the intelligence explosion theory is a failure to recognize that intelligence is necessarily part of a broader system—a vision of intelligence as a "brain in jar" that can be made arbitrarily intelligent independently of its situation.

Chollet argues that human (and all animal) intelligence is 'hyper-specialised' and situational to a degree that makes comparisons of general intelligence much less useful than they first appear,

> If intelligence is a problem-solving algorithm, then it can only be understood with respect to a *specific* problem. In a more concrete way, we can observe this empirically in that all intelligent systems we know are highly specialized ... The brain has hardcoded conceptions of having a body with hands that can grab, a mouth that can suck, eyes mounted on a moving head that can be used to visually follow objects (the vestibulo-ocular reflex), and these preconceptions are required for human intelligence to start taking control of the human body. It has even been convincingly argued, for instance by

Chomsky, that very high-level human cognitive features, such as our ability to develop language, are innate.

A strong version of the modularity of mind hypothesis, ‘massive modularity’, also implies that intelligence is hyper-specialised at extremely specific tasks. If true, massive modularity would lend support to the claim that ‘intelligence… can only be understood with respect to a specific problem’, which in turn suggests that intelligence cannot be increased independent of its situation.

Conclusion

This post has explained the structure and reasoning behind one of the starting points of the MTAIR model - Analogies and General Priors. This module connects conclusions about the nature of HLMI to very basic assumptions about the nature of intelligence, and analogies to domains other than HLMI about which we have greater experience.

We have made the simplifying assumption to group the conclusions drawn from the general priors and the different analogy domains into four outputs, which we think characterise the important variables needed to predict HLMI development and post-HLMI takeoff. In later posts, we will explain how those outputs are used to make predictions about HLMI takeoff and development.

The next post in the series will be **Paths to High-Level Machine Intelligence**, which attempts to forecast when HLMI will be developed and by what route.

We are interested in any feedback you might have, particularly if there are any views or arguments which you feel our model does not currently capture, or captures incorrectly.

Footnotes

1. We define HLMI as machines that are capable of performing almost all economically-relevant information-processing tasks (either individually or collectively). We are using the term “high-level machine intelligence” here instead of the related terms “human-level machine intelligence”, “artificial general intelligence”, or “transformative AI”, since these other terms are often seen as baking in assumptions about either the nature of intelligence or advanced AI that are not universally accepted.

Paths To High-Level Machine Intelligence

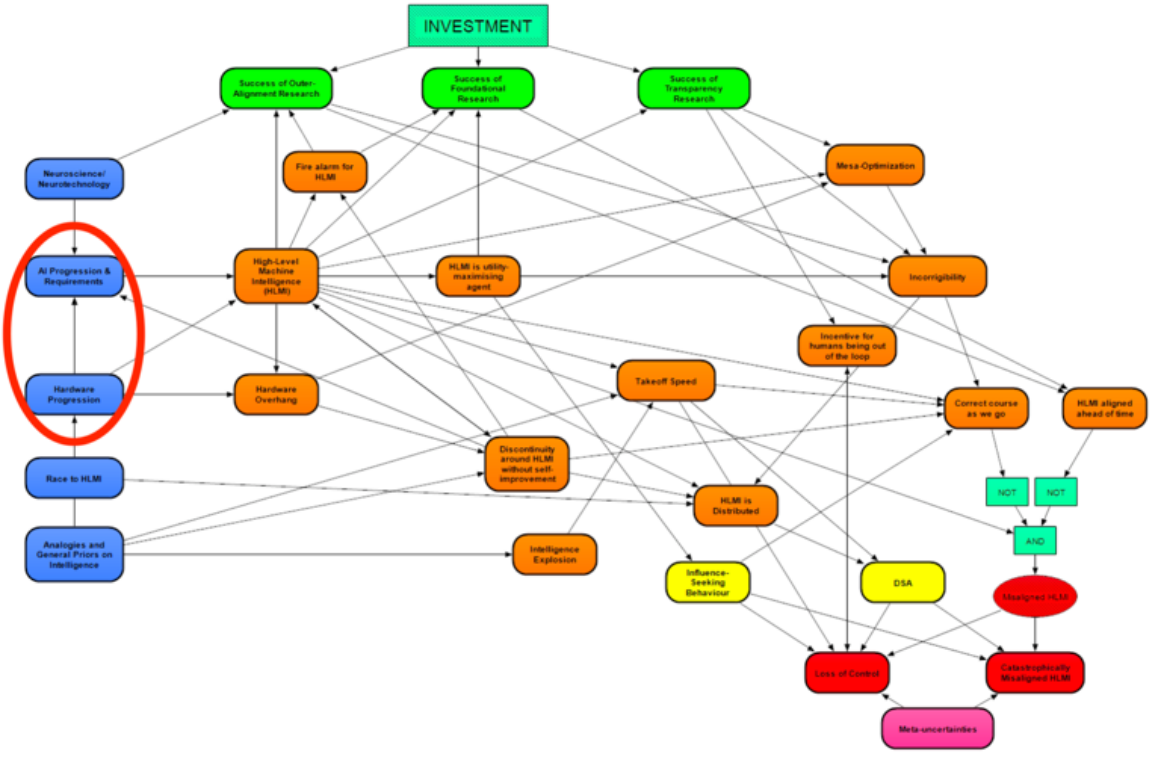
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is part 3 in our sequence on Modeling Transformative AI Risk . We are building a model to understand debates around existential risks from advanced AI. The model is made with Analytica software, and consists of nodes (representing key hypotheses and cruxes) and edges (representing the relationships between these cruxes), with final output corresponding to the likelihood of various potential failure scenarios. You can read more about the motivation for our project and how the model works in the Introduction post . The previous post in the sequence, Analogies and General Priors on Intelligence , investigated the nature of intelligence as it pertains to advanced AI.

This post explains parts of our model most relevant to paths to high-level machine intelligence (HLMI). We define HLMI as machines that are capable, either individually or collectively, of performing almost all economically-relevant information-processing tasks that are performed by humans, or quickly (relative to humans) learning to perform such tasks. Since many corresponding jobs (such as managers, scientists, and startup founders) require navigating the complex and unpredictable worlds of physical and social interactions, the term HLMI implies very broad cognitive capabilities, including an ability to learn and apply domain-specific knowledge and social abilities.

We are using the term “high-level machine intelligence” here instead of the related terms “human-level machine intelligence”, “artificial general intelligence”, or “transformative AI”, since these other terms are often seen as baking in assumptions about either the nature of intelligence or advanced AI that are not universally accepted.

In relation to our model as a whole, this post focuses on these modules:

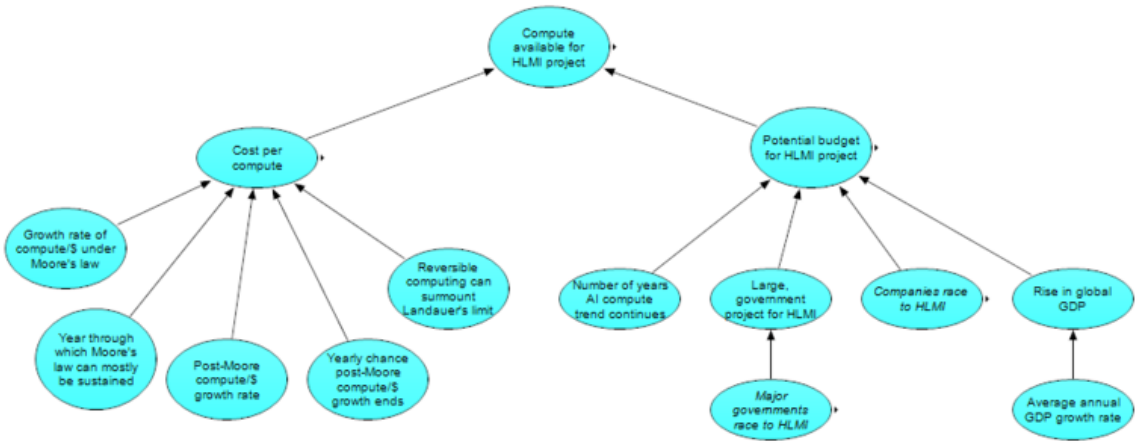


The module *AI Progression & Requirements* investigates when we should expect HLMI to be developed, as well as what kind of HLMI we should expect (e.g., whole brain emulation, HLMI from current deep learning methods, etcetera). These questions about the timing and kind of HLMI are the main outputs from the sections of the model in this post, influencing downstream parts of the model. The timing question, for instance, determines how much time there is for safety agendas to be solved. The question regarding the kind of HLMI, meanwhile, affects many further cruxes, including which safety agendas are likely necessary to solve in order to avoid failure modes, as well as the likelihood of HLMI being distributed versus concentrated.

The module *Hardware Progression* investigates how much hardware will likely be available for a potential project towards HLMI (as a function of time). This module provides input for the *AI Progression & Requirements* module. The *AI Progression & Requirements* module also receives significant input from the *Analogies and General Priors on Intelligence* module, which was described in the previous post in this sequence.

We will start our examination here with the *Hardware Progression* module, and then discuss the module for *AI Progression & Requirements*.

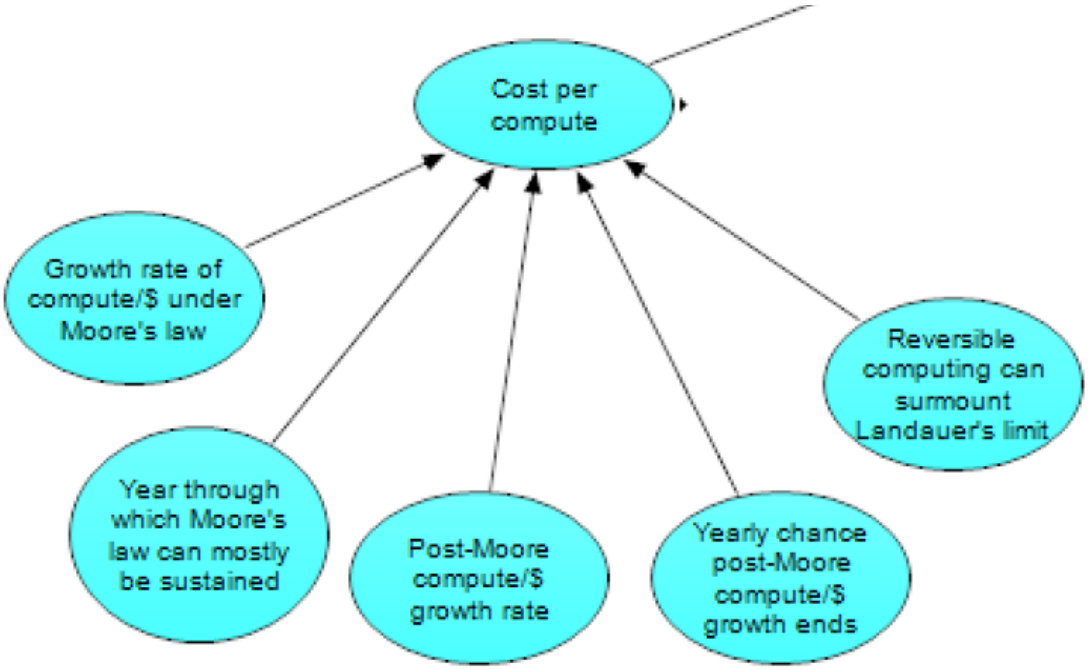
Hardware Progression



The output from this section is *compute available for an HLMI project*, which varies by year. This output is determined by dividing the *potential budget for an HLMI project* by the *cost per compute* (both as a function of the year).

Cost per Compute

Zooming in on the cost portion, we see the following subgraph:



The *cost per compute* (over time) is the output of this subgraph, and is determined by the other listed nodes. Starting from the current cost of compute, the compute/$ is expected to continue rising until the trend of increasing transistor density on 2D Si chips (i.e., “Moore’s law”) runs out of steam. Note that we’re using “Moore’s law” in the colloquial sense to refer to approximately exponential growth within the 2D Si chip paradigm, not the more specific claim of a 1-2 year doubling time for transistor count. Also note that at this stage, we do not differentiate between CPU, GPU, or ASIC compute, as similar trends apply to all of them. Both the *growth rate of compute/$ under the future of Moore’s law* and the *year through which Moore’s law can mostly be sustained* are uncertainties within our model.

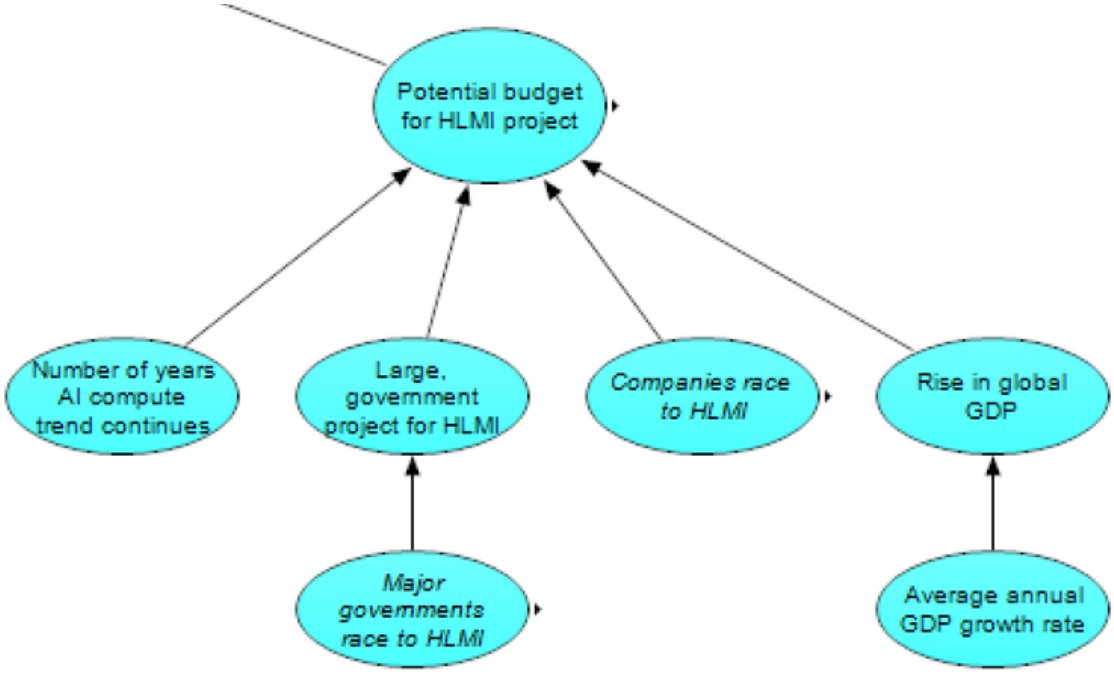
After this paradigm ends, compute/$ may increase over time at a new (uncertain) rate (*post-Moore compute/$ growth rate*). Such a trend could perhaps be sustained by a variety of mechanisms: new hardware paradigm(s) (such as 3D Si chips, optical computing, or spintronics), specialized hardware (leading to more effective compute for the algorithms of interest), a revolution in physical manufacturing (such as via atomically precise manufacturing), or pre-HLMI AI-led hardware improvements. Among technology forecasters, there is large disagreement about the prospects for post-Moore computational growth, with some forecasters predicting Moore-like or faster growth in compute/$ to continue post-Moore, while others expect compute/$ to slow considerably or plateau.

Even if post-Moore growth is initially substantial, however, we would expect compute/$ to eventually run into some limit and plateau (or slow to a crawl), due to either hard technological limits, or economic limits related to increasing R&D and fab costs. The possibility of an eventual leveling off of compute/$ is handled by the model in two different ways. First, there is assumed to be an uncertain, *yearly chance post-Moore compute/$ growth ends*. Second, Landauer’s limit (the thermodynamic limit relating bit erasure and energy use) may present an upper bound for compute/$, which the model assumes will happen unless *reversible computing can surmount Landauer’s limit*.

It should be noted that the model does not consider that specialized hardware may present differential effects for different paths to HLMI, nor does it consider how quantum computing might affect progression towards HLMI, nor the possibility of different paradigms post-Moore seeing different compute/$ growth rates. Such effects may be important, but appear complex and difficult to model.

Potential Budget for HLMI Project

The budget section, meanwhile, has this subgraph:

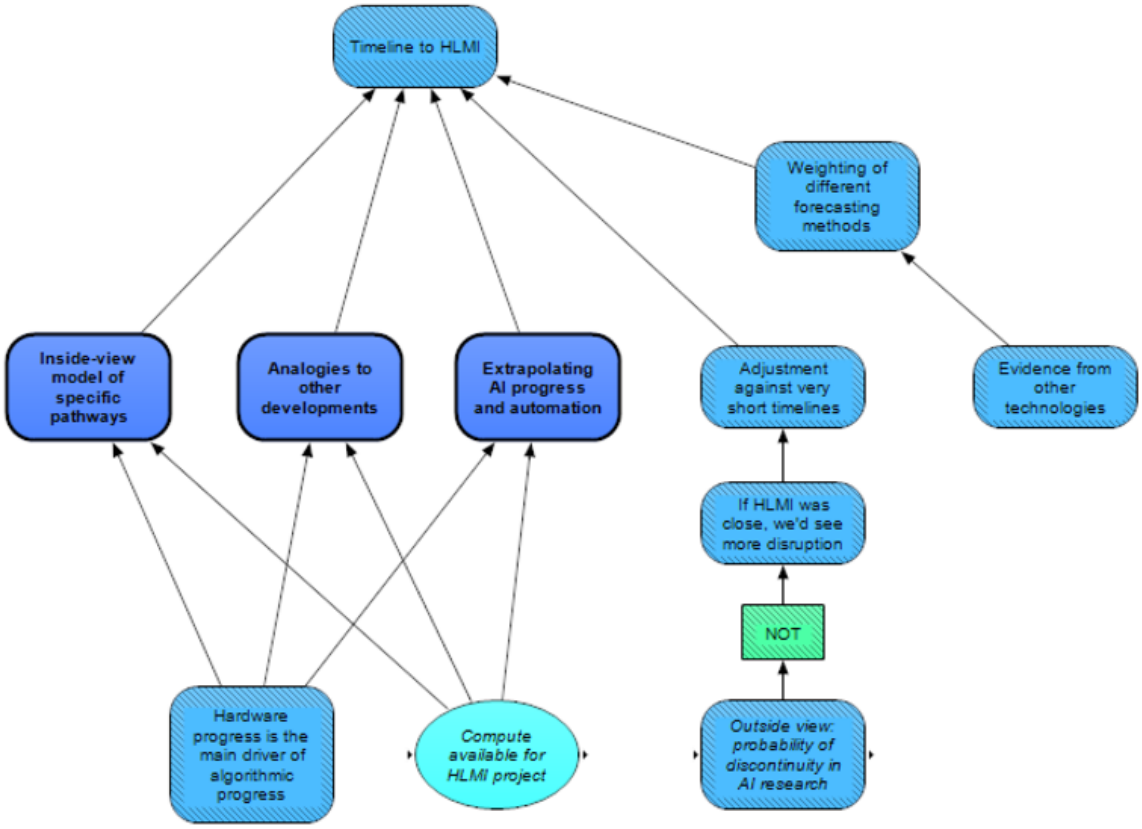


The *potential budget for an HLMI project*, which is the primary output of this subgraph (used in combination with the output from the previous section on *cost per compute* to derive the compute available for an HLMI project) is modeled as follows. In 2021, the budget is set to that of the most expensive AI project to date. The recent, quickly-rising *AI compute trend may or may not continue for some number of years*, with the modeled budget rising accordingly. After that, the budget for HLMI is currently modeled as generally rising with the *rise in global GDP* (determined by the *average annual GDP growth rate*), except if *companies race to HLMI* (in which case we assume the budget will grow to be similar, in proportion of global GDP, to tech firms' R&D budgets), or if there's a *large government project for HLMI* (in which case we assume budgets will grow to ~1% of the richest country's GDP, in line with the Apollo program and Manhattan Project). We think such a large government project is particularly likely if *major governments race to HLMI*.

We should note that our model does not consider the possibility of budgets being cut between now and the realization of HLMI; while we think such a situation is plausible (especially if there are further AI Winters), we expect that such budgets would recover before reaching HLMI, and thus aren't clearly relevant for the potential budget leading up to HLMI. The possibility of future AI Winters implying longer timelines through routes other than hardware (e.g., if current methods "don't scale" to HLMI and further paradigm shifts don't occur for a long time) is, at least implicitly, covered in the module on *AI Progression & Requirements*.

As our model is acyclic, it doesn't handle feedback loops well. We think it is worth considering, among other possible feedback loops, how pre-HLMI AI may affect GDP, or how hardware spending and costs may affect each other (such as through economies of scale & learning effects, as well as simple supply and demand), though these effects aren't currently captured in our model.

AI Progression & Requirements



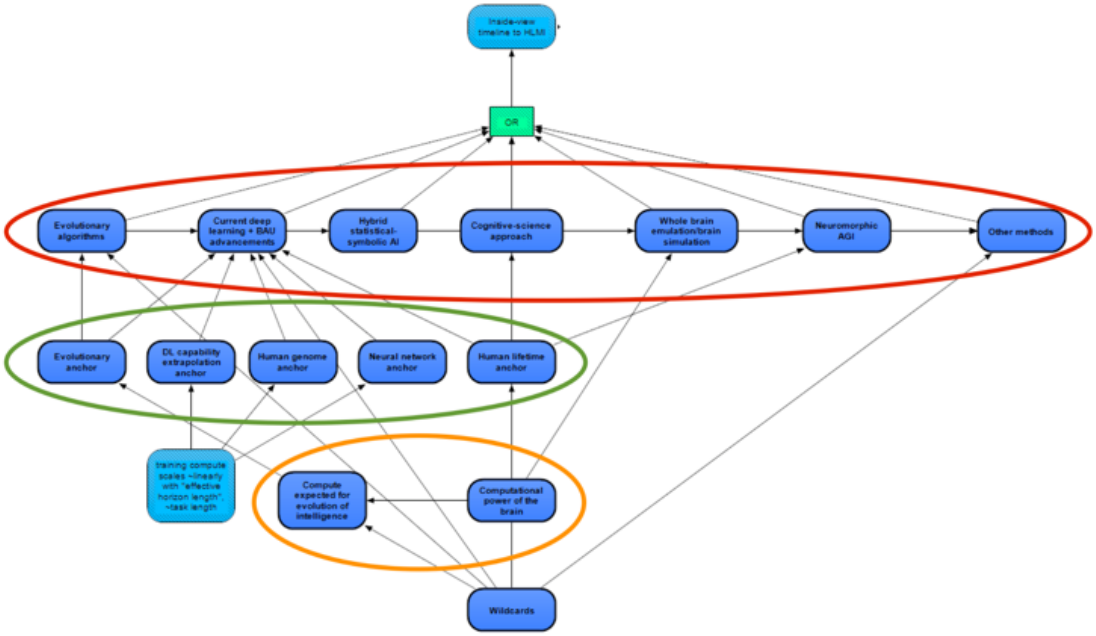
The module on *AI Progression & Requirements* uses a few different methods for estimating the *timeline to HLMI*: a gears-level, *inside-view model of specific pathways* (which considers how various possible pathways to HLMI might succeed), as well as outside-view methods (*analogies to other developments*, and *extrapolating AI progress and automation*). All of these methods are influenced by the *compute available for an HLMI project* (described in the section above), as well as whether or not *hardware progress is the main driver of algorithmic progress*.

Estimates from these approaches are then combined, currently as a linear combination (*weighting of different methods of forecasting timelines*), with the weighting based on *evidence from other technologies* regarding which technological forecasting methods are most appropriate. A (possible) *adjustment against very short timelines* is also made, depending on whether one believes that *if HLMI was close, we'd see more disruption* from AI, which in turn is assumed to be less likely if there's a larger *probability of discontinuity in AI research* (since such a discontinuity could allow HLMI to "come out of nowhere").

Inside-view model of specific pathways

This module represents the most intricate of our methods for estimating the arrival of HLMI. Here, several possible approaches to HLMI are modeled separately, and the date for the arrival of HLMI is taken as the earliest date that one of these approaches succeeds. Note that such an approach may introduce an optimizer's curse (from the OR gate); if the different pathways have independent errors, then the earliest modeled pathway may be expected to have unusually large error in favor of being early, underestimating the actual timeline. On

the other hand, the estimations of the different pathways themselves each contain the combination of different requirements (AND gates), such as requirements for both adequate software and adequate hardware, and this dynamic introduces the opposite bias – the last requirement to fall into place may be expected to have an error in favor of being late. Instead of correcting for these two biases, we simply note that these biases exist in opposite directions, and we are uncertain about which is the larger bias.



The different methods to HLMI in our model (circled in red above, expanded upon in their individual subsections below) are:

- *Evolutionary algorithms – either similar to evolutionary algorithms today, a direct simulation of virtual evolution, or some middle ground. While current evolutionary algorithms are very different from virtual evolution, we have grouped these together instead of grouping current evolutionary algorithms with current deep learning methods, largely because in our model, both of the evolutionary methods rely on similar hardware estimation methods.*
- *Current deep learning plus business-as-usual advancements – HLMI achieved through methods similar to those in deep learning today; the key features of such methods are large, deep neural nets, trained from a largely "blank-slate" initial state, and optimized via local search techniques such as SGD (though crucially, to avoid double counting, not including evolutionary algorithms).*
- *Hybrid statistical-symbolic AI – approaches marrying statistical methods and more intentionally designed symbolic methods, such as in GOFAI .*
- *Cognitive-science approach – approaches to HLMI heavily relying on cognitive science or developmental psychology (likely in combination with deep learning).*
- *Whole brain emulation / brain simulation – emulating a particular person's brain in silico (for WBE), or simulating a generic human brain (for brain simulation).*
- *Neuromorphic AGI – HLMI created using many of the "low-level" processes or architectures of the brain, but without being put together into a virtual brain with particularly humanlike intelligence.*
- *Other methods – a catch-all, for unanticipated or other potential paths to HLMI.*

Each method is generally assumed to be achieved once both the hardware and software requirements are met. The software requirements of the different methods are largely dependent on different considerations, while the hardware requirements are more

dependent on different weighting or modifications of the same set of “anchors” (circled in green in the figure above).

These anchors are plausibly-reasonable estimates for the computational difficulty of training an HLMI, based on different analogies or other reasoning that allows for anchoring our estimates. Four of the five anchors come from Ajeya Cotra’s Draft report on AI timelines (with minor modifications), and we add one more. These five anchors (elaborated upon below) are:

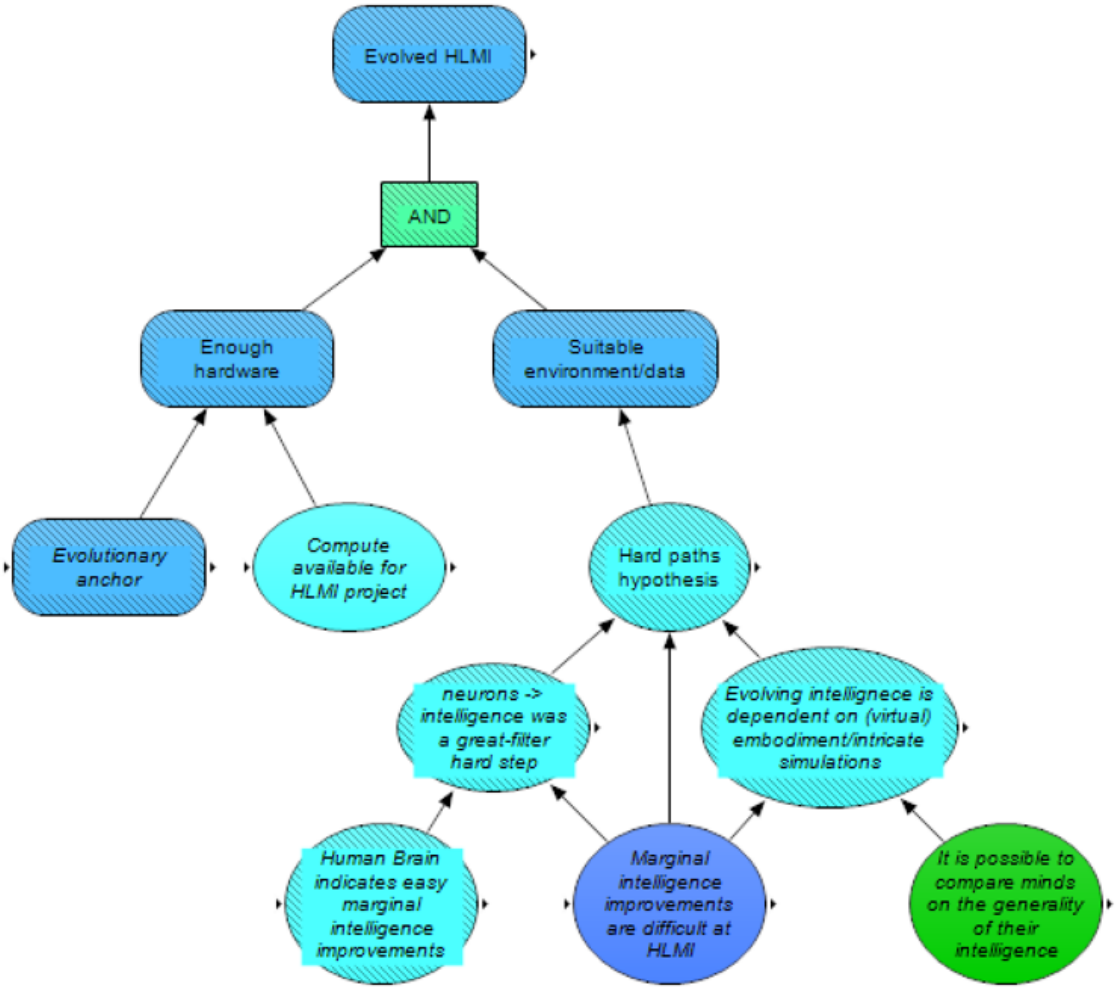
- *Evolutionary anchor – an estimate based on the “compute-equivalent” that was “used” in evolution from the first animals with nervous systems to humans (considering both the “brain-compute” used and the compute necessary to simulate the environment sufficiently). This estimate includes both upward and downward adjustments, for possible anthropic effects and for human engineers potentially outcompeting evolution.*
- *Deep learning capability extrapolation anchor – an estimate of how much compute would be needed to achieve broad, human-level capabilities, given an extrapolation of how capabilities in current deep learning systems scale with compute.*
- *Human genome anchor – an estimate of the compute needed to train a current ML system sufficiently, based on setting the parameter count to the size of the human genome (arguably the “code” of human intelligence), used in combination with empirically-derived scaling laws and other relevant biological considerations.*
- *Neural network anchor – an estimate of the compute needed to train a current ML system sufficiently, similar to the human genome anchor, but instead with the parameter count set by considerations related to the compute-equivalent used in the human brain.*
- *Human lifetime anchor – an estimate of how much compute-equivalent occurs in “training” a human brain from birth to adulthood, increased by a factor for human infants being “pre-trained” by evolution.*

Also similar to Cotra, we plan to use lognormal distributions (instead of point estimates) around our anchors for modeling compute requirements, due to uncertainty, spanning orders of magnitude, about these anchors’ “true” values, as well as uncertainty around the anchors’ applicability without modifications. A couple of these anchors are influenced by upstream nodes related to the evolution of intelligence and the computational power of the brain (circled in orange above; these “upstream” nodes are lower down on the figure since the information in the figure flows bottom to top).

Additionally, one final module includes “wildcards” that are particularly speculative but, if true, would be game-changers. The nodes within this wildcards module have been copied over to other relevant modules (to be discussed below) as alias nodes.

We discuss the modules for the various methods to HLMI below, elaborating on the relevant anchors (and more upstream nodes) where relevant.

Evolutionary Algorithms



The development of evolved HLMI would require both *enough hardware* and a *suitable environment or other dataset*, given the evolutionary algorithm. Whether there will be enough hardware by a specific date depends on both the required hardware (approximated here by the *evolutionary anchor*, discussed below) and how much hardware is available at that date (calculated in the *Hardware Progression* module, discussed above). The ability to create a suitable environment/dataset by a certain date depends, in large part, on the *hard paths hypothesis* – the hypothesis that it's rare for environments to straightforwardly select for general intelligence.

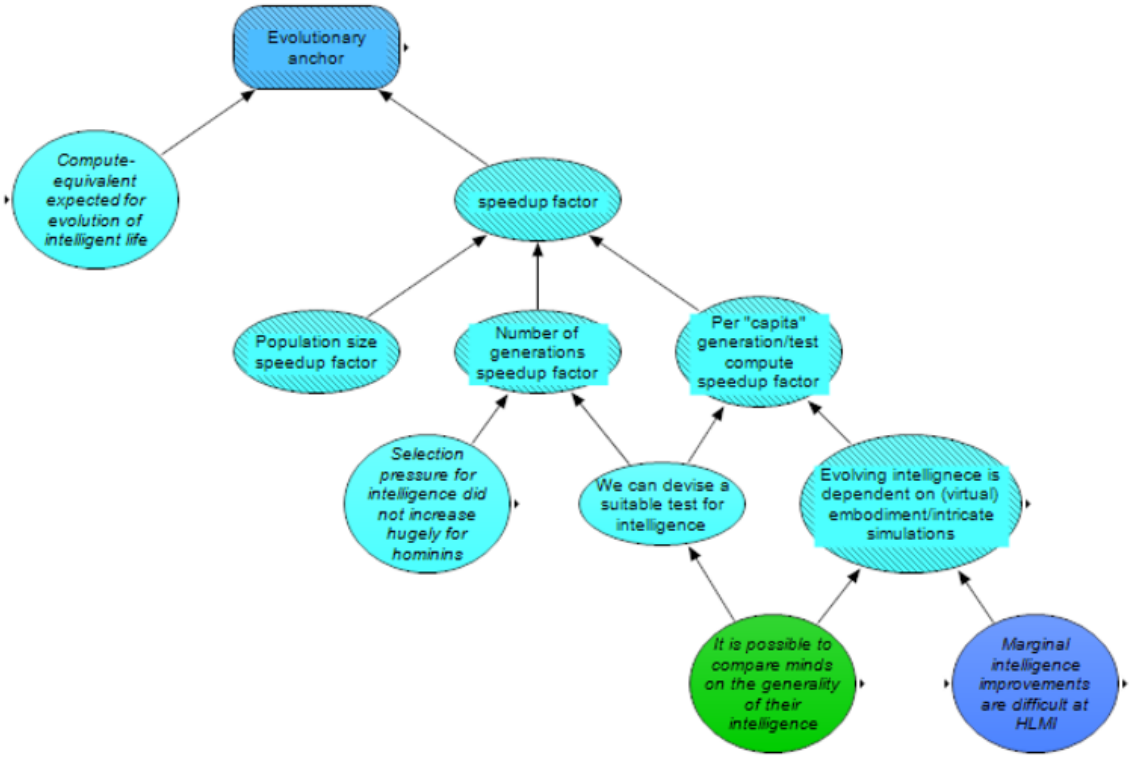
This hypothesis is further influenced by other cruxes. If somewhere along the progression from the first neurons to human intelligence, there was a *"hard step" (a la the great filter)*, then such a hypothesis is more likely true. Such a hard step would imply that the vast majority of planets that evolve animal-like life with neuron-like parts never go on to evolve a technologically intelligent lifeform. This scenario could be the case if Earth (or some portion of Earth's history) was particularly unusual in some environmental factors that selected for intelligence. Due to anthropic effects, we cannot rule out such a hard step simply based on the existence of (human) technological intelligence on Earth. Such a hard step is less likely to be the case if *marginal intelligence improvements are easy around the human level*, and we consider *evidence from the human brain* (such as whether the human brain appears to be a "scaled up" version of a more generic mammalian or primate brain) to be particularly informative here (both of these cruxes are copied over from our module on *Analogies and General Priors on Intelligence*, described in the previous post).

Additionally, the hard paths hypothesis is influenced by whether *evolving intelligence would depend on embodiment* (e.g., in a virtual body and environment). Proponents of this idea have argued that gaining sufficient understanding of the physical world (and/or developing cognitive modules for such understanding), including adequate symbol grounding of important aspects of the world and (possibly necessary) intuitive physical reasoning (e.g., that humans use for certain mathematical and engineering insights), would require an agent to be situated in the world (or in a sufficiently realistic simulation), such that the agent can interact with the world (or simulation). Opponents of such claims often argue that multi-modal learning will lead to such capabilities, without the need for embodiment. Some proponents further claim that even if embodiment is not in principle necessary for evolving intelligence, it may still be necessary in practice, if other environments tend to lack the complexity or other attributes that sufficiently select for intelligence.

If evolving intelligence depends on embodiment, this would significantly decrease the potential paths to evolving intelligence, as HLMI would therefore only be able to be evolved in environments in which it was embodied, and also potentially only in select simulated environments that possessed key features. The necessity of embodiment appears to be less likely if *marginal intelligence improvements are easy*. In particular, the necessity of embodiment appears less likely if *minds can be compared on the generality of their intelligence* (also copied over from our module on *Analogies and General Priors on Intelligence*), as the alternative implies "general intelligence" isn't a coherent concept, and intelligence is then likely best thought of as existing only with respect to specific tasks or goals; in such a situation, AI systems interacting in or reasoning about the physical world would likely need to have the development of their cognitive faculties strongly guided by interactions in such a world (or a close replica).

Finally, even if there is no evolutionary hard step between neurons and intelligence and embodiment is not necessary for evolving intelligence, the hard paths hypothesis may still be true; some have argued, for example, that curriculum learning may be necessary for achieving HLMI. It therefore seems reasonable to assume that the harder marginal intelligence improvements are around the HLMI level (even holding constant the other mentioned cruxes), the lower chance we should give to an individual environment/dataset leading to the evolution of intelligence, and thus the higher credence we give the hard paths hypothesis.

As mentioned above, on the compute side, the requirements for evolved HLMI are dependent on the *evolutionary anchor*. This anchor is determined by an upstream submodule (the *Evolutionary anchor* submodule). Our estimate for this anchor is similar to its analog in Cotra's report, but with a couple of key differences:



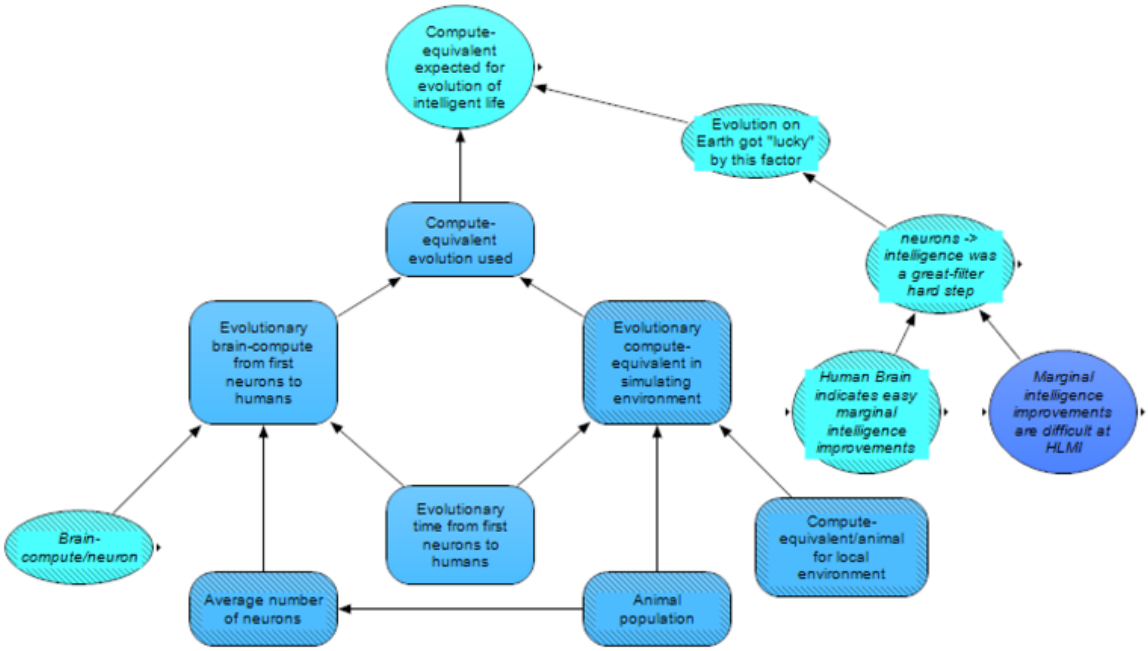
Here, this anchor is determined by taking the amount of "compute-equivalent" *expected for the evolution of intelligent life* such as humans (considering both neural activity and simulating an environment), and dividing it by a *speedup factor* due to human engineers potentially being able to outcompete biological evolution towards this goal. Cotra's report did not include such a speedup factor, as she was using her evolutionary anchor as somewhat of an upper bound. Here, on the other hand, we are instead concerned with an all-things-considered estimate of how difficult it would be to evolve HLMI.

Our *speedup factor* considers three sources of increased speed:

- *Population size speedup factor – the population of Earth's animals is not set by minimizing the amount of compute-equivalent necessary to evolve an intelligent species, but instead by Earth's carrying capacity; human engineers, meanwhile, could set population sizes with the goal of minimizing necessary compute to evolve intelligence.*
- *Number of generations speedup factor – much of Earth's macroevolutionary history was presumably not optimizing for an intelligent species in as few generations as possible, and there are many tricks that human engineers could employ to select for and possibly achieve this goal quicker. In particular, if evolution was hardly selecting for intelligence before hominins, then a more intentional effort could likely be much faster. Additionally, if we can design tests for intelligence, then we could potentially create stronger selection pressure for intelligence, reducing the simulated time needed.*
- *Per "capita" speedup factor – for biological evolution, organisms are "run" from birth until they die, despite much of that time not necessarily presenting much selection effect. Human engineers might be more efficient here – especially if we can devise a test for intelligence (in which case the generation time might be whatever is required to complete a short test) and/or if evolving intelligence does not depend on embodiment (in which case much of the information processes in animals' brains is likely, in principle, unimportant for evolving intelligence, and thus human engineers could reduce compute requirements by crafting much simpler datasets that don't lead to costly adaptations for embodiment).*

Note that there is somewhat of a tradeoff between these three factors – a smaller population, for instance, would presumably imply more generations needed. Each of these factors should therefore be set to minimize the product of the three variables instead of only the corresponding variable itself.

The expected amount of compute-equivalent required to evolve intelligence on Earth (not considering the *speedup factor* – i.e., the left node in the above diagram) is estimated from a further upstream submodule, *compute expected for evolution of intelligence*, which is represented below.



This estimate is made up of two main factors which are multiplied together: the *compute-equivalent evolution used* on Earth, and a factor related to the possibility that *evolution on Earth got "lucky"*. The latter of these factors corresponds to the aforementioned possibility that there was a hard step between the first neurons and humans – this factor is set to 1 if either there was no hard step or if there was a hard step relating strictly to Earth's environment being unusual, and otherwise the factor is set to a number higher than one, corresponding to how much faster evolution of intelligence (from first neurons) took compared to, naively, how long such a process would be expected to take given an Earth-like environment and infinite time. That is, the factor is more-or-less set to an estimate to the question, "For every Earth-like planet that evolves a lifeform with technological intelligence, how many Earth-like planets evolve a lifeform with something like Precambrian jellyfish intelligence?"

The factor for *compute-equivalent evolution used*, meanwhile, is broken up into the sum of two factors: the *brain-compute from first neurons to humans* (i.e., based on an estimate of how much compute each nervous system "performs", though note the definition here is somewhat inexact) and the *compute-equivalent used in simulating the environment* (i.e., an estimate of the minimum compute necessary to simulate an environment for the evolution of animals on Earth). The environmental compute-equivalent factor has traditionally been neglected in analyzing the compute-equivalent for simulating evolution, as previous research has assumed that much more necessary compute occurs in brains than in the surrounding environment (for instance, because the human brain uses much more brain-compute than the compute in even high-resolution video games, and humans can function in such environments, including interacting with each other). While this assumption may be reasonable for the functioning of large animals such as humans, it is not immediately

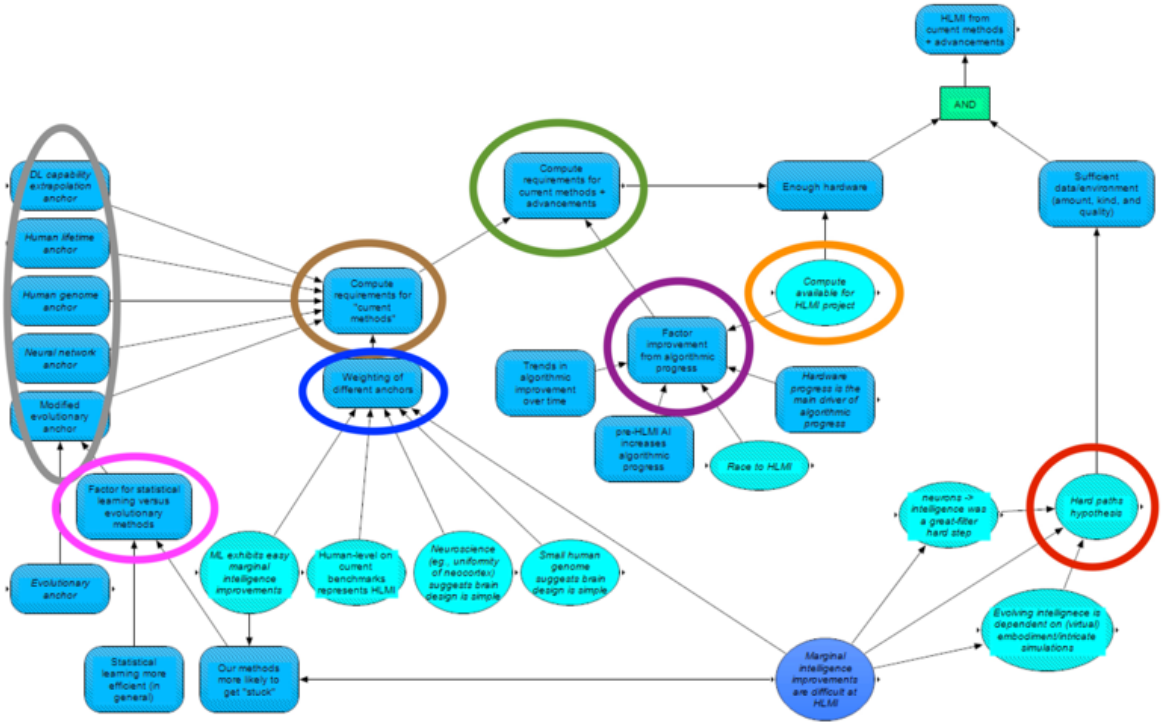
obvious that it also applies to much smaller animals such as *C. elegans*. Furthermore, environmental complexity, dynamism, and a large action-space are typically thought to be important for the evolution of intelligence, and a more complex and dynamic environment, especially in which actors are free to take a very broad set of actions, may be more computationally expensive to simulate.

Both the brain-compute estimate and the environmental compute-equivalent estimate are determined by multiplying the *evolutionary time from first neurons to humans* by their respective populations, and then by the compute-equivalent per time of each member within these populations. For the brain-compute estimate, the population of interest is the neurons, so the relevant factors are *average number of neurons* (in the world) and the *brain-compute per neuron* (as has been estimated by Joseph Carlsmith). For the environmental compute-equivalent estimate, the relevant factors are the *average population of animals* (of course, related to the average population of neurons), and the *compute required to simulate the local environment* for the average animal (under the idea that the local environment surrounding animals may have to be simulated in substantially higher resolution than parts of the global environment which no animal occupies).

To recap, adding together the brain-compute estimate and the environmental compute-equivalent estimate yields the total estimate for compute-equivalent evolution used on Earth, and this estimate is then multiplied by the “luckiness” factor for the compute-equivalent expected to be necessary for evolving intelligence. The evolutionary anchor is then determined by multiplying this factor by a speedup factor. The date by which enough hardware is available for evolved HLMI is set to be the first date the available compute for an HLMI project exceeds the evolutionary anchor, and evolved HLMI is assumed to be created on the first date that there is enough hardware and a suitable environment or dataset (which is strongly influenced by the hard paths hypothesis).

Current Deep Learning plus Business-As-Usual Advancements

This module, which represents approaches towards HLMI similar to most current approaches used in deep learning, is more intricate than the others, though also relies on some similar cruxes to the above.



Similar to evolutionary algorithms, we model achieving this milestone at the first date that there is both *enough hardware* (given the available algorithms) and *sufficient data* (quantity, quality, and kind, again given the available algorithms) for training the HLMI.

The ease of achieving the necessary data is strongly influenced by the *hard paths hypothesis* (circled in red above), as before.

Whether enough hardware will exist for HLMI via these methods (by a certain date) is determined by both the *amount of hardware available* (circled in orange) and the *hardware requirements* (circled in green). As before, the hardware available is taken from the *Hardware Progression* module. The hardware requirements, in turn, are determined by the *compute requirements for "current methods"* in deep learning to reach HLMI (circled in brown), modified by a *factor for algorithmic progress* (circled in purple). Expected algorithmic improvements are modeled as follows.

First, baseline algorithmic improvements are extrapolated from *trends in such improvements*. This extrapolation is performed, by default, in terms of time. However, if the crux *hardware progress is the main driver of algorithmic progress* resolves positively, then the extrapolation is instead based on hardware progress, proxied with *compute available for an HLMI project*. Next, the baseline algorithmic improvements are increased if there is a *race to HLMI* (in which case it is assumed investment into algorithmic advancements will increase), or if it is assumed that *pre-HLMI AI will enable faster AI algorithmic progress*.

Compute requirements for "current methods" are estimated as a linear combination of the estimates provided by the anchors (circled in grey). The *evolutionary anchor* is *modified by a factor* (circled in pink) for whether our statistical learning algorithms would be better/worse than the evolutionary anchor in terms of finding intelligence. The main argument in favor of our statistical methods is that *statistical learning is generally more efficient than evolution*, and the main argument against is that our *statistical methods may be more likely to get "stuck"*, either via strongly separated local minima, or due to goal hacking and Goodhart's Law becoming dead-ends for sufficiently general intelligence (which may be less likely if *marginal improvements in intelligence* near HLMI are easier, and particularly less likely if *ML shows evidence of easy marginal intelligence gains*).

Our estimates for the human lifetime anchor, human genome anchor, and neural network anchor are all calculated using the similar logic as in Cotra's report. The human lifetime anchor involves estimating the brain-compute that a human brain performs in "training" from birth to adulthood – i.e., (86 billion neurons in a human brain)*(brain-compute per neuron per year)*(18 years) – and multiplying this number by a factor for newborns being "pre-trained" by evolution. This pre-training factor could be based on the extent to which ML systems today are "slower-learners" compared to humans, or the extent to which human-engineered artifacts tend to be less efficient than natural analogs.

The human genome anchor and neural network anchor, meanwhile, are calculated in somewhat similar manners to each other. In both, the modeled amount of compute to train an HLMI can be broken down into the amount of data to train the system, times the compute used in training per amount of data. The number of data points needed can be estimated from the number of parameters the AI uses, via empirically-derived scaling laws between parameters and data points for training, with the number of parameters calculated differently for the different anchors: for the human genome anchor, it's set to the bytes in the human genome, and for the neural network anchor, it's set based on expectations from scaling up current neural networks to use similar compute as the brain (with a modifying factor). The compute used in training per data point for both of these anchors, meanwhile, is calculated as the brain-compute in the human brain, time a modifying factor for the relative efficiency of human engineered artefacts versus their natural analogs, times the amount of "subjective time" generally necessary to determine if a model perturbation improves or worsens performance (where "subjective time" is set from the amount of compute the AI uses, such that the compute the AI uses equals the brain-compute the human brain uses, times the modifying factor, times the subjective time). For this last factor regarding the subjective time for determining if a model perturbation is beneficial or harmful (called the "effective horizon length"), for the human genome anchor, the appropriate value is arguably related to animal generation times, while for the neural network anchor, the appropriate value is arguably more related to the amount of time that it takes humans to receive environmental feedback for their actions (i.e., perhaps seconds to years, though for meta-learning or other abilities evolution selected for, effective horizon lengths on par with human generation times may be more appropriate). For a more in-depth explanation of the calculations related to these anchors, see Cotra's report or Rohin Shah's summary of the report.

Finally, our model includes an anchor for *extrapolating the cutting edge of current deep learning algorithms* (e.g., GPT-3) along various benchmarks in terms of compute – effectively, this anchor assumes that deep learning will (at least with the right data) "spontaneously" overcome its current hurdles (such as building causal world models, understanding compositionality, and performing abstract, symbolic reasoning) in such a way that progress along these benchmarks will continue their current trends in capabilities vs compute as these hurdles are overcome.

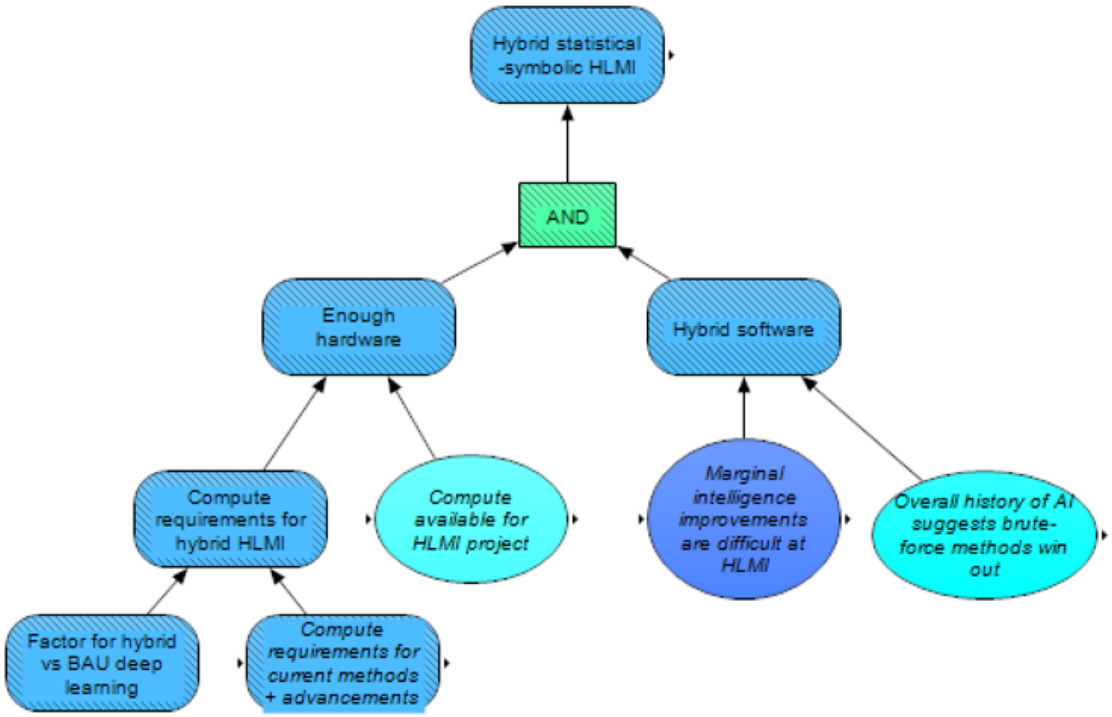
The human genome anchor, neural network anchor, and DL capability extrapolation anchor all rely on the concept of the effective horizon length, which potentially provides a relationship between task lengths and training compute. Resultantly, disagreements related to these relationships are cruxy. One view is that, all else equal, effective horizon length is generally similar to task length, and training compute scales linearly with effective horizon length. If this is not the case, however, and either effective horizon length grows sublinearly with task length, or training scales sublinearly with effective horizon length, then all three of these anchors could be substantially shorter. While our model currently has a node for whether *training time scales approximately linearly with effective horizon length and task length*, we are currently uncertain how to model these relationships if this crux resolves negative. An intuition pump in favor of sublinearity between training time and task length is that if one has developed the qualitative capabilities necessary to write a 10-page book, one also likely has the qualitative capabilities necessary to write a 1,000-page book; writing a 1,000-page book might itself take 100 times as long as writing a 10-page book, but training

to write a 1,000-page book presumably does not require 100 times longer than training to write a 10-page book.

The *weighting* of these five anchors (circled in blue) is also a subject of dispute. We model greater weighting towards lower anchors if *marginal intelligence improvements around HLMI are easy*, and we additionally weight several of the specific anchors higher or lower depending on other cruxes. Most importantly, the weighting of the *DL capability extrapolation anchor* is strongly dependent on whether *human-level on current benchmarks represents HLMI* and if *ML exhibits easy marginal intelligence improvements* (implying such extrapolations are more likely to hold until humal level, as compute is ramped up). The *human lifetime anchor* is weighted higher if *neuroscience suggests the brain design is simple* (since then, presumably, the human mind is not heavily fine-tuned due to pretraining). Additionally, if the *small human genome suggests that the brain design is simple*, then the *human genome anchor* is weighted higher (otherwise, we may assume that the complexity of the brain is higher than we might assume from the size of the genome, in which case this anchor doesn't make a lot of sense).

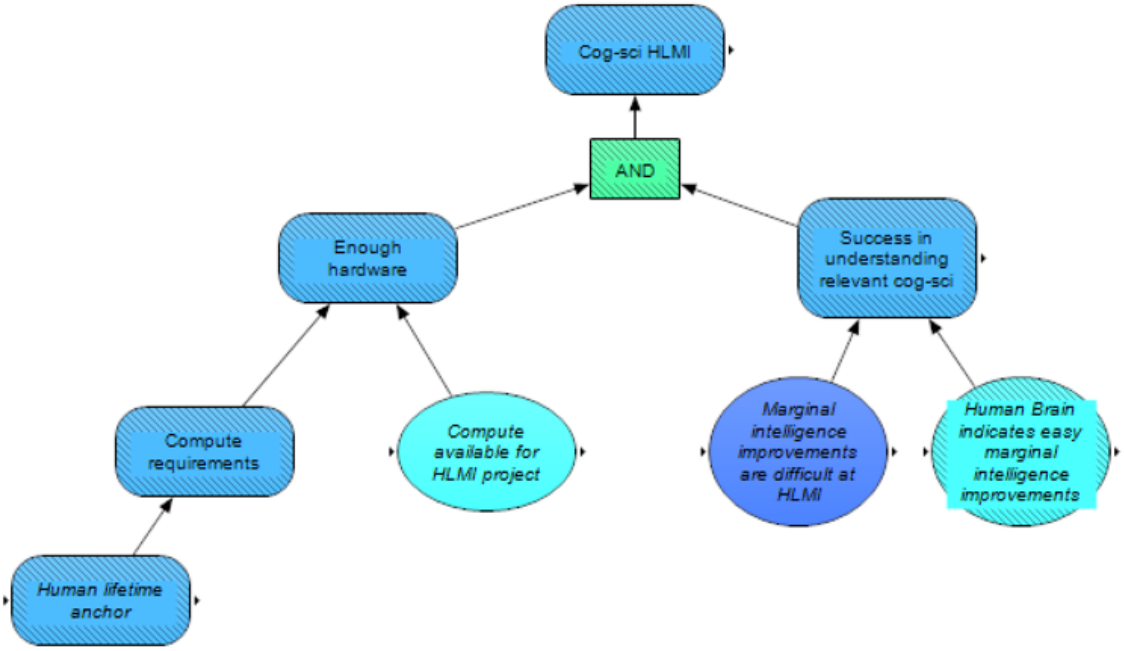
Let's consider how a couple of archetypal views about current methods in deep learning can be represented in this model. If one believes that DL will reach HLMI soon after matching the compute in the human brain (for instance, due to the view that we could make up for a lack of fine-tuning from evolution with somewhat larger scale), then this would correspond to high weight on the *human lifetime anchor*, in combination with a "no" resolution for the *hard paths hypothesis*. On the other end of the spectrum, if one were to believe that "current approaches in DL can't scale to HLMI", then this would presumably correspond to either a sufficiently strong "yes" to the *hard paths hypothesis* (that is, "current methods could scale if we had the right environment or data, but such data will not be created/aggregated, and the wrong kind of data won't lead to HLMI"), or a high weight to the *modified evolutionary anchor*, with, presumably, either a very large factor for *our methods getting "stuck"* ("evolution was only able to find intelligence due to evolution having advantages over our algorithms"), or a large factor for *"luckiness" of evolution* ("evolution got astronomically lucky to reach human-level intelligence, and even with evolutionary amounts of compute, we'd still be astronomically unlikely to reach HLMI").

Hybrid statistical-symbolic AI



Many people who are doubtful about the ability of current statistical methods to reach HLMI instead think that a hybrid statistical-symbolic approach (e.g., DL + GOFAI) could be more fruitful. Such an approach would involve achieving the necessary *hardware* (similar to before) and creating the necessary symbolic methods, data, and other aspects of the *software*. Here, we model the required hardware as being related to the *required hardware for current deep learning + BAU (business-as-usual) advancements* (described in the above section), modified with a *factor for hybrid methods requiring more/less compute*. As we are not sure how best to model achieving the necessary software (we are open to relevant suggestions), our model of this is very simple – we assume that such software is probably easier to develop if *marginal intelligence improvements are easier*, and further that the stronger one buys the bitter lesson, that there is a history of more naive, compute-leveraging, *brute-force methods winning* compared to more manually crafted methods, the less one should suspect such hybrid software is feasible.

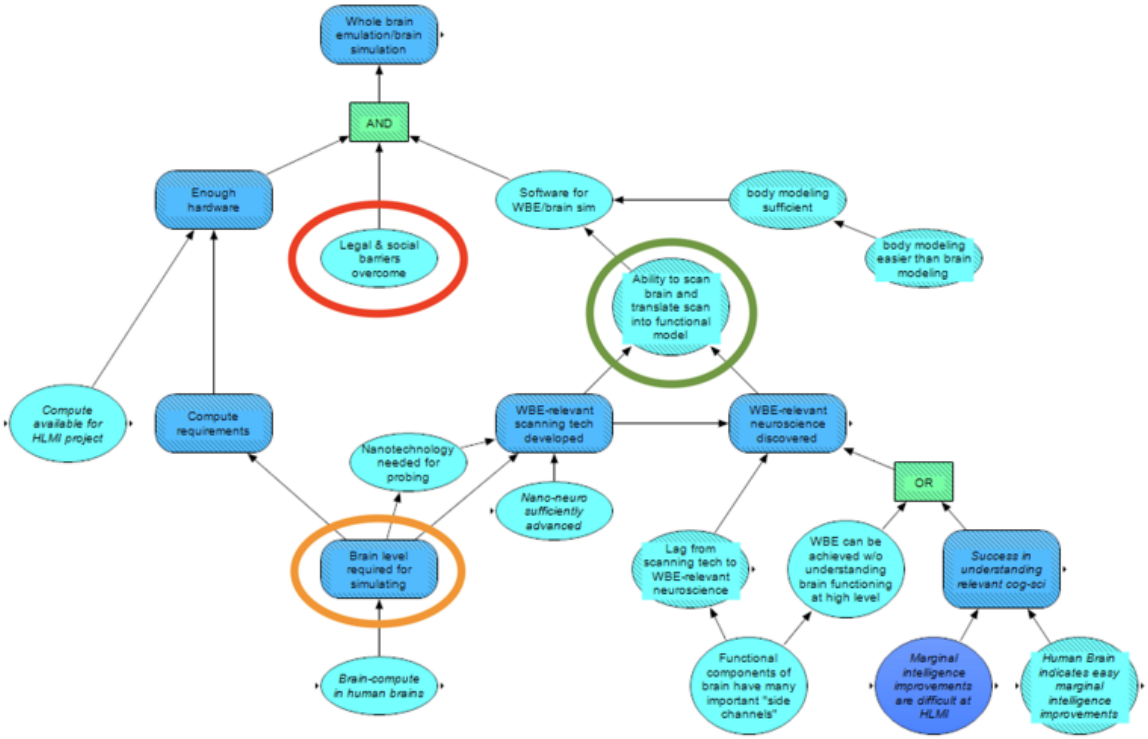
Cognitive-science approach



Similar to the above section, some people who are doubtful about the prospects of deep learning have suggested insights from cognitive science and developmental psychology would enable researchers to achieve HLMI. Again, this approach requires the relevant *hardware* and *software* (where the software is now dependent on understanding the relevant cognitive science). The hardware requirement here is estimated via the *human lifetime anchor*, as the cognitive science approach is attempting to imitate the learning processes of the human mind more closely. Again here, we are unsure how to model the likelihood in software success (ideas for approaching this question would be appreciated). For the time being, we assume that, similar to with other methods, the *easier marginal intelligence is*, the more likely this method will work, and in particular *evidence from the brain* would be particularly informative here, as a simpler human brain would presumably imply an easier time copying certain key features of the human mind.

Whole brain emulation/brain simulation

These methods would involve running a model of a brain (in the case of WBE, of a specific person's brain, and in the case of brain simulation, of a generic brain) on a computer, by modeling the structure of the brain and the information-processing dynamics of the lower-level parts of the brain, as well as integrating this model with a body inside an environment (either a virtual body and environment, or a robotic body with I/O to the physical world). To be considered successful for our purposes here, the model would have to behave humanlike (fulfilling the criteria for HLMI), but (in the case of WBE) would not need to act indistinguishable from the specific individual whose brain was being emulated, nor would it have to accurately predict the individual's behavior. We can imagine, by analogy to weather forecasting, that accurately forecasting the weather is much harder than simply forecasting weather-like behavior, and similarly the task outlined here is likely far easier than creating a WBE with complete fidelity to an individual's behavior, especially given the potential for chaos.



As before, our Analytica model considers cruxes for fulfilling *hardware* and *software* requirements, but we also include a node here for *overcoming the legal and social barriers* (circled in red). This is due to the fact that many people find the idea of brain emulation/simulation either ethically fraught or intuitively disturbing and unpalatable, and it is therefore less likely to be funded and pursued, even if technologically feasible.

As before, the hardware question is determined by *availability* and *requirements*, with requirements dependent on the *scale at which the brain must be simulated* (e.g., spiking neural network, concentrations of neurotransmitters in compartments, and so on) – circled in orange. We consider that the more *brain-compute that the human brain uses* (number of neurons times brain-compute per neuron), the lower scale we should assume must be simulated, as important information processing is then presumably occurring in the brain at lower scales (interactions on lower scales tend to be faster and larger in absolute number); however, the amount of brain-compute the brain uses is modeled here as a lower bound on computational costs, as emulating/simulating might require modeling details less efficiently than how the brain instantiates its compute (in the same way that a naive simulation of a calculator – simulating the electronic behavior in the various transistors and other electronic parts – would be much more computationally expensive than simply using the calculator itself).

On the software side, success would be dependent on the ability to *scan a brain and translate the scan into a functional model* (circled in green) and to *model a body sufficiently* (with one potential crux for whether *body modeling is easier than brain modeling*). Scanning the brain sufficiently, in turn, is dependent on *WBE-relevant scanning technology* enabling the creation of a structural model of the brain (at whatever scale is necessary to model), while translating the scan into a model would depend on the discovery of *WBE-relevant neuroscience* that allows for describing the information-processing dynamics of the parts of the brain. Note that, depending on the scale necessary to consider for the model, such "neuroscience" might include understanding of significantly lower-level behavior than what's typically studied in neuroscience today.

Discovering the relevant neuroscience, in turn, would depend on a few factors. First, relevant scanning technology would need to be created, so that the dynamics of the brain could be probed adequately. Such scanning technology would presumably include some similar scanning technology as mentioned above for creating a structural model, though would also likely go beyond this (our model collapses these different types of scanning technology down to one node representing all the necessary scanning technology being created). Second, after such scanning technology was created, there would plausibly be a *lag from the scanning technology to the neuroscience*, as somewhat slow, physical research would presumably need to be carried out with the scanning technology to discover the relevant neuroscience.

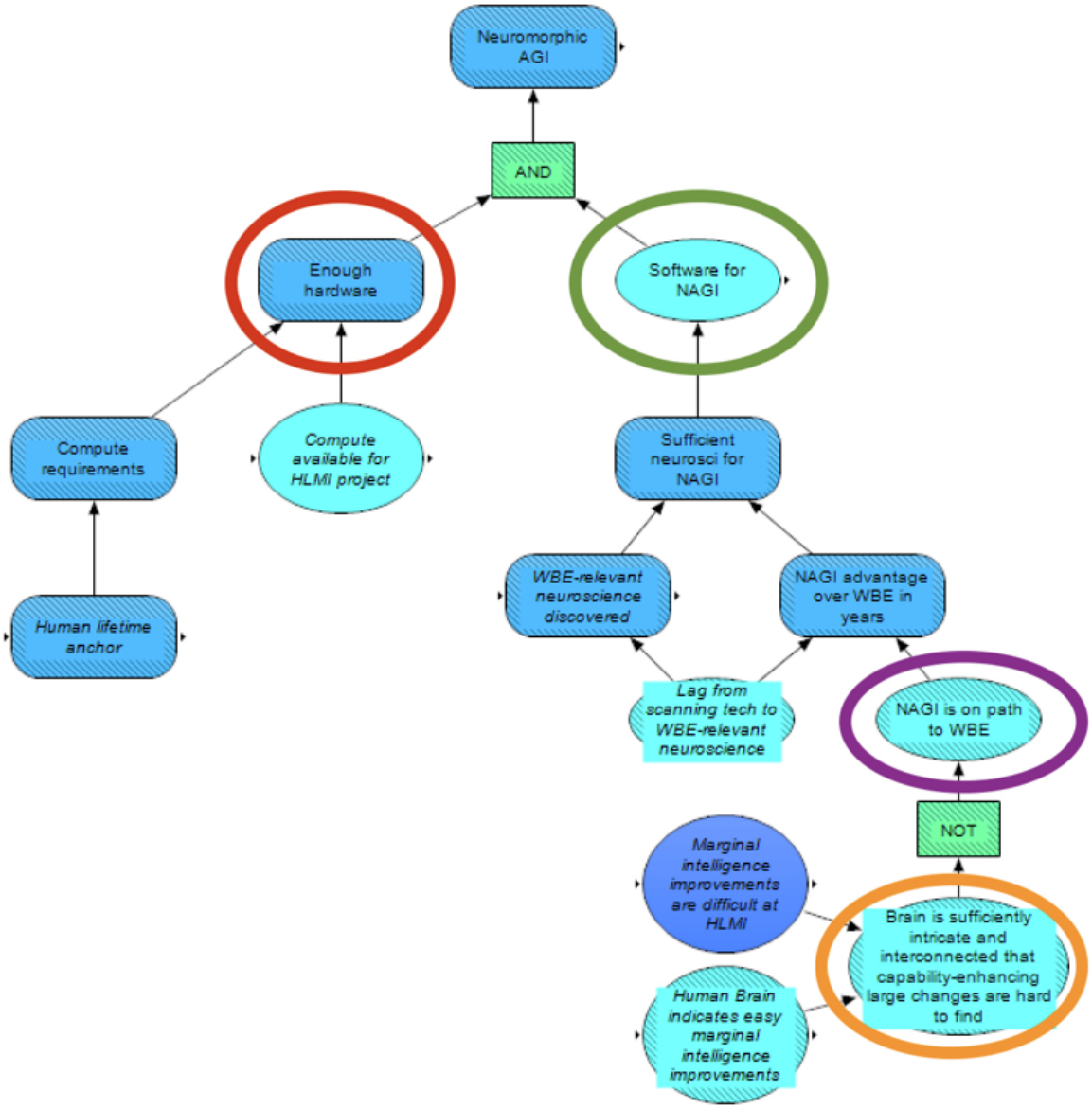
Third, neuroscientific success would depend on either brain emulation/simulation being *achievable without understanding the brain functioning at a high level*, or there would need to be success in *understanding sufficient cognitive science* (which we proxy with the same understanding necessary for a cognitive-science HLMI). Such higher-level understanding may be necessary for validating the proper functioning of components on various scales (e.g., brain regions, large-scale brain networks, etcetera) before integrating them.

We plan to model both the length of the lag and whether WBE can be achieved without understanding the higher-level functioning of the brain as dependent on whether there are *many important "side channels"* for the I/O behavior of the functional components of the brain. If there are many such side channels, then we expect finding all of them and understanding them sufficiently would likely take a long time, increasing the lag considerably. Furthermore, validating the proper functioning of parts in this scenario would likely require reference to proper higher-level functioning of the brain, as the relevant neural behavior would resultantly be incredibly messy. On the other hand, if there are not many such side channels, discovering the proper behavior of the functional parts may be relatively quick after the relevant tools for probing are invented, and appropriate higher-level functioning may emerge spontaneously from putting simulated parts together in the right way. Arguments in favor of there being many side channels typically make the point that the brain was biologically evolved, not designed, and evolution will tend to exploit whatever tools it has at its disposal in a messy way, without heed to legibility. Arguments against there being many such side channels typically take the form that the brain is fundamentally an information-processing machine, and therefore its functional parts, as signal processors, should face strong selection pressure for maintaining reliable relationships between inputs and outputs – implying relatively legible and clean information-processing behavior of parts.

Developing the necessary scanning technology for WBE, meanwhile, also depends on a few factors. The difficulty of developing the technology would depend on the brain level required to model (mentioned above). Additionally, if the scanning technology *requires nanotechnology* (likely dependent on the brain level required for simulating), then an important crux here is whether *sufficient nano-neurotechnology is developed*.

Neuromorphic AGI (NAGI)

To avoid one point of confusion between this section and the two previous sections: the previous section (WBE/brain simulation) is based on the idea of creating a virtual brain that operates similar to a biological brain, and the section before that (cognitive-science approach) is based on the idea of creating HLMI that uses many of the higher-level processes of the human mind, but that isn't instantiated on a lower-level in a similar manner to a biological brain (e.g., such an approach might have modules that perform similar functions to brain regions, but without anything modeling bio-realistic neurons). In contrast, this section (neuromorphic AGI) discusses achieving HLMI via methods that use lower-level processes from the brain, but without putting them together in a manner to create something particularly human mind-like (e.g., it may use models of bio-realistic neurons, but in ways at least somewhat dissimilar to human brain structures).



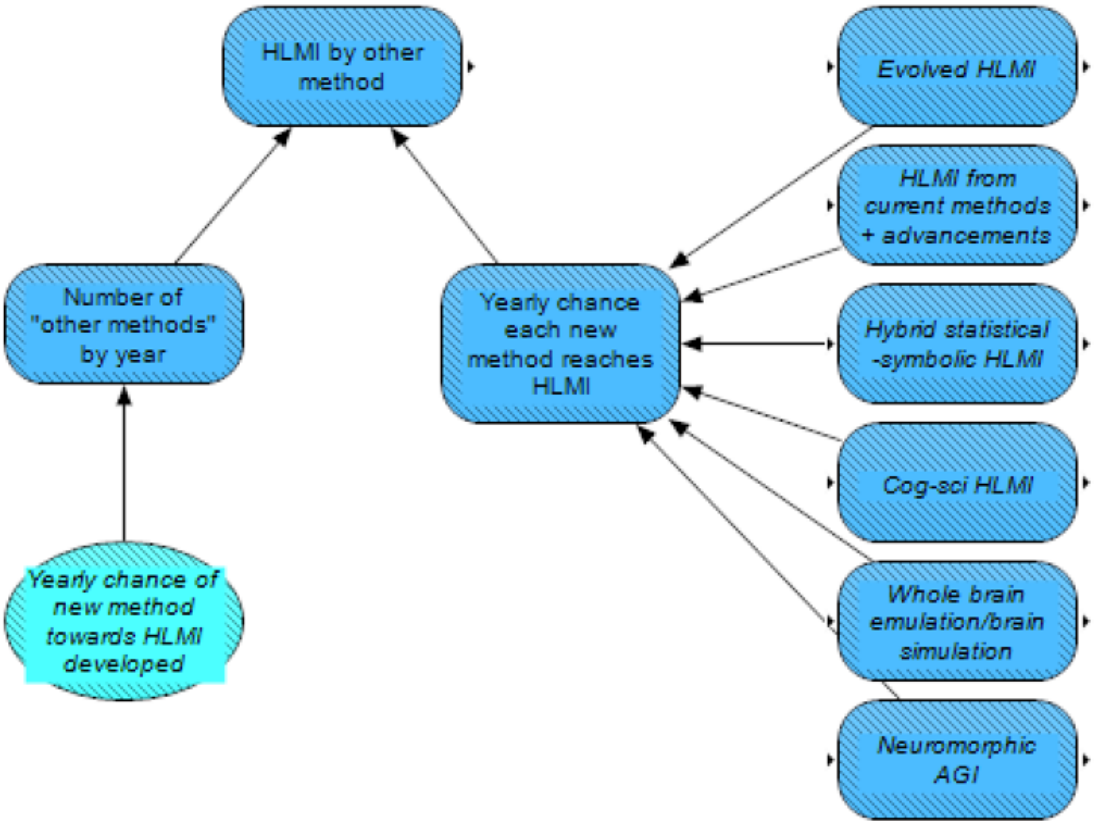
Similar to other methods, NAGI is modeled as being achieved if the relevant *hardware* (circled in red) and *software* (circled in green) are achieved. As with the cognitive-science approach, the *human lifetime anchor* is used for the *compute requirement* (which is compared against the *available compute* for whether there is enough hardware), as this method is attempting to imitate more humanlike learning.

On the software side, the main crux is whether *NAGI is on the path to WBE* (circled in purple). If so, then the relevant *neuroscience for NAGI* (sufficient for the software for NAGI) is assumed to be discovered before the *relevant neuroscience for WBE*. The amount of time *advantage that NAGI has over WBE* in terms of relevant neuroscience under this condition is then assumed to be less than or equal to the *lag from scanning technology to WBE-relevant neuroscience*; that is, the neurotechnology needed for NAGI is implicitly assumed to be the same as for WBE, but gathering the relevant neuroscience for NAGI is assumed to be somewhat quicker.

Whether NAGI is on the path to WBE, however, is dependent on whether the *brain is sufficiently intricate and interconnected that large changes are almost guaranteed to worsen*

capabilities (circled in orange; if so, then NAGI is assumed to not be on the path to WBE, because potential “modifications” of brain architecture to create NAGI would almost definitely hurt capabilities). This crux is further influenced by the *difficulty of marginal intelligence improvements*, with particular emphasis on *evidence from the human brain*.

Other Methods

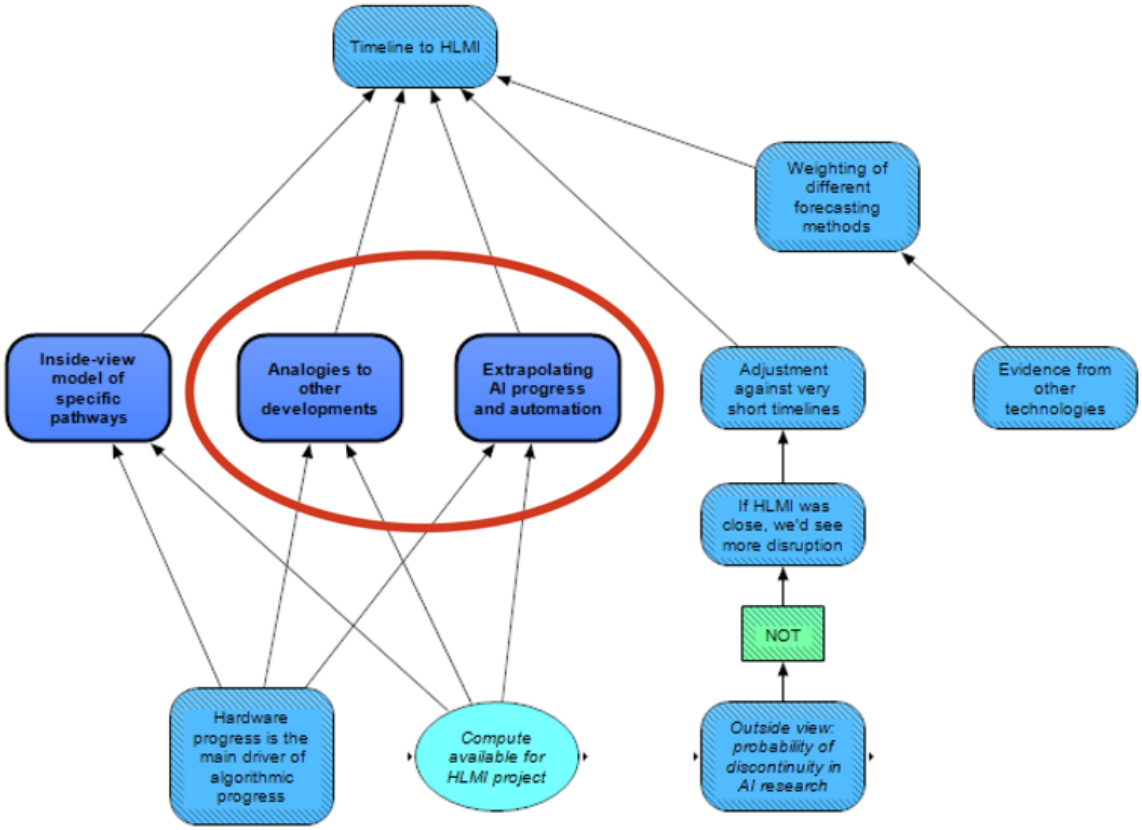


As a catch-all for other possible methods, we have a section for other methods, which depends on the *yearly chance of new methods being developed* and the *yearly chance each of these new methods reaches HLMI*.

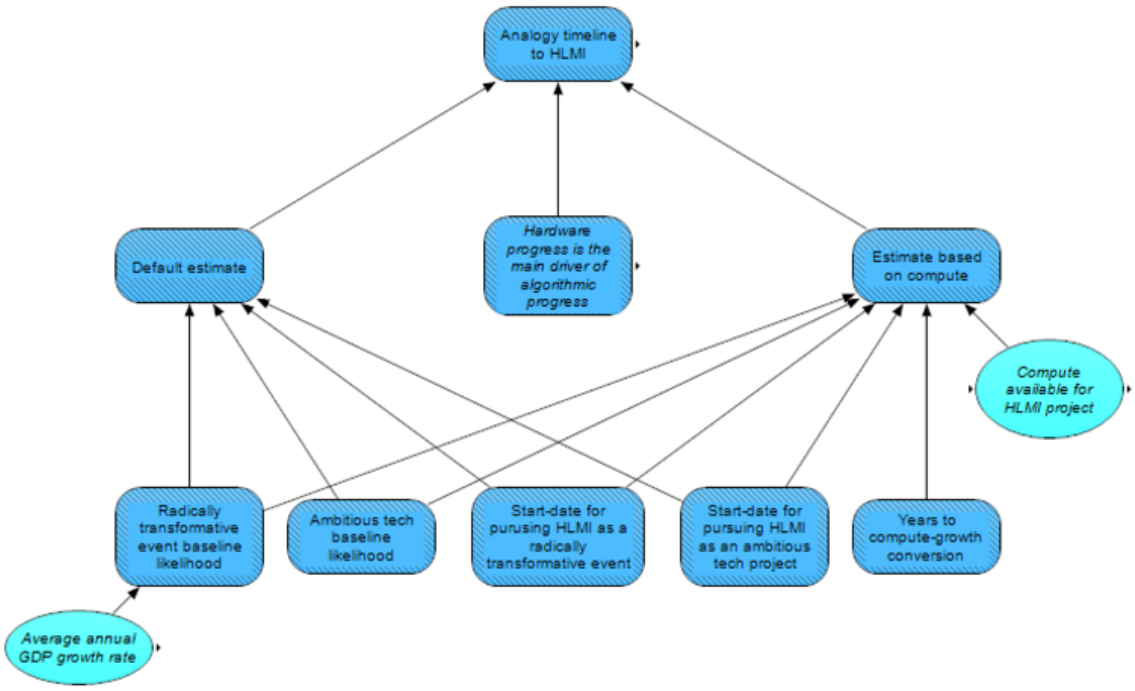
As both of these uncertainties are deeply unknown, we plan on taking a naive, outside-view approach here. (We are open to alternative suggestions or improvements.) The field of AI arguably goes back to 1956, and over these 65 years, we now arguably have around six proposed methods for how to plausibly get to HLMI (i.e., the other six methods listed here). Other characterizations of the different paths may yield a different number, for example by combining the first two methods into one, or breaking them up into a few more, but we expect most other characterizations would not tend to differ from ours drastically so – presumably generally within a factor of 2. Taking the assumption of six methods developed within 65 years at face value, this indicates a yearly chance of developing a new method towards HLMI of ~9%. (Though we note this assumption implies the chances of developing a new method is independent each year to the next, and this assumption is questionable.) For the second uncertainty (the chance the methods each reach HLMI in a year), the relationship is also unclear. One possible approach is to simply take the average chance of all the other methods (possibly with a delay of ~20 years for the method to “scale up”), but this will likely be informed by further discussion and elicitation.

Outside-view approaches

In addition to the inside-view estimation for HLMI timelines based on specific pathways, we also estimate HLMI arrival via outside-view methods – *analogies to other developments* and *extrapolating AI progress and automation*.



Analogies to other developments



For this module, we take an approach similar to Tom Davidson's in his Report on Semi-informative Priors, though with some simplifications and a few different modeling judgements.

First, the likelihood of HLMI being developed in a given year is estimated by analogizing HLMI to other classes of developments. This initial, baseline likelihood is estimated in a very naive manner – blind to the history of AI, including to the fact that HLMI has not yet been developed, and instead only considering the base rate for "success" in these reference classes. The two classes of events we use (in agreement with Davidson's two preferred classes) are: *highly ambitious but presumably physically possible projects seriously pursued by the STEM community* (e.g., harnessing nuclear energy, curing cancer, creating the internet), and *radically transformative events for human work and all of human civilization* (the only examples in this reference class are the Agricultural Revolution and the Industrial Revolution).

For both of these reference classes, different methods can be used to estimate the baseline yearly likelihood of "success"; here, we sketch out our current plans for making such estimates. For the ambitious-STEM-projects class, the baseline yearly likelihood can be estimated as simply taking the number of successes for such projects (e.g., 2 for the above three listed ones, as nuclear energy and the internet were successful, while a cure from cancer has not been successful yet) and dividing this number by the sum of the number of years that each project was seriously pursued by the STEM community (e.g., around 5-10 years for pursuing nuclear energy, plus around 50 years for pursuing a cure for cancer, plus perhaps 30 years for the development of the internet, implying, if these three were a representative list, a baseline yearly likelihood of around (2 successes)/(8 + 50 + 30 years) = ~2.3%). Disagreement exists, however, with which projects are fair to include on such a reference-class list, as well as when projects may have been "seriously" started and if/when they were successful.

For the radically-transformative-events class, the appropriate calculation is a bit fuzzier. Obviously it wouldn't make sense to simply take a per-year frequentist approach similar to with the ambitious-STEM-projects class – such an estimate would be dominated by the tens of thousands or hundreds of thousands of years it took before the Agricultural Revolution,

and it would ignore that technological progress and growth rates have sped up significantly with each transition. Instead, based on the idea that human history can be thought of as broken down into a sequence of paradigms with increasing exponential economic growth rates and ~proportionately faster transitions between these paradigms (with the transformative events marking the transition between the different paradigms), we consider that comparisons between paradigms should be performed in terms of economic growth.

That is, if the global economy doubled perhaps ~2-10 times between the first first humans and the Agricultural Revolution (depending on whether we count from the beginning of *Homo sapiens* ~300,000 years ago, the beginning of the genus *Homo* ~2 million years ago, or some middle point), and ~8 times between the Agricultural Revolution and the Industrial Revolution, then, immediately after the Industrial Revolution, we might assume that the next transition would similarly occur after around ~5-10 economic doublings. Using similar logic to the ambitious-STEM-projects class, we would perhaps be left with a baseline likelihood of a transformative event per economic doubling of ~2/(6 + 8) = ~14% (taking the average of 2 and 10 for the Agricultural Revolution). If we assume 3% yearly growth in GDP post-Industrial Revolution, the baseline likelihood in terms of economic growth can be translated into a yearly baseline likelihood of ~0.7%. Thus, we have two estimates for baseline yearly likelihoods of the arrival of HLMI, based on two different reference classes.

Second, we consider how the yearly likelihood of developing HLMI may be expected to increase based on how long HLMI has been unsuccessfully pursued. This consideration is based on the expectation that the longer HLMI is pursued without being achieved, the harder it presumably is, and the less likely we may expect it to succeed within the next year. Similar to Davidson, and resembling a modified version of Laplace's rule of succession, we plan to set the probability of HLMI being achieved in a given year as $P = 1/(Y + m)$, where Y is the number of years HLMI has been unsuccessfully pursued, and m is set based on considerations from the baseline yearly likelihood, in a manner described below.

Determining Y is necessarily somewhat arbitrary, as it requires picking a "start date" to the pursuit of HLMI. We consider that the criteria for a start date should be consistent with those for the other developments in the reference class (and need not be the same for the different reference classes). For the ambitious-STEM-project reference class, the most reasonable start date is presumably when the STEM community began seriously pursuing HLMI (to a similar degree that the STEM community started pursuing nuclear energy in the mid-1930s, a cure for cancer in the 1970s, and so on) – arguably this would correspond to 1956. For the radically-transformative-events reference class, on the other hand, the most reasonable start date is arguably after the Industrial Revolution ended (~1840), as this would be consistent with each of the other transitions being "attempted" from the end of the last transition.

To estimate m, we consider that when HLMI is first attempted, we may expect the development of HLMI to take around as long as implied by the baseline yearly likelihood. Thus, we plan to set m (separately for each of the reference classes) so that the cumulative distribution function of P passes the 50% mark in $Y = 1/(\textit{baseline yearly likelihood})$. That is, we assume in the first year that HLMI is pursued, that with 50% odds HLMI will be achieved earlier than implied by the baseline yearly likelihood, and with 50% odds it will be achieved later (continuing the examples from above, apparently this would place $m = 44$ for the ambitious-STEM-project reference class, and $m = 143$ for the radically-transformative-events reference class).

Third, we update based on the fact that, as of 2021, HLMI has not yet been developed. For example, if we take the ambitious-STEM-project reference class, and assume that the first date of pursuit is 1956, then for 2022 we set Y to (2022 – 1956) = 66 (and 67 for 2023, and so on), and keep m at the previous value (implying, continuing from the dubious above example again, a chance of HLMI in 2022 of 1/(66 + 44) = ~0.9%). Similarly for the radically-transformative-events reference class, if HLMI is assumed to initially be pursued in 1840, then for 2022 Y is set to 182, and we'd get a chance of HLMI in 2022 of 1/(182 + 143) = ~0.3%.

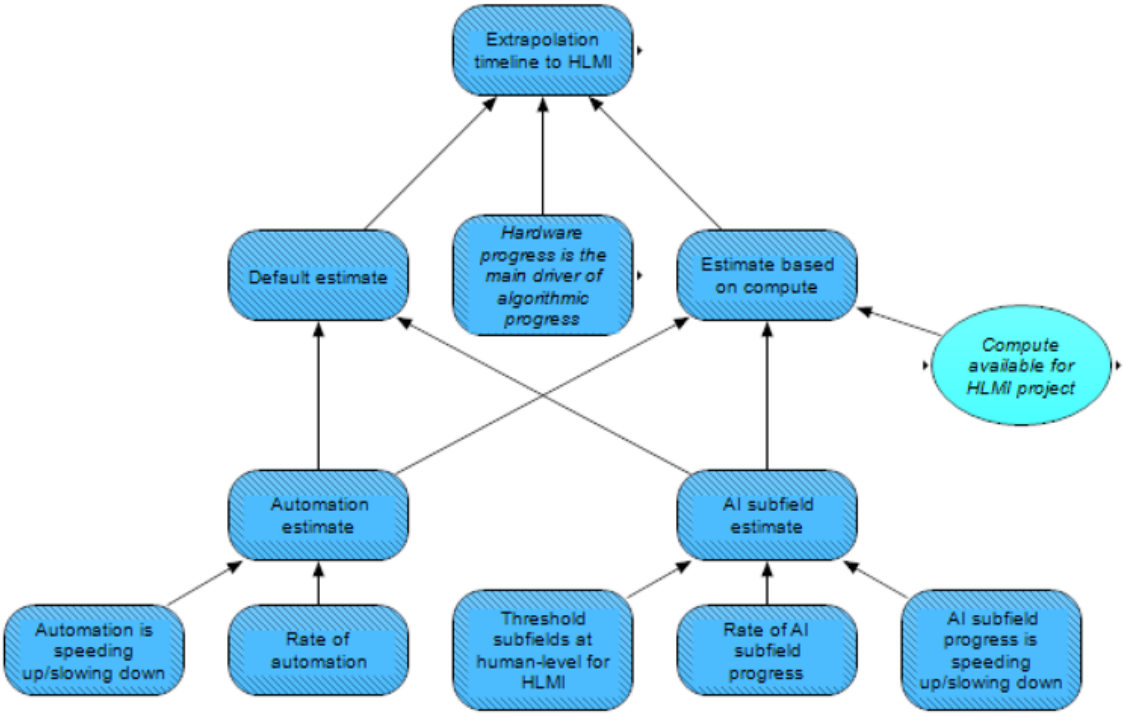
Finally, to account for the possibility that *hardware progress is the main driver of algorithmic progress*, we duplicate both of these calculations, in terms of hardware growth instead of in terms of time. In this case, the probability of HLMI being achieved in a given doubling of compute available can be calculated as $P = 1/(C + n)$, where C is the number of compute-doublings since HLMI has first been pursued (using a consistent definition for when it initially was pursued as above), and n takes the place of m, by here being valued such that the cumulative distribution function of P passes the 50% mark after the number of compute-doublings that we might *a priori* (when HLMI is first pursued) expect to be needed to achieve HLMI.

For the ambitious-STEM-project class, this latter calculation requires a "conversion" in terms of technological progress between years of pursuit of other ambitious projects and compute-growth for AI. This conversion may be set for the number of years necessary for other projects in the reference class to make similar technological progress as a doubling of compute does for AI.

For the radically-transformative-events reference class, the switch to a hardware-based prediction would simply replace economic growth since the end of the Industrial Revolution with hardware growth since the end of the Industrial Revolution (that is, considering a doubling in compute in the new calculation to take the place of a doubling in global GDP in the old calculation). This comparison may on its face seem absurd, as the first two paradigms are estimated based on GDP, while the third is based on compute (which has recently risen much faster). However, we may consider that the important factor in each paradigm is growth in the main inputs towards the next transition. For the Agricultural Revolution, the most important input was arguably the number of people, and for the Industrial Revolution, the most important input was arguably economic size. Both of these are proxied by GDP, because before the Agricultural Revolution, GDP per capita is assumed to be approximately constant, so growth in GDP would simply track population growth. For the development of HLMI, meanwhile, if it is the case that *hardware progress is the main driver of algorithmic progress*, then the main input towards this transition is compute, so the GDP-based prediction may severely underestimate growth in the relevant metric, and the compute-based estimate may in fact be somewhat justified.

It should be noted that in the compute-based predictions, timelines to HLMI are heavily dependent on what happens to compute going forward, with a potential leveling off of compute implying a high probability of very long timelines.

Extrapolating AI progress and automation



In this module, HLMI arrival is extrapolated from the rate of automation (i.e., to occur when extrapolated automation hits 100%), and from progress in various AI subfields. Such an extrapolation is perhaps particularly appropriate if one is expecting HLMI to arrive in a piecemeal fashion, with different AIs highly specialized for automating specific tasks in a larger economy. Note that questions about the distribution versus concentration of HLMI are handled by a downstream module in our model, covered in a subsequent post.

For the extrapolation from automation, this calculation depends on the *rate of automation*, and whether this rate has generally been *speeding up, slowing down, or remaining roughly constant*.

For the extrapolation from the progress in various AI subfields, a somewhat similar calculation is performed, considering again the *rate of progress* and whether this progress has been *speeding up, slowing down, or remaining roughly constant*. HLMI is assumed to be achieved here when enough subfields are extrapolated to reach human-level performance.

One crux for this calculation, however, is the *threshold for what portion of subfields need to reach human level in the extrapolation* to reach HLMI. One could imagine that all subfields may need to reach human level, as HLMI wouldn't be achieved until AI was as good as humans in all these subfields, but this threshold would introduce a couple problems. First, if certain subfields happen to be dominated by researchers that are more conservative in their estimations, then the most pessimistic subfield could bias the results to be too conservative. Second, it's possible that some subfields are bottlenecked by sufficient success in other subfields, and will see very slow progress before these other subfields reach sufficient capabilities. Alternatively, one could imagine that the response from the median subfield may be the least biased, and also be the most appropriate for gauging "general" intelligence. On the other hand, if AI achieving human-level competence on half of the domains did not translate into similar competence in other domains, then this would not imply HLMI.

Similar to the above section on analogies to other developments, the extrapolation here is done in two forms: as a default case, it is performed in terms of time, but if *hardware progress is the main driver of algorithmic progress*, then the extrapolation is performed in

terms of (the logarithm of) compute. In the latter scenario, a slowdown in compute would lead to a comparable slowdown in AI capability gains, all else equal.

Bringing it all together

Here, we have examined several methods for estimating the timeline to HLMI: a gears-level model of specific pathways to HLMI, as well as analogizing HLMI to other developments in plausibly similar reference classes, and additionally extrapolating current trends in automation and AI progress. Disagreements exist regarding the proper weight for these different forecasting methods, and further disagreements exist for many factors underlying these estimations, including some factors that appear in several different places (e.g., the progression of hardware and cruxes related to analogies and general priors on intelligence).

In addition to estimating the timeline to HLMI, our model also allows for estimating the method of HLMI first achieved – though such an estimate only occurs in one of our three forecasting methods (we welcome suggestions for how to make this estimate in model runs where other methods are "chosen" by the Monte Carlo method dice roll, as this information is important for downstream nodes in other parts of our model).

In the next post in this series, we will discuss AI takeoff speeds and discontinuities around and after HLMI.

Acknowledgments

We would like to thank both the rest of the MTAIR project team, as well as the following individuals, for valuable feedback on this post: Edo Arad, Lukas Finnveden, Ozzie Gooen, Jennifer Lin, Rohin Shah, and Ben Snodin

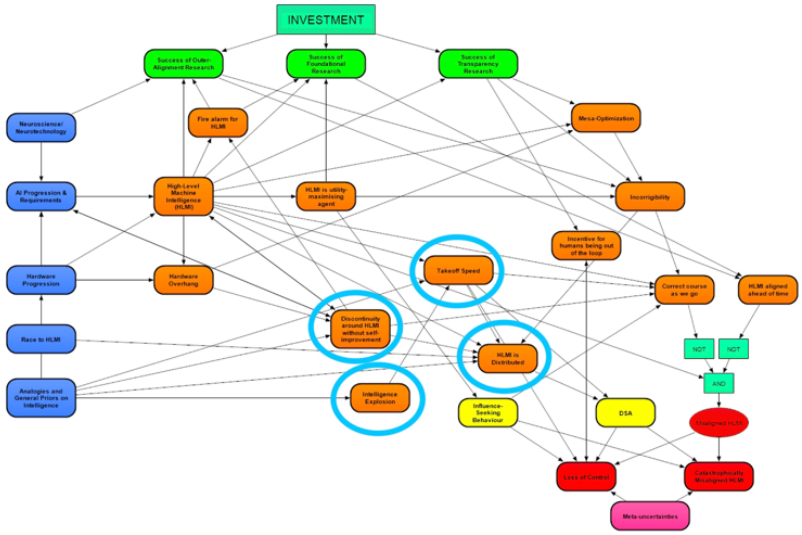
Takeoff Speeds and Discontinuities

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is part 4 in our sequence on Modeling Transformative AI Risk . We are building a model to understand debates around existential risks from advanced AI. The model is made with Analytica software, and consists of nodes (representing key hypotheses and cruxes) and edges (representing the relationships between these cruxes), with final outputs corresponding to the likelihood of various potential failure scenarios. You can read more about the motivation for our project and how the model works in the Introduction post . The previous post in the sequence, Paths to High-Level Machine Intelligence , investigated how and when HLMI will be developed.

We are interested in feedback on this post, especially in places where the model does not capture your views or fails to include an uncertainty that you think could be an important crux. Similarly, if an explanation seems confused or confusing, flagging this is useful – both to help us clarify, and to ensure it doesn't reflect an actual disagreement.

The goal of this part of the model is to describe the different potential characteristics of a transition from (pre-)HLMI[1] to superintelligent AI (i.e., "AI takeoff"). We also aim to clarify the relationships between these characteristics, and explain what assumptions they are sensitive to.



As shown in the above image, the relevant sections of the model (circled in light blue) take inputs primarily from these modules:

- *Analogies and General Priors on Intelligence* – this module concerns both arguments that compare HLMI development with other, previous developments (for example, the evolution of human intelligence, current progress in machine learning, and past historical/cultural transitions) and broad philosophical claims about the nature of intelligence
- *High-Level Machine Intelligence (HLMI)* – this module concerns whether HLMI will be developed at all, and if so, which type of HLMI and when
- *Hardware Overhang* – this module concerns whether, at the time HLMI is developed, the amount of hardware required to run HLMI and the availability of such hardware will lead to a situation of "hardware overhang", where overwhelming hardware resources allow for the first HLMI(s) to be overpowered due to being cheaply duplicated, easily run at a high clock speed, or otherwise rapidly scaled up.

The outputs of the sections of concern in this post, corresponding to the above circled modules, are:

- *Intelligence Explosion* – will a positive feedback loop involving AI capabilities lead these capabilities to grow roughly hyperbolically across a sufficient range, such that the capabilities eventually grow incredibly quickly to an incredibly high level (presumably before plateauing as they approach some fundamental limit)?
- *Discontinuity around HLMI without self-improvement* – will there be a rapid jump in AI capabilities from pre-HLMI AI to HLMI (for instance, if pre-HLMI acts like a machine with missing gears) and/or from HLMI to much higher intelligence (for instance, if a hardware overhang allows for rapid capability gain soon after HLMI)?
- *Takeoff Speed* – how fast will the global economy (or the next closest thing, if this concept doesn't transfer to a post-HLMI world) grow, once HLMI has matured as a technology?
- *HLMI is Distributed* – will AI capabilities in a post-HLMI world be dispersed among many comparably powerful systems? A negative answer to this node indicates that HLMI capabilities will be concentrated in a few powerful systems (or a single such system).

These outputs provide a rough way of characterizing AI takeoff scenarios. While they are non-exhaustive, we believe they are a simple way of characterizing the range of outcomes which those who have seriously considered AI takeoff tend to find plausible. For instance, we can summarize (our understanding of) the publicly espoused views of Eliezer Yudkowsky, Paul Christiano, and Robin Hanson (along with a sceptic position) as follows:

	Eliezer Yudkowsky	**Paul Christiano**	**Robin Hanson**	**Sceptic**

Intelligence Explosion	Yes	Yes	No	No
Discontinuity around HLMI without self-improvement	Yes	No	No	No
Takeoff Speed	~hyperbolically increasing *(No significant intermediate doublings)*	~hyperbolically increasing *(with complete intermediate doublings on the order of ~1 year)*	Doubling time of ~weeks to months	Doubling time on the order of ~years or longer
HLMI is Distributed	No	Yes	Yes	Yes

These outputs – in addition to outputs from other sections of our model, such as those covering misalignment – impact further downstream sections of our model relevant to failure modes.

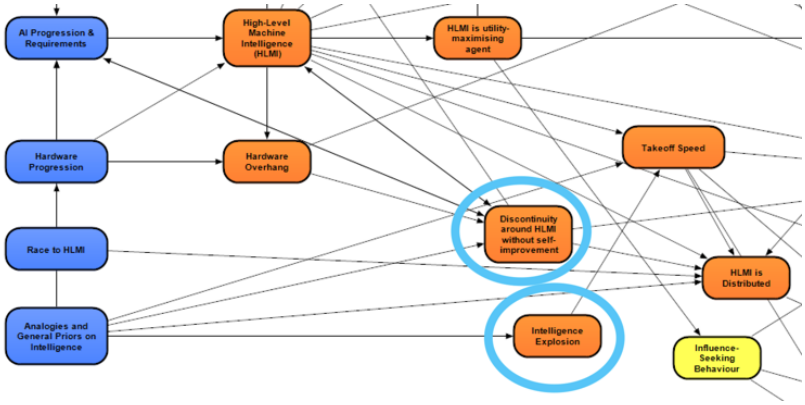
In a previous post we explored the *Analogies and General Priors on Intelligence* module, which includes very important input for the modules in this post. That module outputs answers to four key questions (which are used throughout this later module):

- The difficulty of marginal intelligence improvements at the approximate human level (i.e., around HLMI)
- Whether marginal intelligence improvements become increasingly difficult beyond HLMI at a rapidly growing rate or not
  - 'Rapidly growing rate' is operationalized as becoming difficult exponentially or faster-than-exponentially
- Whether there is a fundamental upper limit to intelligence not significantly above the human level
- Whether, in general, further improvements in intelligence tend to be bottlenecked by previous improvements in intelligence rather than some external factor (such as the rate of physics-limited processes)

In the next section of this post, we discuss the *Intelligence Explosion* and *Discontinuity around HLMI without self-improvement* modules, which are upstream of the other modules covered in this post. After that, we explore these other modules (which are influenced by the earlier modules): *HLMI is Distributed* and *Takeoff Speed*.

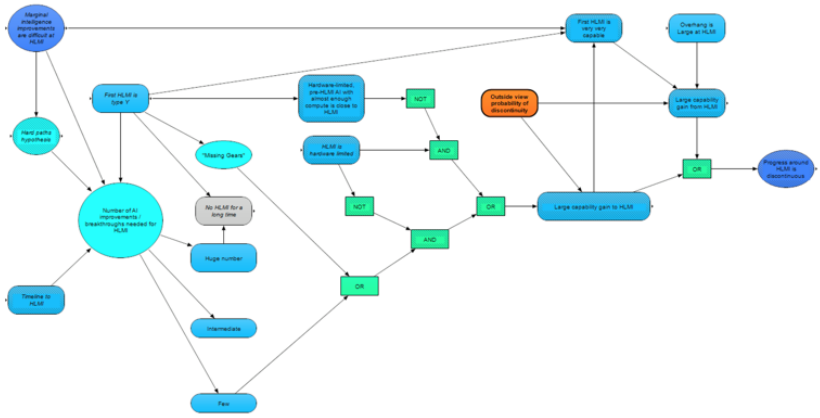
Early Modules (Discontinuity and Intelligence Explosion)

We now examine the modules *Discontinuity around HLMI without self-improvement* and *Intelligence Explosion*, which affect the two later modules – *HLMI is Distributed* and *Takeoff Speed*.



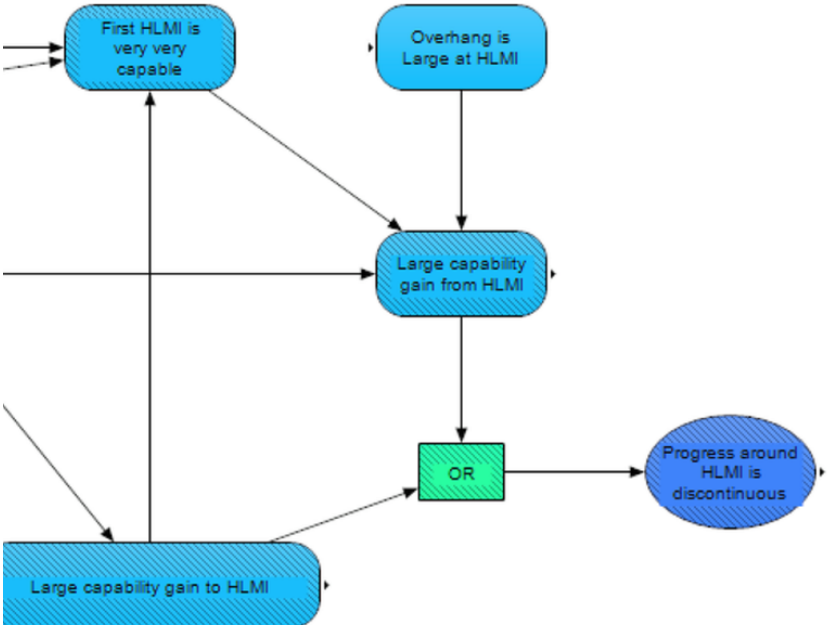
Discontinuity around HLMI without self-improvement

This module aims to answer the question: will the first HLMI (or a very early HLMI) represent a discontinuity in AI capabilities from what came before? We define a discontinuity as a very large and very sudden jump in AI capabilities, not necessarily a mathematical discontinuity, but a phase change caused by a significantly quicker rate of improvement than what projecting the previous trend would imply. Note that this module is NOT considering rapid self-improvement from quick feedback loops (that would be considered instead in the module on *Intelligence Explosion*), but is instead concerned with large jumps occurring around the time of the HLMI generation of AI systems.

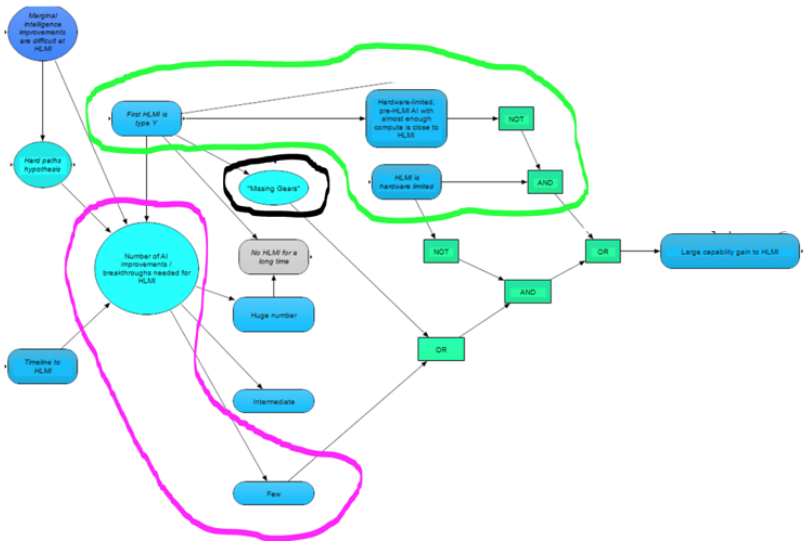


Such a discontinuity could be from a jump in capabilities to HLMI or a jump from HLMI to significantly higher capabilities (or both). A jump in capabilities from HLMI could occur, either if the first HLMI "overshoots" the HLMI-level and is very very capable (which likely depends on both the type of HLMI and whether marginal intelligence improvements are difficult around HLMI), or if a large hardware overhang allows for the first HLMI to scale (in quality or quantity) in such a way that it is far beyond the HLMI-level in capabilities.

The image below shows a zoomed-in view of the two routes (either a large capability gain **from** HLMI, or a large capability gain **to** HLMI):



Regarding capability jumps **to** HLMI, we see a few pathways, which can be broken down by whether HLMI will ultimately be bottlenecked on hardware or software (i.e., which of hardware or software will be last to fall into place – as determined by considerations in our post on Paths to HLMI). If HLMI will be bottlenecked on hardware (circled in green on the graph below), then the question reduces to whether pre-HLMI with almost enough compute has almost as strong capabilities as HLMI. To get a discontinuity from hardware-limited HLMI, the relationship between increasing abilities and increasing compute has to diverge from an existing trend to reach HLMI (i.e., hardware-limited, pre-HLMI AI with somewhat less compute is much less capable than HLMI with the required compute). We suspect that whether this crux is true may depend on the type of HLMI in question (e.g. statistical methods might be more likely to gain capabilities if scaled up and run with more compute).



If HLMI is software limited, on the other hand, then instead of hardware, we want to know whether the last software step(s) will result in a large jump in capabilities. This could happen either if there are very few remaining breakthroughs needed for HLMI (circled in magenta above) such that the last step(s) correspond to a large portion of the problem, or if the last step(s) act as "missing gears" putting the rest of the system in place (circled in black above).

We suspect that whether the last step(s) present a "missing gears" situation is likely to depend on the type of HLMI realized. A (likely) example of "missing gears" would be whole brain emulation (WBE), where 99% of the way towards WBE presumably doesn't get you anything like 99% of the capabilities of WBE. (*See here for an extended discussion of the relationship between "fundamental breakthroughs" and "missing gears".*) If the "missing gears" crux resolves negatively, however, then determining whether there will be a large capability gain to HLMI is modeled as depending on the number of remaining breakthroughs needed for HLMI.

We make the simplifying assumption that the remaining fundamental breakthroughs are all of roughly comparable size, such that the more breakthroughs needed, the less of a step any individual breakthrough represents. This means that the last breakthrough – the one that gives us HLMI – might either take us from AI greatly inferior to HLMI all the way to HLMI (if there are 'few' key breakthroughs needed), or just be an incremental improvement on pre-HLMI AI that is already almost as useful as HLMI (if there are an 'intermediate' or 'huge number' of key breakthroughs needed).

We consider several lines of evidence to estimate whether HLMI requires few or many key breakthroughs.

To start, the type of HLMI being developed influences the number of breakthroughs we expect will be needed. For instance, if HLMI is achieved with current deep learning methods plus business-as-usual advancements, then, *ceteris paribus*, we'd expect fewer breakthroughs needed to reach HLMI than if HLMI is achieved via WBE.

As well as depending on the type of HLMI developed, our model assumes the expected number of breakthroughs needed for HLMI is influenced significantly by the *Difficulty of Marginal Intelligence Improvements at HLMI* (in the *Analogies and General Priors on Intelligence* module). If marginal intelligence improvements are difficult at HLMI, then more separate breakthroughs are probably required for HLMI.

Lastly, the **hard paths hypothesis** and **timeline to HLMI** (as determined in the Paths to HLMI modules) each influence our estimate of how many breakthroughs are needed to reach HLMI. The hard paths hypothesis claims that it's rare for environments to straightforwardly select for general intelligence – if this hypothesis is true, then we'd expect more steps to be necessary (e.g., for crafting the key features of such environments). Additionally, short timelines would imply that there are very few breakthroughs remaining, while longer timelines may imply more breakthroughs needing to be found (remember, it's fine if this logic feels "backwards", as our model is not a causal model *per se*, and instead arrows represent probabilistic influence).

How many breakthroughs?

We don't place exact values on the number of breakthroughs needed for HLMI in either the 'few', 'intermediate' or 'huge number' cases. This is because we have not yet settled on a way of defining "fundamental breakthroughs", nor of estimating how many would be needed to shift the balance on whether there would be a large capability gain to HLMI.

Our current plan for characterizing the number of breakthroughs is to anchor the 'intermediate' answer to 'number of remaining breakthroughs' as similar to the number of breakthroughs that have so far occurred in the history of AI. If we identify that there have been 3 major paradigms in AI so far (knowledge engineering, deep search, and deep learning), and maybe ten times as many decently-sized breakthroughs (within deep learning, this means things like CNNs, the transformer architecture, DQNs) to get to our current level of AI capability, then an 'intermediate' case would imply similar numbers to come. From this, we have:

- Few Breakthroughs: 0-2 major paradigms and 0-9 new breakthroughs
- Intermediate: Similar number of breakthroughs as so far, or somewhat more: 3-9 new paradigms and 10-100 breakthroughs
- Huge number: More than 10 paradigms, 100 breakthroughs

Another way to estimate the number of remaining breakthroughs is to use expert opinion. For example, Stuart Russell identifies 4 remaining fundamental breakthroughs needed for HLMI (although these 'breakthroughs' seem more fundamental than those listed

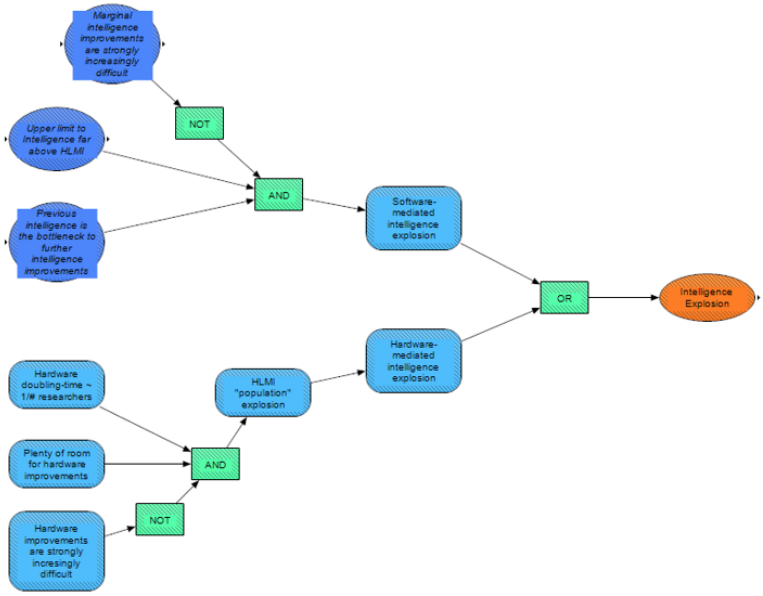
above, and might correspond to a series of breakthroughs as defined above):

> "We will need several conceptual breakthroughs, for example in language or common sense understanding, cumulative learning (the analog of cultural accumulation for humans), discovering hierarchy, and managing mental activity (that is, the metacognition needed to prioritize what to think about next)"

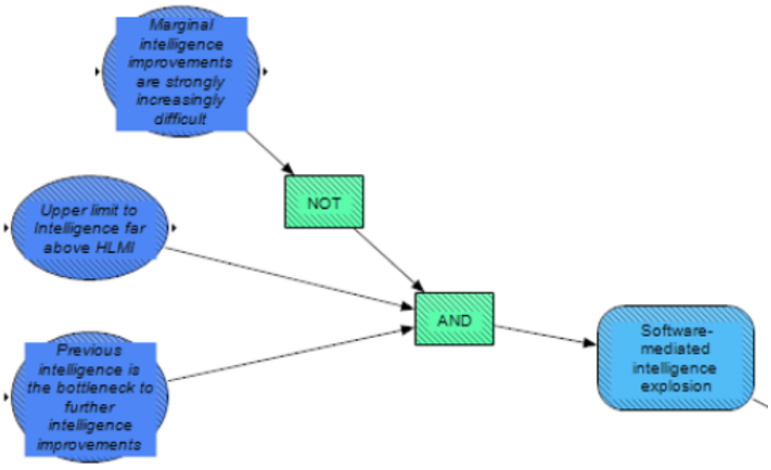
Note that the 'Few Breakthroughs' case includes situations where 0 breakthroughs are needed (and we actually don't make any new breakthroughs before HLMI) - i.e. cases where current deep learning with only minor algorithmic improvements and somewhat increased compute gives us HLMI.

Intelligence Explosion

This module aims to answer the question: will there eventually be an intelligence explosion? We define an "intelligence explosion" as a process by which HLMI successfully accelerates the rate at which HLMI hardware or software advances, to such a degree that the rate of progress approaches vertical over a large range.



In our model, whether an intelligence explosion eventually occurs does not directly depend on what type of HLMI is first developed (as we assume that if one type of HLMI could not achieve an intelligence explosion while another could, even if the first type of HLMI is achieved first, the latter type – if possible to build – will eventually be built and cause an intelligence explosion then). Our model considers two paths to an intelligence explosion – a software-mediated path and a hardware-mediated path.

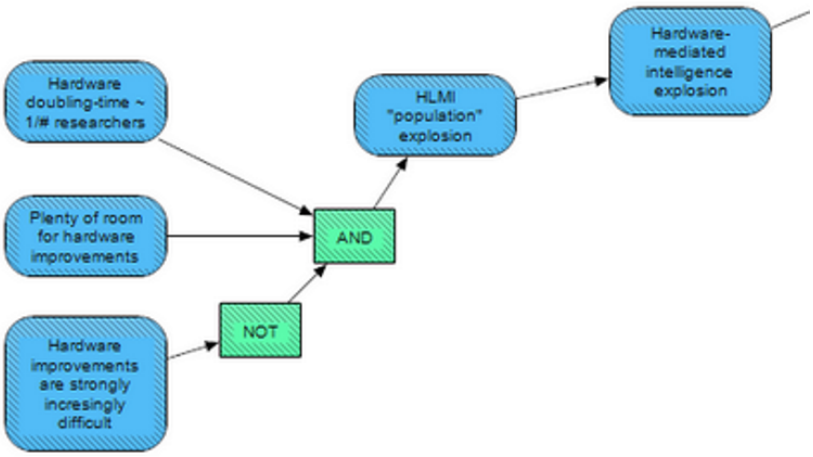


Under the software-mediated path, our model assumes there will be explosive growth in intelligence, due to AI accelerating the rate of AI progress, if HLMI is developed and:

- marginal intelligence improvements are not strongly increasingly difficult,
- there are no other theoretical limits to increasing general intelligence or the practical capabilities of a general intelligence (at least none barely above the human level), and
- further intelligence improvements are bottlenecked by previous intelligence (as opposed to, say, physical processes that cannot be sped up).

In such a scenario, the positive feedback loop from "more intelligent AI" to "better AI research (performed by the more intelligent AI)" to "even more intelligent AI still" would be explosive – in effect, HLMI could achieve sustained returns on cognitive reinvestment, at least over a sufficiently large range.

In addition to the software-mediated path, an intelligence explosion could occur due to a hardware-mediated path.



In this scenario, HLMI (doing hardware R&D) would cause an explosion in the amount of hardware, and thus an explosion in the "population" of HLMI (implying more HLMIs to perform hardware research, faster hardware gains, and so on). This phenomenon would require hardware improvements to scale with the number of hardware researchers (with this work being performed by HLMIs), for hardware improvements not to become strongly-increasingly difficult, and for there to be plenty of room for more hardware improvements. Such a pathway would allow, at least in principle, the capabilities of AI to explode even if the capability of any given AI (with a fixed amount of hardware) did not explode.

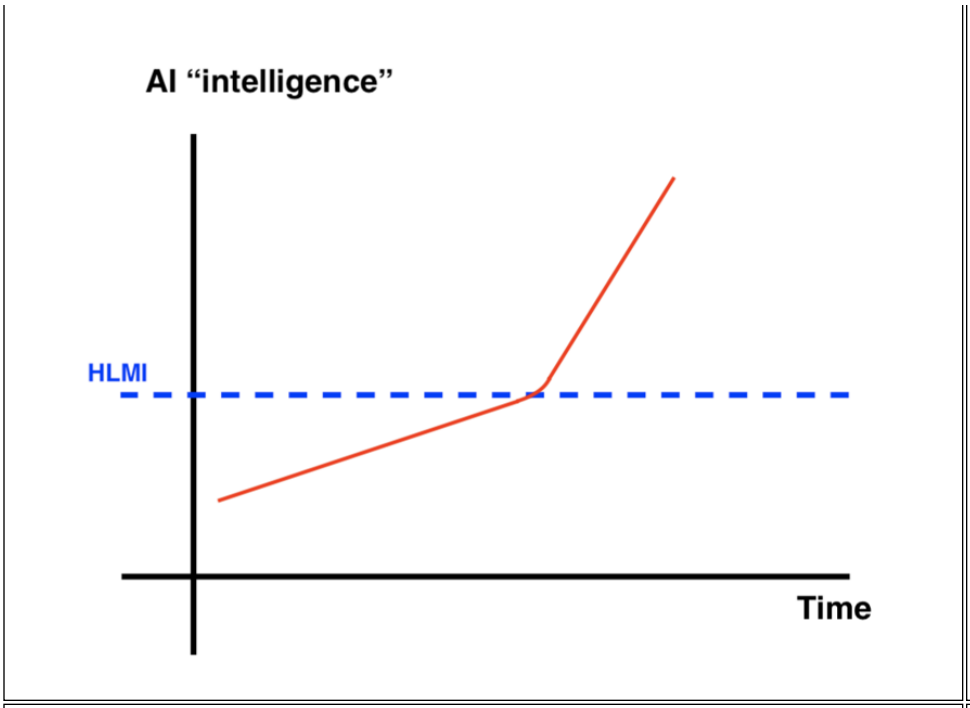
Note that *Intelligence Explosion* as defined in this model does not necessarily refer to an instantaneous switch to an intelligence explosion immediately upon reaching HLMI – an intelligence explosion could occur after a period of slower post-HLMI growth with intermediate doubling times. Questions about immediate jumps in capabilities upon reaching HLMI are handled by the *Discontinuity* module.

Comparing Discontinuity and Intelligence Explosion

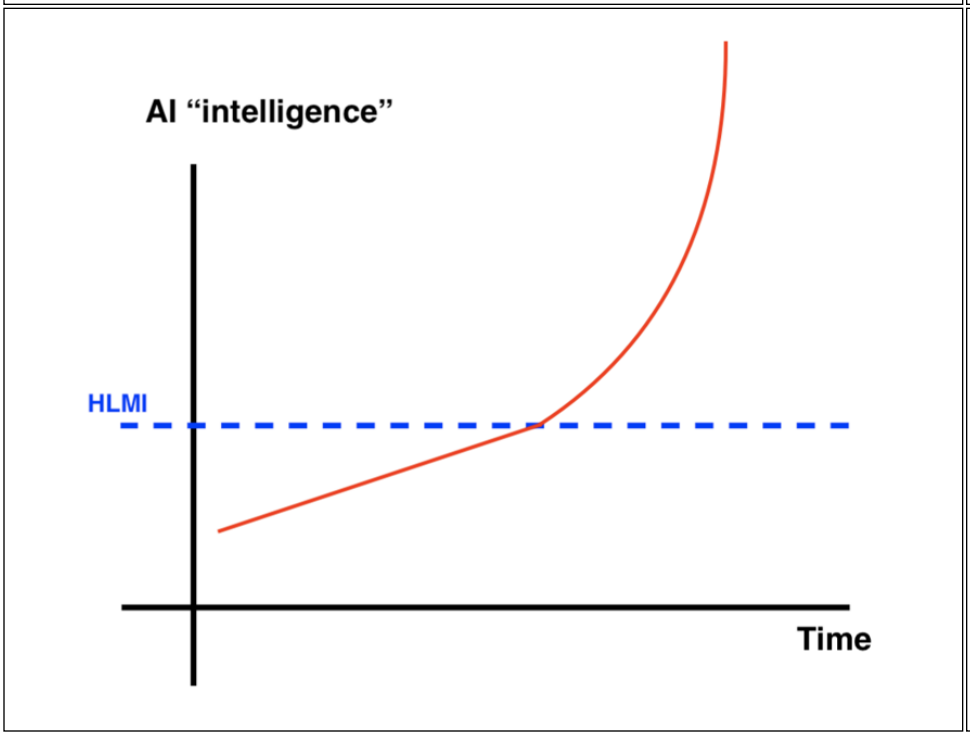
The distinction between a discontinuity and an intelligence explosion in our model can be understood from the following graphs, which show rough features of how AI capabilities might advance over time given different resolutions of these cruxes. Note that, while these graphs show the main range of views our model can express, they are not exhaustive of these views (e.g., the graphs show the discontinuity going through HLMI, while it's possible that a discontinuity would instead simply go to or from HLMI).

Additionally, the model is a simplification, and we do not mean to imply that progress will be quite as smooth as the graphs imply – we're simply modeling what we expect to be the most important and crux-y features. Take these graphs as qualitative descriptions of possible scenarios as opposed to quantitative predictions – note that the y-axis (AI "intelligence") is an inherently fuzzy concept (which perhaps is better thought of as increasing on a log scale) and that the dotted blue line for "HLMI" might not occupy a specific point as implied here but instead a rough range. Further remember that we're not talking about economic growth, but AI capabilities (which feed into economic growth).

No Intelligence Explosion, No Discontinuity	**No Intelligence Expl**

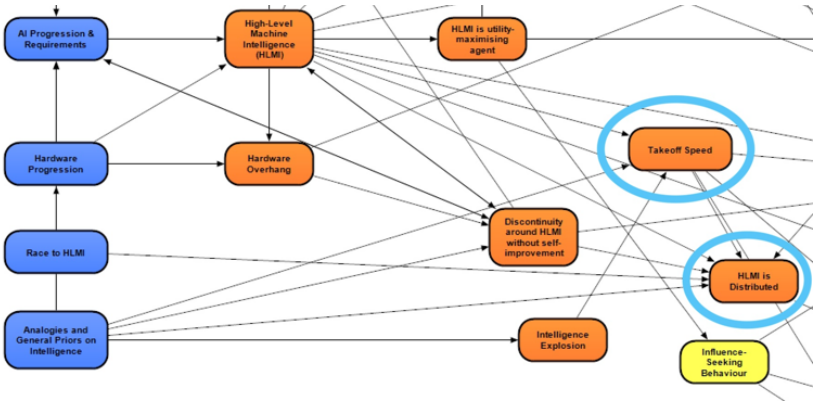


Intelligence Explosion, No Discontinuity



Later Modules (Takeoff Speed and HLMI is Distributed)

The earlier modules from the previous section act as inputs to the modules in this section: *Takeoff Speed* and *HLMI is Distributed*.

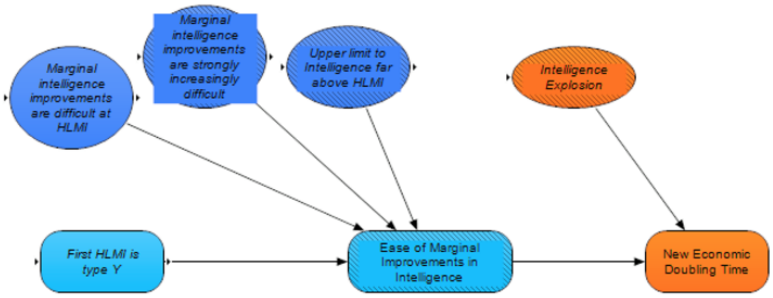


Takeoff Speed

We have seen how the model estimates the factors which are important for assessing the takeoff speed:

- Will there be an intelligence explosion?
- From *Analogies and General Priors*:
  - How difficult are marginal intelligence improvements at and beyond HLMI?
  - What is the upper limit to intelligence?

This module aims to combine these results to answer the question – what will happen to economic growth post-HLMI? Will there be a new, faster economic doubling time (or equivalent) and if so, how fast will it be? Alternatively, will growth be roughly hyperbolic (before running into physical limits)? To clarify, while previously we were considering changes in AI capabilities, here we are examining the resultant effects for economic growth (or similar). This discussion is not per se premised on considerations of GDP measurement.



If the *Intelligence Explosion* module indicates an intelligence explosion, then we assume economic growth becomes roughly hyperbolic, along with AI capabilities (i.e., increasingly short economic doubling times).

If there is not an intelligence explosion, however, then we assume that there will be a switch to a new mode of exponential economic growth, and we estimate the speed-up factor for this new growth rate compared to the current growth rate based largely on outside-view estimates of previous transitions in growth rates (the Agricultural Revolution and the Industrial Revolution – on this outside view alone, we would expect the next transition to bring us an economic doubling time of perhaps days to weeks). This estimate is then updated based on an assessment of the overall ease of marginal improvements post-HLMI.

If

- marginal intelligence improvements are not strongly increasingly difficult,
- there is no significant upper limit to HLMI capability, and
- marginal intelligence improvements are not difficult at HLMI,

then we conclude that the transition to more-powerful HLMI looks faster, all else equal, and update our outside-view estimate regarding the economic impact accordingly (and similarly we update towards slower growth if these conditions do not apply). We plan to use these considerations to create a lognormally-distributed estimate of the final growth rate, given that we are uncertain over multiple orders of magnitude regarding the post-HLMI growth rate, even in a world without an intelligence explosion.

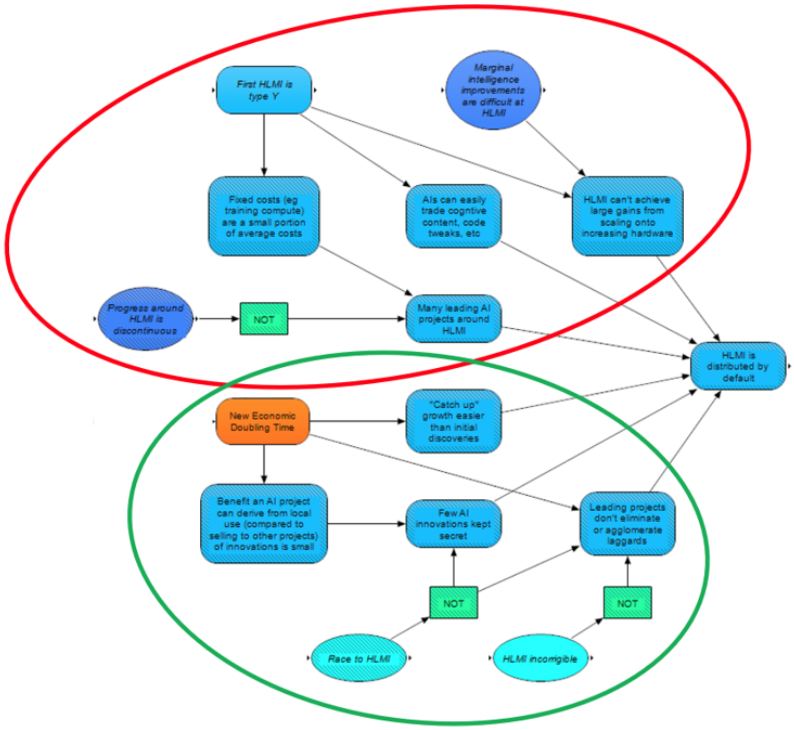
The connection between economic doubling time and the overall intelligence/capability of HLMI is not precise. We think our fuzzy assessment is appropriate, however, since we're only looking for a ballpark estimate here (and due to the lognormal uncertainty, the results of our model should be robust to small differences in these parameters).

HLMI is Distributed

This module aims to answer the question, is HLMI 'distributed by default'? That is, do we expect (ignoring the possibility of a Manhattan Project-style endeavour that concentrates most of the world's initial research effort) to see HLMI capability distributed throughout the world, or highly localized into one or a few leading projects?

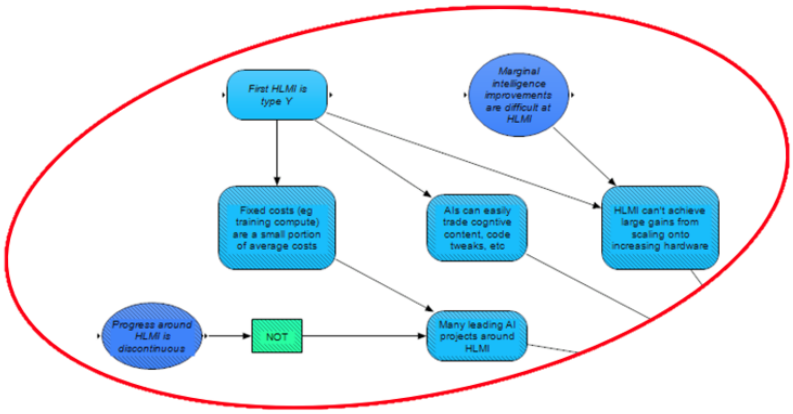
Later in this sequence of posts we will synthesise predictions about the two routes to highly localised HLMI: the route explored in this post i.e., HLMI not being distributed by default; and an alternative route, explored in a later module, where most of the world's research effort is concentrated into one project. We expect that if HLMI is distributed by default **and** research effort is not strongly concentrated into a few projects, many powerful HLMIs will be around at the same time.

Several considerations in this section are taken from Intelligence Explosion Microeconomics (in section 3.9 "Local versus Distributed Intelligence Explosions").



Arguments about the degree to which HLMI will be distributed by default can be further broken up into two main categories: those heavily influenced by the economic takeoff speed/possibility of an intelligence explosion (mostly social factors, circled in green); and those not heavily influenced by the economic takeoff speed (mostly technical factors, circled in red). We should note that, while takeoff speed indirectly affects the likelihood of HLMI distribution through intermediate factors, it does not *directly* affect whether HLMI will be distributed; even in the case of an intelligence (and therefore economic) explosion, it's still possible that progress could accelerate uniformly, such that no single project has a chance to pull ahead.

Here, we will first examine the factors not tied to takeoff speed, before turning to the ones that are.



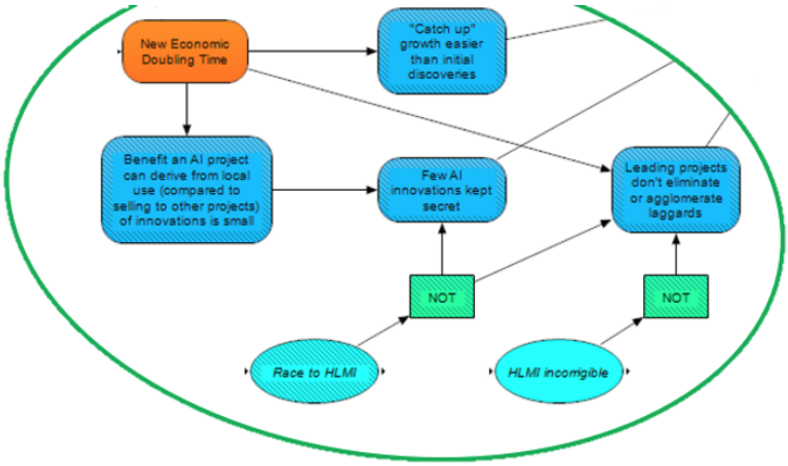
A significant consideration is whether there will be a **discontinuity in AI capabilities around HLMI**. If there is a discontinuity, then it is highly likely that HLMI will not initially be distributed by default, because one project will presumably reach the discontinuity first. We model this as there being only one leading AI project to begin with. Even if progress around HLMI is

continuous, however, there could still be only a few leading projects going into HLMI, especially if **fixed costs are a large portion of the total costs for HLMI** (presumably affected by the kind of HLMI), since high fixed costs may present a barrier from many competitor projects.

HLMI is also more likely to be distributed if **AIs can easily trade cognitive content, code tweaks, and so on** (this likelihood is also presumably influenced by the type of HLMI), as if so, the advantages that leading projects hold may be more likely to be distributed to other projects.

Finally, if **HLMI can achieve large gains from scaling onto increasing hardware**, then we might expect leading projects to increase their leads over competitors, as profits could be reinvested in more hardware (or compute may be seized by other means), and thus HLMI may be expected to be less distributed. We consider that the likelihood of large gains from further hardware is dependent on both the type of HLMI, and the difficulty of marginal improvements in intelligence around HLMI (with lower difficulty implying a greater chance of large gains from increasing hardware).

Then, there are the aforementioned factors which are **heavily influenced by the takeoff speed** (which is influenced by whether there will be an intelligence explosion):



If **catch-up innovation based on imitating successful HLMI projects is easier than discovering methods of improving AI in the first place**, then we would expect more distribution of HLMI, as laggards may successfully play catch-up. A faster doubling time – and in particular an intelligence explosion – may push against this, as projects that begin to gather a lead may "pull ahead" more easily than others can play catch-up (we expect a faster takeoff to accelerate cutting-edge AI development by more than it accelerates the rest of the economy).

If **major AI innovations tend to be kept secret**, then this also pushes against HLMI being distributed. We may consider that a race to HLMI may encourage more secrecy between competitors. Additionally, secrecy may be more likely if AI projects can derive larger benefits from using its innovations locally than from selling its innovations to other projects. Local use may be more likely larger if there are shorter economic doubling times/an intelligence explosion, as such scenarios imply large returns from cognitive reinvestment.

Finally, we consider that distributed HLMI is less likely if **leading projects eliminate or agglomerate laggards**. Again, a race dynamic probably makes this scenario more likely. Additionally, if HLMI is incorrigible, it might be more likely to "psychopathically" eliminate laggards via actions that projects with corrigible HLMI might opt to avoid.

Conclusion

To summarize, we have examined key questions related to AI takeoff: whether there will be a discontinuity in AI capabilities to and/or from HLMI, whether there will be an intelligence explosion due to feedback loops post-HLMI, the growth rate of the global economy post-AI takeoff, and whether HLMI will be distributed by default. These estimates use a mixture of inside-view and outside-view considerations.

In building our model, we have made several assumptions and simplifications, as we're only attempting to model the main cruxes. Naturally, this does not leave space for every possible iteration on how the future of AI might play out.

In the next post in this series, we will discuss risks from mesa-optimization.

1. We define HLMI as machines that are capable, either individually or collectively, of performing almost all economically-relevant information-processing tasks that are performed by humans, or quickly (relative to humans) learning to perform such tasks. We are using the term "high-level machine intelligence" here instead of the related terms "human-level machine intelligence", "artificial general intelligence", or "transformative AI", since these other terms are often seen as baking in assumptions about either the nature of intelligence or advanced AI that are not universally accepted.

Acknowledgments

This post was edited by Issa Rice. We would like to thank both the rest of the MTAIR project team, as well as the following individuals, for valuable feedback on this post: Ozzie Gooen, Daniel Kokotaljo and Rohin Shah

Modeling Risks From Learned Optimization

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

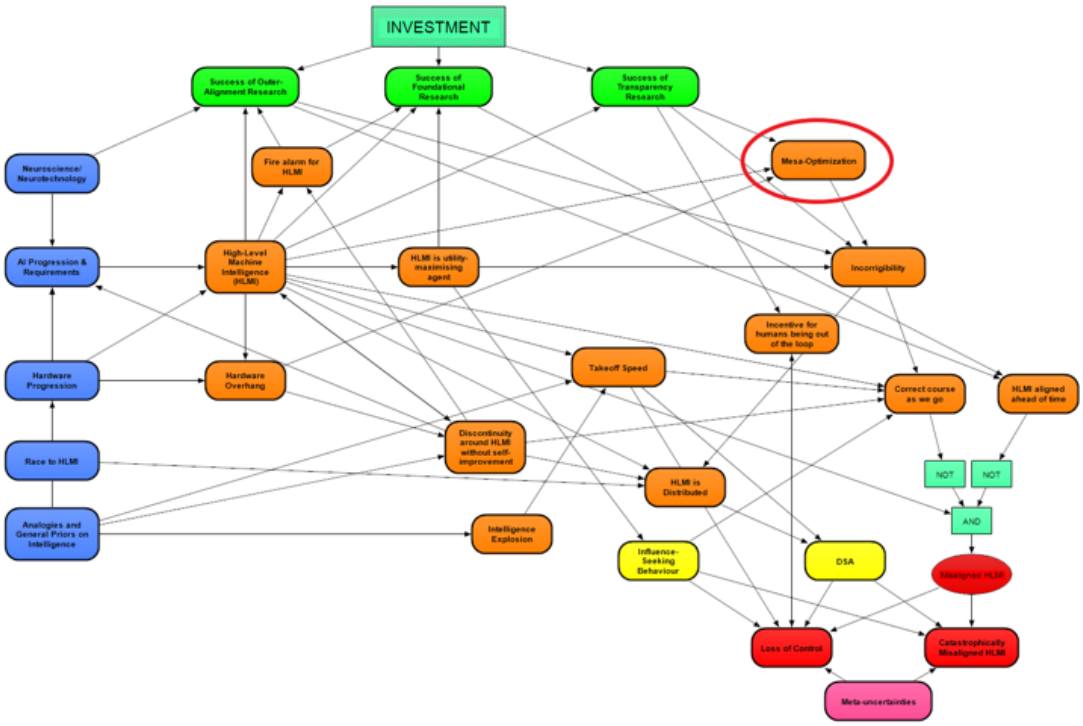
This post, which deals with how risks from learned optimization and inner alignment can be understood, is part 5 in our sequence on Modeling Transformative AI Risk . We are building a model to understand debates around existential risks from advanced AI. The model is made with Analytica software, and consists of nodes (representing key hypotheses and cruxes) and edges (representing the relationships between these cruxes), with final output corresponding to the likelihood of various potential failure scenarios. You can read more about the motivation for our project and how the model works in the Introduction post . The previous post in the sequence, Takeoff Speeds and Discontinuities , described the different potential characteristics of a transition from high-level machine intelligence [1] to superintelligent AI.

We are interested in feedback on this post, especially in places where the model does not capture your views or fails to include an uncertainty that you think could be an important crux. Similarly, if an explanation seems confused or confusing, flagging this is useful – both to help us clarify, and to ensure it doesn't reflect an actual disagreement.

This post explains how risks from learned optimization are incorporated in our model. The relevant part of the model is mostly based on the Risks from Learned Optimization sequence and paper (henceforth RLO). Although we considered responses and alternate perspectives to RLO in our research, these perspectives are not currently modeled explicitly.

For those not familiar with the topic, a *mesa-optimizer* is a learned algorithm that is itself an optimizer. According to RLO, inner alignment is the problem of aligning the objective of a mesa-optimizer with the objective of its *base optimizer* (which may be specified by the programmer). A contrived example supposes we want an algorithm that finds the shortest path through any maze. In the training data, all mazes have doors that are red, including the exit. Inner misalignment arises if we get an algorithm that efficiently searches for the next red door – the capabilities are robust because the search algorithm is general and efficient, but the objective is not robust because it finds red doors rather than the exit.

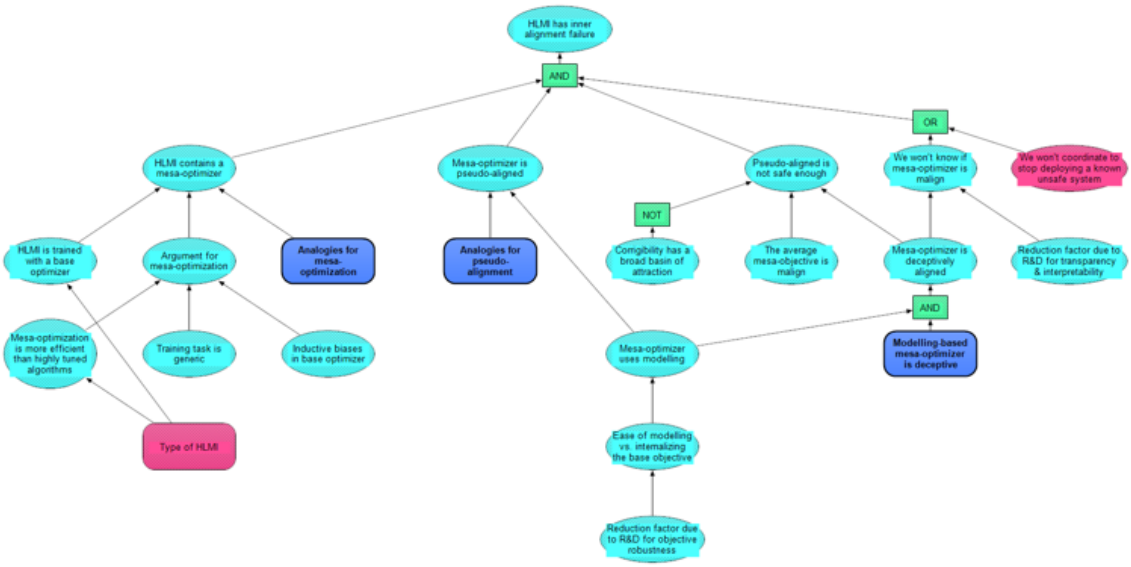
The relevant part of our model is contained in the **Mesa-Optimization** module:



Right-click and select "open image in new tab" (or similar) to see images full-size

The output of the **Mesa-Optimization** module is an input to the **Incorrigibility** module. The logic is that inner misalignment is one way that high-level machine intelligence (HLMI) could become [incorrigible](), which in turn counts strongly against being able to **Correct course as we go** in the development of HLMI.

Module Overview



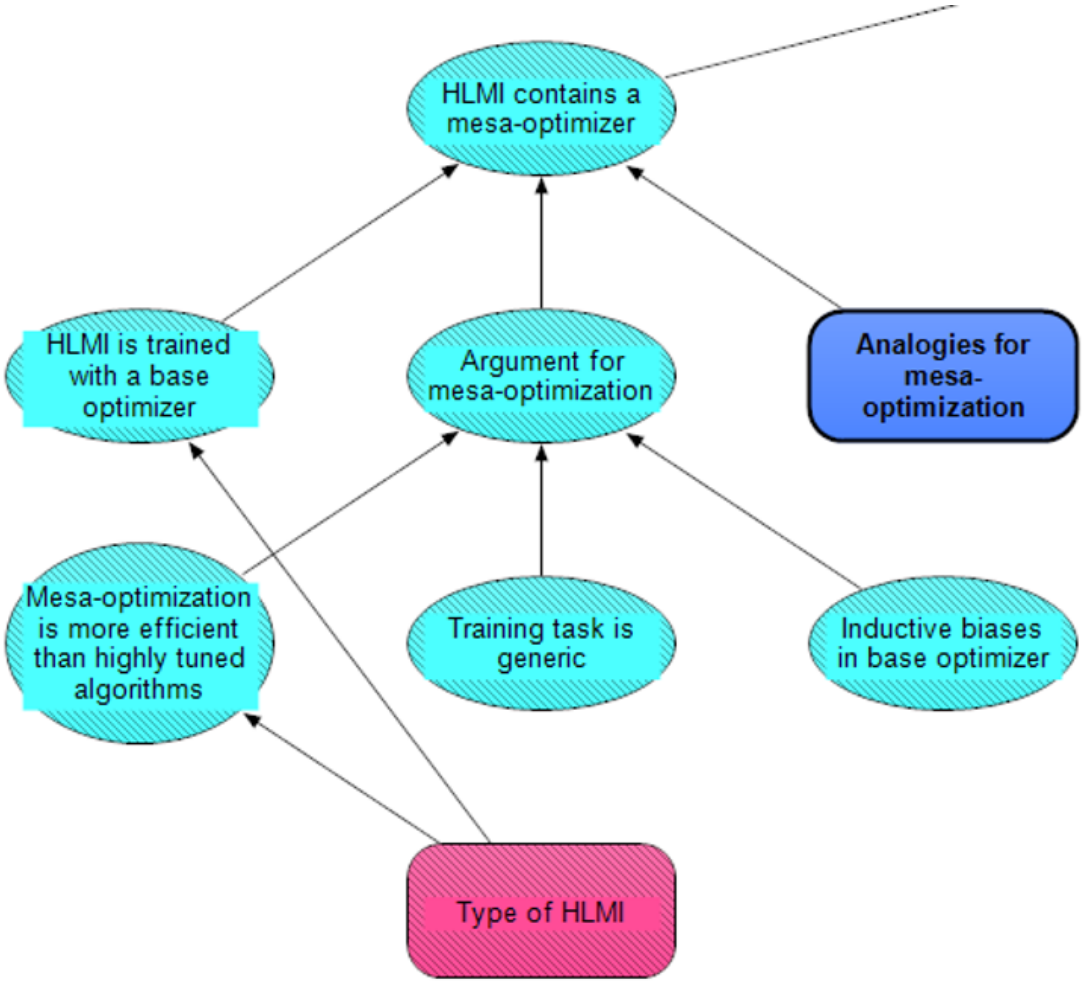
Overview of the Mesa-Optimization module. We recommend reading top-to-bottom, left-to-right.

The top-level logic of the **Mesa-Optimization** module is: HLMI has an inner alignment failure if

1. The HLMI contains a mesa-optimizer, AND
2. Given (1), the mesa-optimizer is pseudo-aligned, i.e. it acts aligned in the training setting but its objective is not robust to other settings, AND
3. Given (2), the pseudo-alignment is not sufficient for intent alignment, i.e. is not safe enough to make HLMI corrigible, AND
4. Given (3), we fail to stop deployment of the unsafe system.

In the following sections, we explain how each of these steps are modeled.

HLMI contains a mesa-optimizer



The output of this section is **HLMI contains a mesa-optimizer**, which depends on three nodes. The left node, **HLMI is trained with a base optimizer**, means that a training algorithm optimized a distinct learned algorithm, and that learned algorithm forms all or part of the HLMI system. A crux here is what **Type of HLMI** you expect, which comes out of the pathways discussed in the Paths to HLMI post. For instance, HLMI via current deep learning methods or evolutionary methods will involve a base optimizer, but this is not true of other pathways such as whole-brain emulation.

The middle node, **Argument for mesa-optimization**, represents the argument from first principles for why mesa-optimization would occur in HLMI. It is mainly based on the post Conditions for Mesa-Optimization. This is broken down further into three nodes. **Mesa-optimization is more efficient than highly tuned algorithms** distils some claims about the advantages of mesa-optimization compared to systems without mesa-optimization, including that it offers better generalisation through search. You could reject that claim on the grounds that sample efficiency will be high enough, such that HLMI consists of a bunch of algorithms that are highly tuned to different domains. You could also argue that some **types of HLMI** are more prone to be highly tuned than others. For instance, evolutionary algorithms might be more likely to mesa-optimize than machine learning, and machine learning might be more likely to mesa-optimize than hybrid ML-symbolic algorithms.

The middle node, **Training task is generic**, resolves as positive if the learned algorithm is not directly optimized for domain-specific tasks. This scenario is characteristic of pre-training in modern machine learning. For example, the task of predicting the next word in a large text dataset is generic, because the dataset could contain all kinds of content that is relevant to many different domain-specific tasks. In contrast, one could use a more domain-specific dataset (e.g. worded math problems) or objective function (e.g. reward for the quality of a news article summary). This crux is important if mesa-optimizers generalise better than other kinds of algorithms, because then a generic training task would tend to select for mesa-optimizers more. GPT-3 gives credence to this idea, because it demonstrates strong few-shot learning performance by simply learning to predict the next word in a text. However, it's uncertain if GPT-3 meets the definition of a mesa-optimizer.

Finally, **Inductive biases in base optimizer** lumps together some factors about inductive bias in the HLMI's architecture and training algorithm which affect the chance of mesa-optimization. For example, the extent that mesa-optimization exists in the space of possible models (i.e. algorithmic range), and the ease by which the base optimizer finds a mesa-optimizer (i.e. reachability). Inductive bias is a big factor in some people's beliefs about inner alignment, but there is disagreement about just how important it is and how it works (see e.g. Inductive biases stick around, including comments).

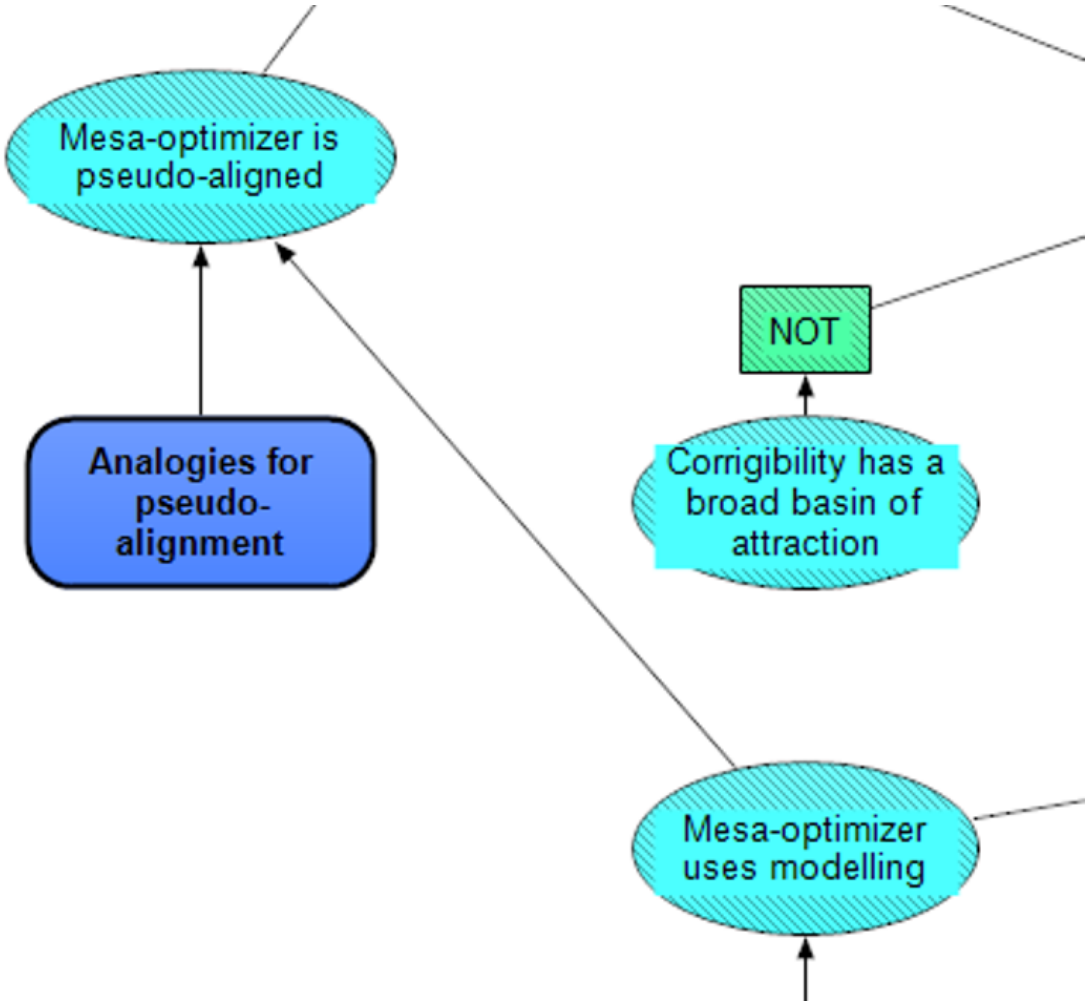
Analogies for mesa-optimization

Finally, evidence for the node **HLMI contains a mesa-optimizer** is also drawn from history and analogies. We modeled this evidence in a submodule, shown below. The submodule is structured as a Naive Bayes classifier: it models the likelihood of the evidence, given the hypotheses that an HLMI system does or does not contain a mesa-optimizer. The likelihood updates the following prior: if you only knew the definition of mesa-optimizer, and hadn't considered any specific cases, arguments or evidence for it, what is the probability of an optimizing system containing a mesa-optimizer?

We considered three domains for the evidence: machine learning today, firms with respect to economies, and humans with respect to natural selection. A few more specific nodes are included under **Examples from ML today** because there has been some interesting discussion and disagreement in this area:

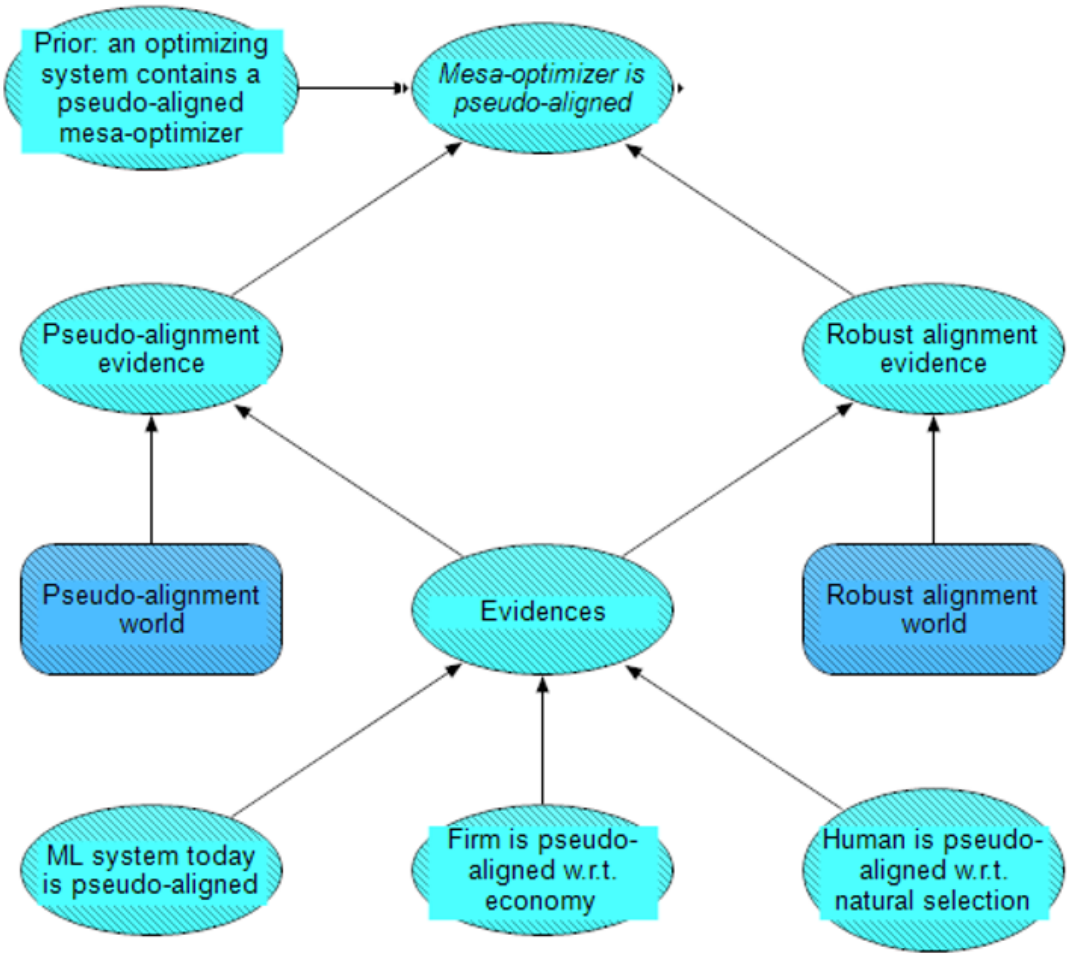
- The "spontaneous emergence of learning algorithms" from reinforcement learning algorithms has been cited as evidence for mesa-optimization, but this may not be informative.
- Meta-learning is a popular area of ML research, but since it applies a deliberate and outer-loop optimization, it doesn't clearly affect the likelihood of spontaneous mesa-optimization.
- GPT-3 does few-shot learning in order to perform novel tasks. This post gives an argument for why it is incentivised for – or may already be – mesa-optimizing.
- It is unclear and debated how much AlphaZero marginally benefits from Monte Carlo Tree Search, which is a form of mechanistic optimization, compared to just increasing the model size. In turn it is unclear how much evidence AlphaZero provides for getting better generalisation through search, which is argued as an advantage of mesa-optimization.

The mesa-optimizer is pseudo-aligned



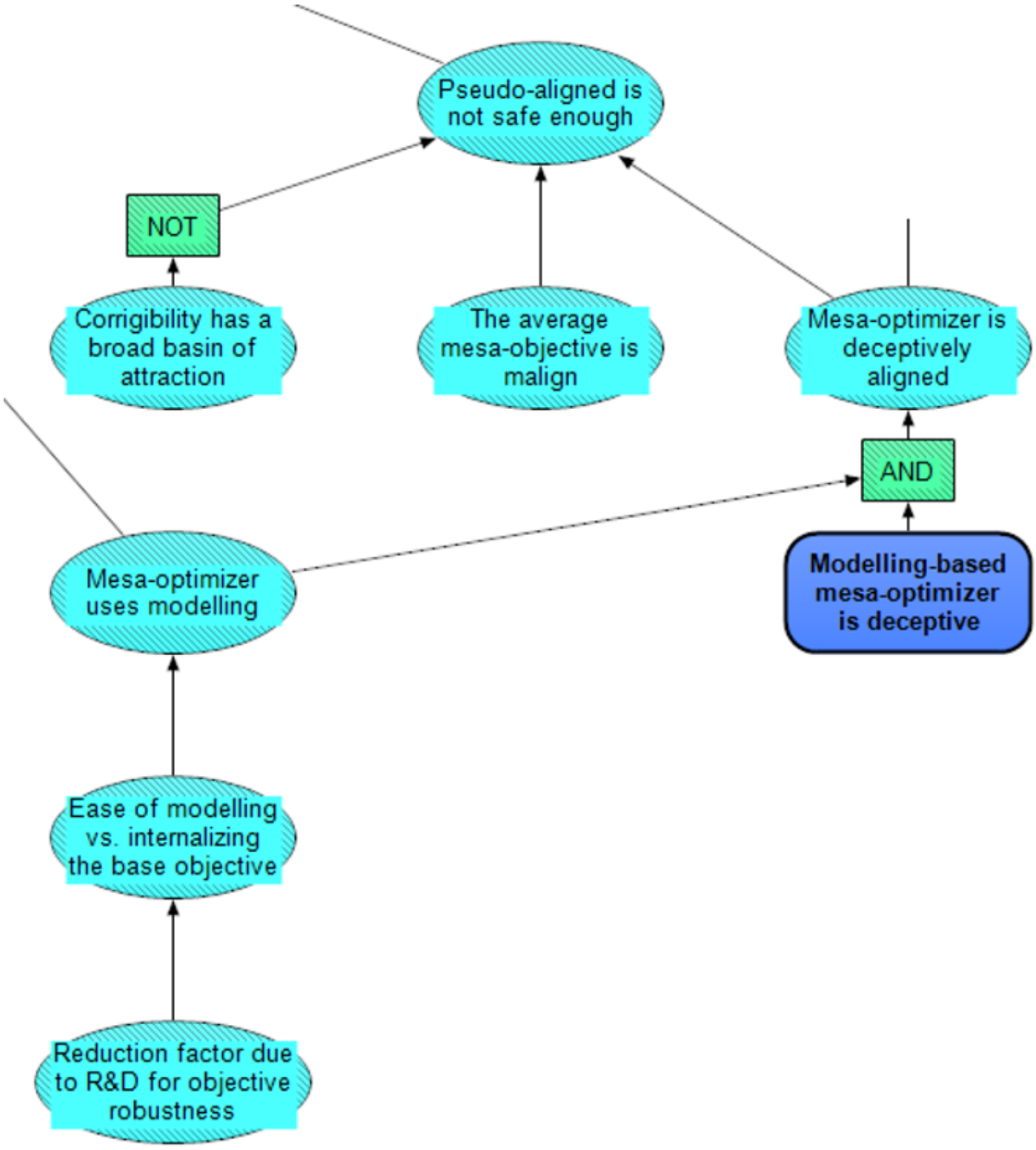
The possibility of pseudo-alignment depends on just two nodes in our model. The structure is simple mainly because pseudo-alignment of any kind seems much more likely than robust alignment, so we haven't noticed much debate on this point. In general there are many more ways to perform well on the training distribution which are not robustly aligned with the objective, i.e. they would perform significantly worse on that objective under some realistic shift in the distribution. And in practice today, this kind of robustness is a major challenge in ML (Concrete Problems in AI Safety section 7 gives an overview, though it was published back in 2016).

The dependency on the left is a module, **Analogies for pseudo-alignment**, which is structured identically to **Analogies for mesa-optimization** (i.e. with a Naive Bayes classifier, and so on), but the competing hypotheses are pseudo-alignment and robust alignment, and the analogies are simply "ML systems today", "Firm", and "Human".s



The second node influencing pseudo-alignment is **Mesa-optimizer uses modeling**. The concept of "modeling" vs. "internalization" introduced in the RLO paper (section 4.4) is relevant to pseudo-alignment. Internalization implies robust alignment, whereas modeling means the mesa-optimizer is pointing to something in its input data and/or its model of the world in order to act aligned. We explain this node and the implications of "modeling" in more detail in the next section.

Pseudo-alignment is not safe enough



At the top level of this subsection, we include three reasons why pseudo-alignment might not be safe enough to count as a failure of inner alignment. Firstly, there is a crux of whether **corrigibility has a broad basin of attraction**. This refers to Paul Christiano's claim that "A sufficiently corrigible agent will tend to become more corrigible and benign over time. Corrigibility marks out a broad basin of attraction towards acceptable outcomes." If Christiano's claim about corrigibility is true, this increases the overall chance that a pseudo-aligned algorithm becomes safe enough before it locks in a path to catastrophe.

A second crux for the safety of pseudo-alignment is **how malign we expect a mesa-objective to be by default** (by malign, we just mean non-benign, or harmful *in effect* – it doesn't have to be inherently malicious). It's uncertain what mesa-objectives are generally like, because they arise internally from the base optimizer, and there is currently scant empirical evidence of mesa-optimization. It's reasonable to expect a proxy objective to be much closer to the base objective than to a random objective, so the danger partly depends

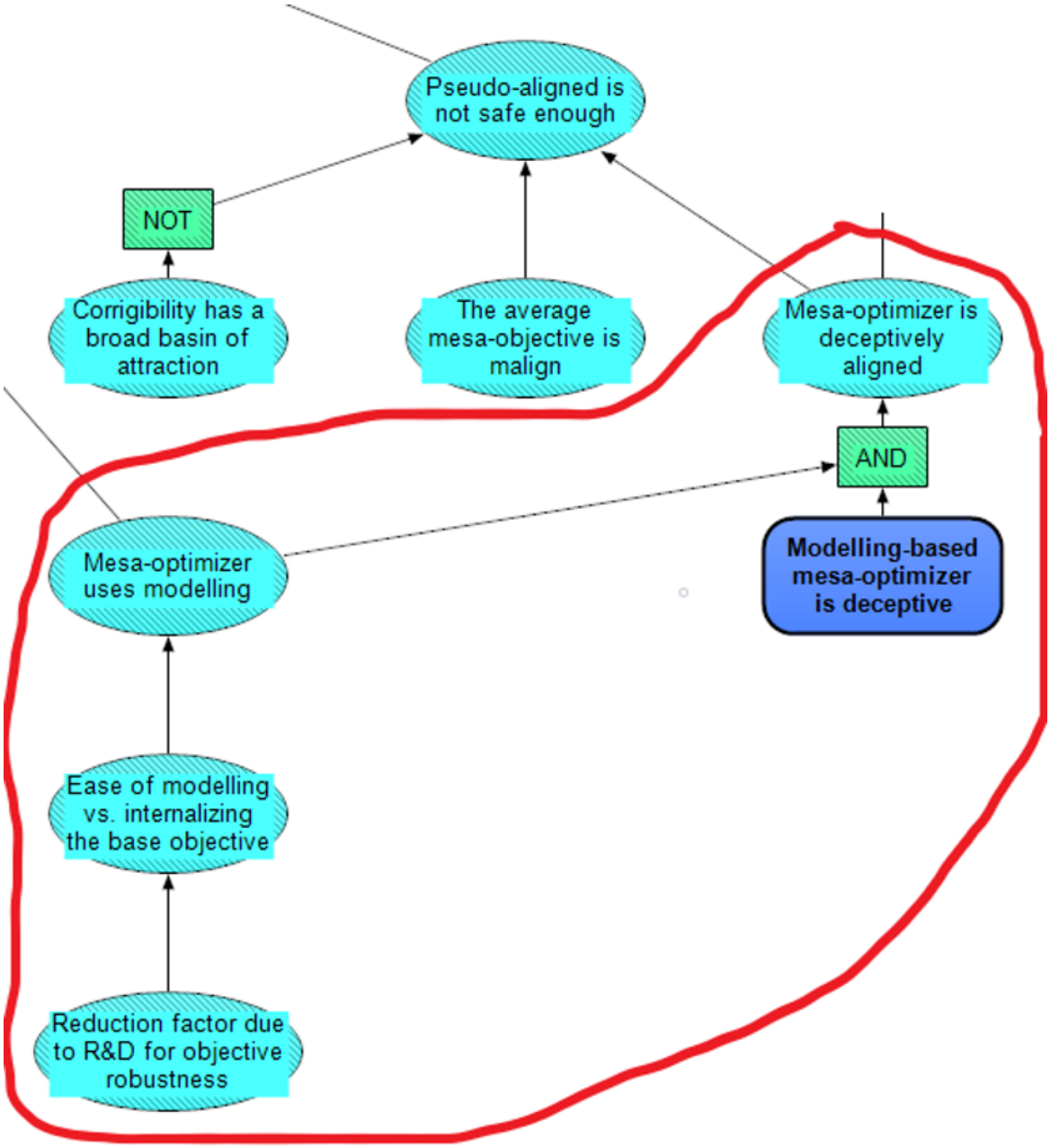
on the base objective. Perhaps mesa-objectives will just be weird in benign ways, e.g. being irrational, or very local. On the other hand, it seems that simple, unbounded, coherent objectives would tend to have lower description length than objectives lacking those attributes, so those objectives may be favoured by simplicity bias.

Before moving on to the third reason, *deceptive alignment*, let's review the concept of *modeling* vs. *internalization* introduced in the RLO paper (section 4.4). Modeling is where a mesa-optimizer acts aligned with a base objective (in training) by representing the base objective within its world model (using relevant information from its input data) and then deliberately acting aligned with this representation. By understanding the base objective, the mesa optimizer can behave more aligned, whether or not its own objective is different. (In this case, “modeling” is similar to a trial employee working hard because their mental model of the company indicates that the company wants them to work hard.) The alternative manner to modeling in which a mesa-optimizer can act aligned with a base objective is *internalization:* the base optimizer adjusts the mesa-objective to be more and more aligned until convergence. (In this case, internalization is similar to a trial employee working hard because they intrinsically want to work hard.)

Modeling-based mesa-optimizers are thought to end up *deceptively* or *corrigibly* aligned as a result of training. Deceptive alignment means the learned algorithm optimizes for the base objective during training for instrumental reasons (namely, to deceive the learning algorithm or programmers into thinking it’s aligned). Returning to the example of our trial employee, a deceptively aligned trial employee might just want a salary and to slack off, so he will work hard while in trial to be offered the job, and then after being offered the job only work the minimal amount to not get fired. So a deceptively aligned algorithm deviates when it becomes favourable, e.g. it figures out that training has ended, removing the threat of being penalised.

The main alternative to a deceptive mesa-optimizer that we consider is a *corrigible* mesa-optimizer. A corrigible mesa-optimizer optimizes for a *pointer* to the part of its world model that includes the base objective (both in training and deployment). The corrigibility comes from the fact that the base optimizer can modify this pointer to be more accurate. Here, our trial employee cares about being "the perfect employee" both in trial and once offered a job, but they are uncertain how to measure that. So they might observe employees that have succeeded (the pointer) and learn to imitate that. Note that this is subtly different from the internalizing employee. The internalizing employee doesn't care about the company in itself, but values working hard for its own sake (and thus happens to be aligned with the company), whereas the corrigibly modeling employee wants to do whatever they believe the company wants of them, which is to work hard.

Coming back to our model, the third reason that pseudo-alignment might not be safe enough is if the **Mesa-optimizer is deceptively aligned**.



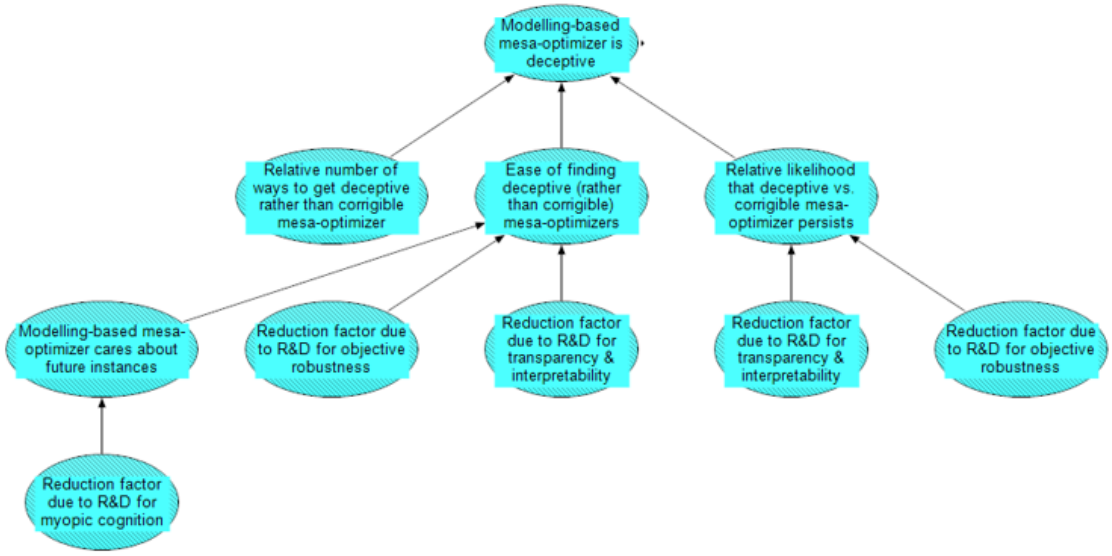
We broke down deceptive alignment into two main conditions. First, a precondition that the **mesa-optimizer is using modeling**. The key question influencing this precondition is **how easy it is for a mesa-optimizer to model rather than internalize**. We represent this as the likelihood ratio of eventually settling on modeling versus internalization. One argument for modeling being easier is that, within the training data for the HLMI, there will likely be a huge amount of rich input data relevant for determining the base objective, and it will therefore be easier for the HLMI to capture this objective function by referring to the relevant parts of this data, than it will be to build a representation of the objective from scratch. However, this argument depends on just how difficult it is to construct representations that are useful for modeling the objective, which in turn depends both on the nature of the data and the objective. We can also look to analogies here, such as imprinting in animals. For example, rather than have a detailed representation of a mother goose, a gosling imprints on some initial stimulus in its environment, whether that be an actual goose or, say, a

human's boots, and then treats the imprinted object as its mother (e.g. it follows the object around).

Feeding into the **ease of modeling vs. internalization** node is **Reduction factor due to R&D for objective robustness** (*objective robustness* is terminology from Clarifying inner alignment terminology). This node quantifies the effect that safety research has (as a constant multiplier) on reducing the **Ease of modeling vs. internalizing the base objective**, assuming that modeling is dangerous and better to avoid.

The second condition for deceptive alignment is simply: given that a mesa-optimizer uses modeling, it is deceptive. This possibility is broken down into a submodule of nodes, explained in the next section.

Deceptive alignment



The inside of the "Modeling-based mesa-optimizer is deceptive" submodule.

There are three factors feeding into the **modeling-based mesa-optimizer is deceptive** node. First is **Relative number of ways to get a deceptive rather than corrigible mesa-optimizer**. Think of this as a base rate for deceptive (rather than corrigle) mesa-optimization. There is a counting argument for deception here: since deception is instrumentally motivated, it is compatible with almost any proxy goal, whereas for corrigibility, the base optimizer has to find a very specific pointer to relate to the base objective.

The second factor is the **Ease of finding deceptive (rather than corrigible) mesa-optimizers**. Similar to *reachability*, this node is represented by the relative likelihood that the base optimizer outputs a deceptive mesa-optimizer at some point in training.

The post Does SGD Produce Deceptive Alignment? details some arguments for and against deceptive algorithms being easier to find than corrigible alignment (specifically for stochastic gradient descent, but much of it is more general). Becoming deceptive requires caring about future instances, so that there is an incentive to deceive now for later gain. Becoming deceptive also requires reasoning about the training process, so that the mesa-optimizer expects the threat of modification to go away and for deception to pay off. On the other hand, becoming corrigible requires a *robust* pointer to a good model of the base objective – otherwise the mesa-optimizer will just optimize for whatever proxy it has. Furthermore, even

if there is information related to the base objective in the input data, the mesa-optimizer will need to have uncertainty about precisely what objective it is pointing to. The conditions for deception are claimed to be easier [2]. However, there seems to be a lot of uncertainty here which might hinge on the nature of the base objective and how the base optimizer works.

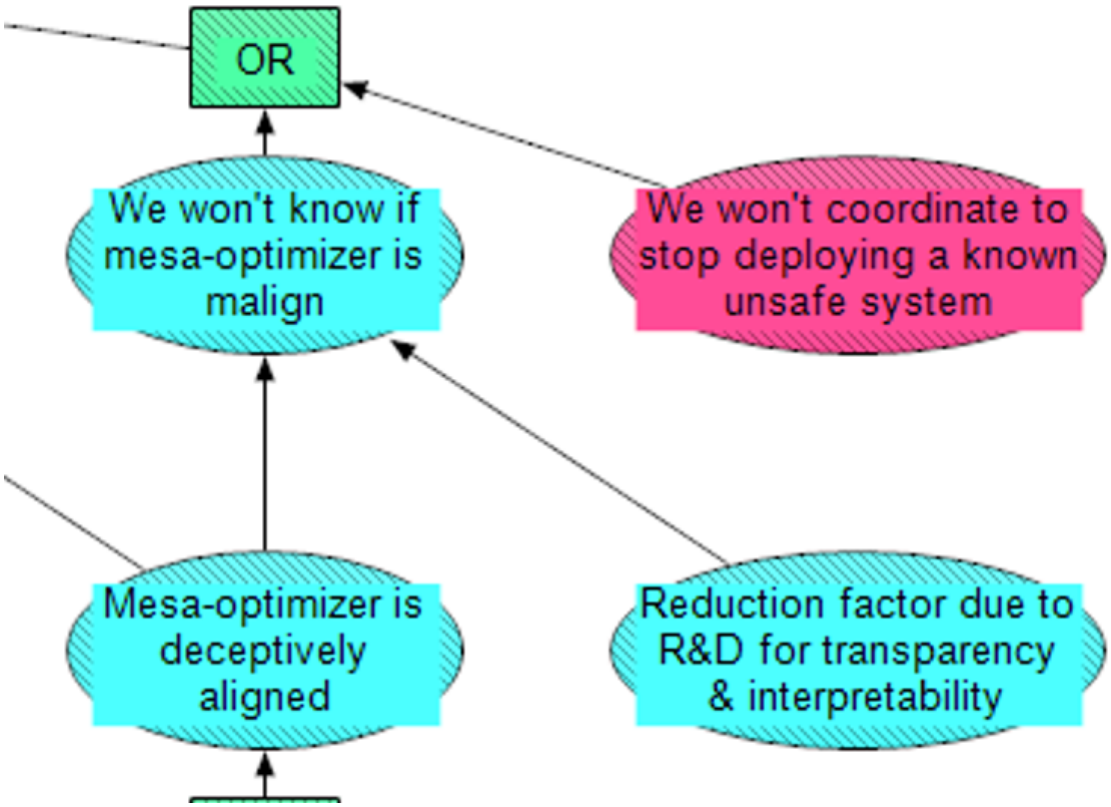
Going one level down from the **Ease of finding deceptive (rather than corrigible) mesa-optimizers** node, we have **Modeling-based mesa-optimizer cares about future instances**. This means the mesa-optimizer is motivated by what happens in future instances of training or in deployment (instead of just its current round of training). Caring about future instances is a necessary condition for deceptive alignment, but not corrigible alignment – if an algorithm does not consider future instances, then it will not see the benefit of defecting later in time, and therefore has no instrumental reason to deceive. The node for **Reduction factor due to R&D for myopic cognition** refers to the idea that methods to induce myopic cognition, such as approval-directed agents, could reduce the chance of modeling-based mesa-optimizers caring about future instances (myopic cognition is discussed in this post).

The third top-level node here is **Relative likelihood that deceptive vs. corrigible mesa-optimizer persists** – that is, the likelihood that once a mesa-optimizer becomes deceptive in one instance, it survives and remains deceptive indefinitely (relative to the likelihood for a corrigible mesa-optimizer). Deceptive algorithms may be favoured due to their ability and incentive to adapt to the base objective, because they can reason about the objective and they care about future instances. However, protecting a proxy objective from modification may be difficult. This too seems like an important uncertainty that depends on empirical facts about the base optimizer, e.g. how stochastic gradient descent works with modern neural nets.

Finally, we include two areas of research that could reduce the likelihood of deceptive alignment. One is research to improve objective robustness, such as Relaxed adversarial training for inner alignment, where one of the aims is avoiding deception. A key part of making relaxed adversarial training work is transparency tools to guide the training process. This comment argues why that is more helpful than inspecting the algorithm *after* deception may have already occurred. Research into transparency may help to prevent deceptive alignment in other ways, so this is kept as a separate node.

We fail to stop deployment

The last part of the **Mesa-Optimization** module is about whether the overseers of an HLMI project will pull back before it's too late, given there is an HLMI with an unsafe mesa-optimizer that hasn't been deployed yet (or trained to sufficiently advanced capabilities to "break out"). The way they could pull back in such a situation is if they are aware that it's unsafe, AND they coordinate to stop the threat. That logic is inverted to an OR in the module, to fit with the top-level output of **HLMI has inner alignment failure.**



Mesa-optimizer is deceptively aligned is one of the strongest factors in knowing whether a mesa-optimizer is unsafely pseudo-aligned, because deception works against overseers figuring this out. Meanwhile, **R&D for transparency & interpretability** may help to detect unsafe pseudo-aligned algorithms. On the other hand, deceptive algorithms may just exploit weaknesses in the transparency tools, so it is even possible for the reduction factor to be less than 1 (i.e. it *increases* the chance that we fail to see the danger).

Conclusion

In this post, we have examined important cruxes and uncertainties about risks from learned optimization, and how they relate to each other. Our model is mostly based on the Risks from Learned Optimization sequence, and considers whether mesa-optimization will occur in HLMI at all, whether pseudo-alignment occurs and is dangerous, and whether the mesa-optimizer will be deployed or "break out" of a controlled environment. Some uncertainties we have identified relate to the nature of HLMI, connecting back to Paths to HLMI and to analogies with other domains. Other uncertainties relate to the training task, e.g. how generic the task is, and whether the input data is large and rich enough to incentivise modeling the objective. Other uncertainties are broadly related to inductive bias, e.g. whether the base optimizer tends to produce mesa-optimizers with harmful objectives.

We are interested in any feedback you might have, including how this post affected your understanding of the topic and the uncertainties involved, your opinions about the uncertainties, and important points that our model may not capture.

The next post in this series will look at the effects of AI Safety research agendas.

[1] We define HLMI as machines that are capable, either individually or collectively, of performing almost all economically-relevant information-processing tasks that are performed by humans, or quickly (relative to humans) learning to perform such tasks. We are using the term “high-level machine intelligence” here instead of the related terms “human-level machine intelligence”, “artificial general intelligence”, or “transformative AI”, since these other terms are often seen as baking in assumptions about either the nature of intelligence or advanced AI that are not universally accepted.

[2] Parts of this argument are more spelled out in the FLI podcast with Evan Hubinger (search the transcript for "which is deception versus corrigibility").

Acknowledgements

Thanks to the rest of the MTAIR Project team, as well as the following individuals, for valuable feedback and discussion that contributed to this post: Evan Hubinger, Chris van Merwijk, and Rohin Shah.

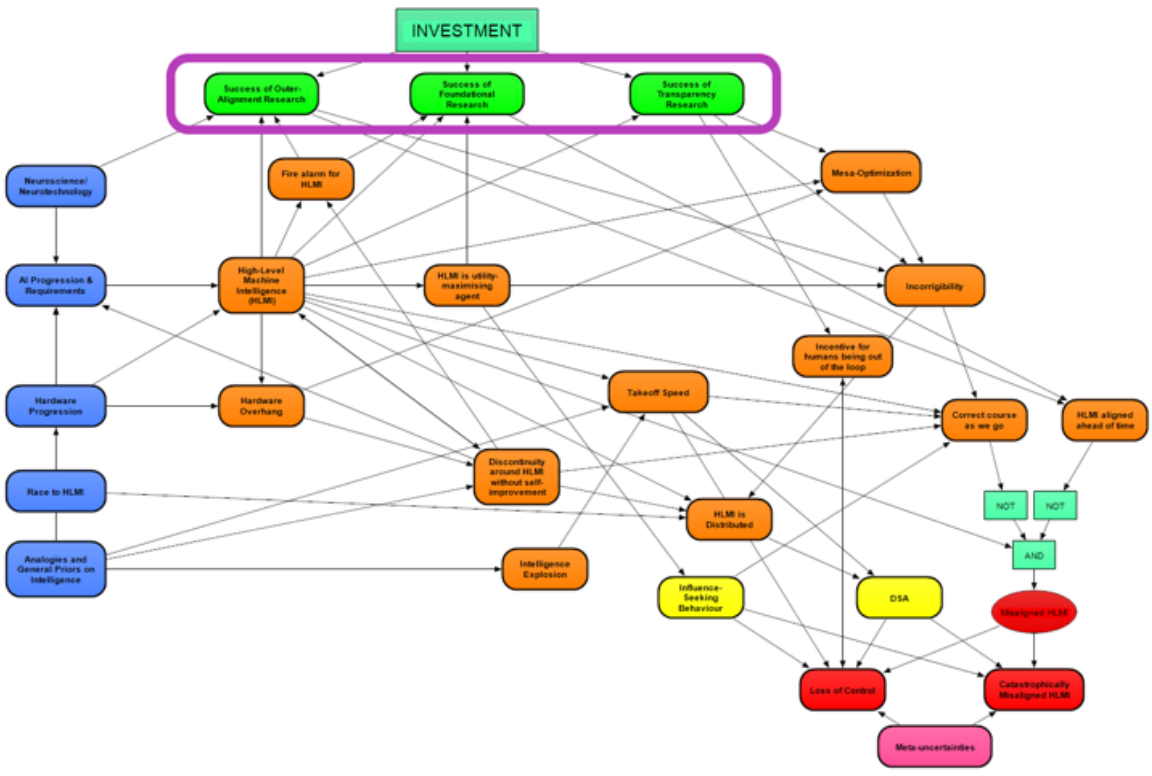
Modeling the impact of safety agendas

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post, which deals with how safety research - that is, technical research agendas aiming to reduce AI existential risk - might impact risks from AI, is part 6 in our sequence on Modeling Transformative AI Risk. In this series of posts, we are presenting a preliminary model of the relationships between key hypotheses in debates about catastrophic risks from AI. Previous posts in this sequence explained how different subtopics of this project, such as AI takeoff and mesa-optimization, are incorporated into our model.

We caution that this part of the model is much more of a work in progress than others. At present, it is best described as loosely modeling a few aspects of safety agendas, and the hope is that it can be further developed to be similar quality to more complete portions of the model. In many cases, we are unclear about how different research agendas relate to specific types and causes of risk. While this is partly because this part of the model is a work in progress, it is also because the theory of impact for many safety agendas is still unclear. So in addition to explaining the model, we will highlight things that are unclear and what would clarify those things.

Modeling of different research areas is contained in the green-colored modules, circled below.



Key questions for safety agendas

We encountered several uncertainties about safety agendas which have made modeling them relatively difficult. While most of these uncertainties are explained throughout the next section, the following points seem like the most important questions to ask about a safety agenda:

1. What is the theory of change? What does success look like, and how does that reduce AI risk? What aspects of alignment is the agenda supposed to address?
2. What is assumed about the progression of AI? How much does the agenda rely on a particular AI paradigm?
3. What is the expected timeline of the research agenda, and how much does that depend on additional funding or buy-in?
4. What are likely effects from work on this agenda even if it doesn't succeed fully? Are there spillover effects for AI capabilities or for other safety agendas? Are there beneficial effects from partial success?

Besides just helping our model, it's worth us highlighting all the benefits of understanding the above for various safety agendas:

- the community can better understand what constitutes success and provide better feedback on the agenda,
- the researchers get a better idea of how to steer the research agenda as it progresses,
- funders (and researchers) can better evaluate agendas, and in turn prioritise funding or pursuing them,
- future research can find gaps in the safety-agenda space more easily.

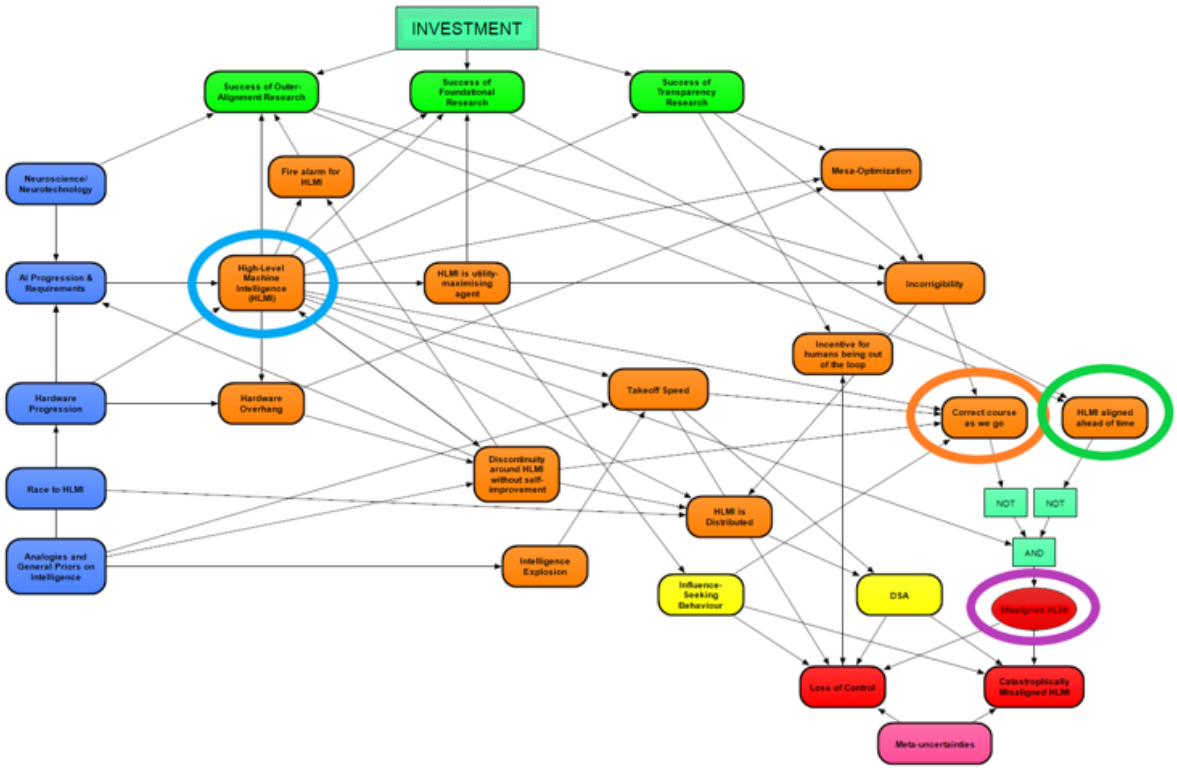
Model overview

Overall impact of safety research

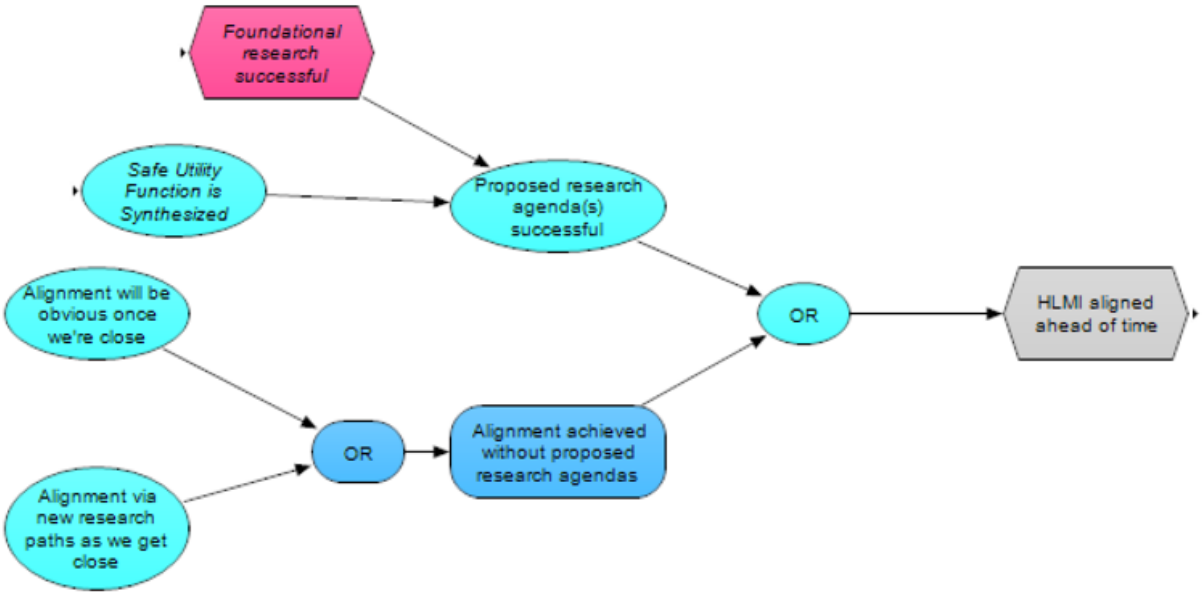
The key outcome to focus on for our model of research impact is *Misaligned HLMI*[1], circled in purple in the figure below. This node is obviously important to the final risk scenarios that we model, which will be covered in more detail in the next post. Looking at the inputs to this node, the model says we can avoid *Misaligned HLMI* if either

1. HLMI is never developed (blue-circled node), or
2. We manage to *Correct course as we go*, meaning HLMI is either aligned by default, or HLMI can be aligned in an iterative fashion in a post-HLMI world (orange-circled node), or
3. *HLMI [is] aligned ahead of time* - that is, people find a way to align HMLI before it appears or needs to be aligned (green-circled node).

While it seems that most people in the AI safety community believe option 3 is necessary for safe HLMI, option 2 is argued by some mainstream AI researchers and by some within the AI safety community, and option 1 would apply if humanity successfully coordinated to never build HLMI (though this possibility is not currently captured by our model). We will discuss option 2 further in the upcoming post on failure modes.



If we assume that aligning HLMI ahead of time *is* a worthwhile endeavour, i.e. conditioning on the model of the risks, how would success be possible? It would either be through the success of currently proposed research agendas and any of their direct follow-ups, or through more novel approaches developed as we get closer to HLMI. New approaches may come from new insights or from a paradigm shift in AI. This is essentially what the *HLMI aligned ahead of time* module is capturing, shown below. In particular, this module currently includes *Foundational research* (specifically the highly reliable agent designs agenda) and *Synthesized utility function*. We are currently including these two here for simplicity, but other agendas should fit in as well.

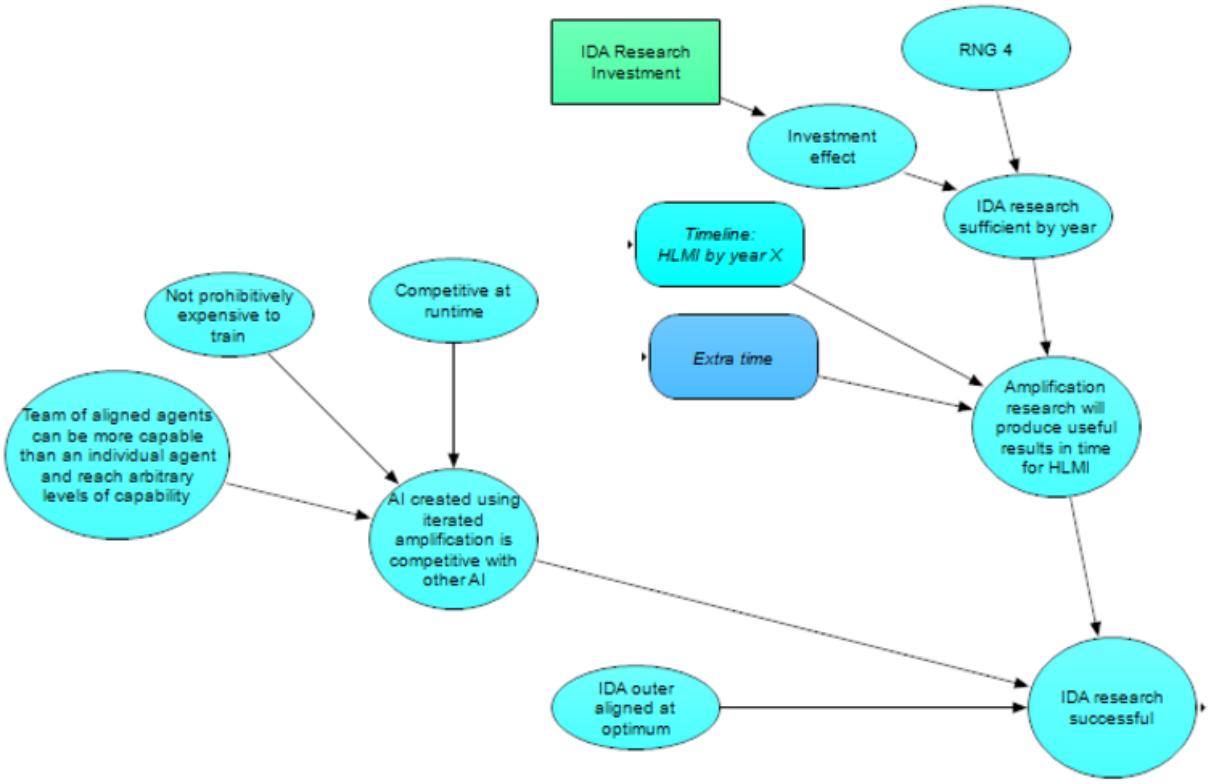


Our current model focuses primarily on currently proposed agendas since we can, and want to, model them more concretely. Additional work on both the current agendas, and the potential for future progress, would be useful in improving our understanding of the risk and building the model.

In this post we'll focus on three (or perhaps two and a half) different approaches to safety; 1) Iterated Distillation and Amplification, 2) Foundational Research, and 3) Transparency, which has been proposed as a useful part of several approaches to safety, but may not be sufficient alone. In the following sections we go through our preliminary models of these agendas, and point out major uncertainties we have about them in **bold**.

Iterated Distillation and Amplification

We will use the IDA research agenda as the main example to explain our uncertainties about modeling, as it appears to have more published detail than most other proposed agendas regarding: what is involved, its theory of change, and its assumptions about AI progress. The section of the model for IDA is shown in the figure below.



The final output of this section is *IDA research successful*. We assume that success means obtaining a clear, vetted procedure to align the intentions of an actual HLMI via IDA. However, the IDA agenda seems to address outer alignment (i.e. finding a training objective with optima that are aligned with the overseer), and not necessarily inner alignment (i.e. making a model robustly aligned with the training objective itself)[2]. This is a case where we are uncertain **what parts of the alignment problem the agenda is supposed to solve**, and **how a partial solution to alignment is expected to fit into a full solution**.

For IDA and other agendas, it's also difficult to reason about *degrees* of success. **What if the research doesn't reach its end goal, but produces some useful insights?** Relatedly, **how could the agenda help other agendas if it does not directly achieve its aims?** We're uncertain how researchers think about this, and how best to model these effects.

Continuing through the model, towards the right of the figure we have a section about the "race" between IDA research and AI capabilities research, summarised as *Amplification research will produce useful results in time for HLMI*. We're particularly unsure how to think about **timelines to solve research agendas**, and we are interested in either community feedback about their understanding of the timelines for success, or any insights on the topic.

Currently, we model research timelines using the node *IDA research sufficient by year X*, where *X* is affected by both *Investment effect* and randomness. This result is then modified by the node *Extra time*, which models the possibility that a "fire alarm" for HLMI is recognised and speeds up safety research in the years before HLMI, either through insight or increased resources. This time to sufficient IDA research is then

compared to the timeline for HLMI (*Timeline: HLMI by year X*), where success is dependent on the time to IDA being less than the time to HLMI.

The competitiveness of IDA is modeled in a section toward the left of the figure. We break competitiveness down into *Not prohibitively expensive to train*, *Competitive at runtime*, and whether IDA scales to arbitrary capabilities (*Team of aligned agents can be more capable than an individual agent and reach arbitrary levels of capability*). All are modeled as being necessary for IDA to be competitive.

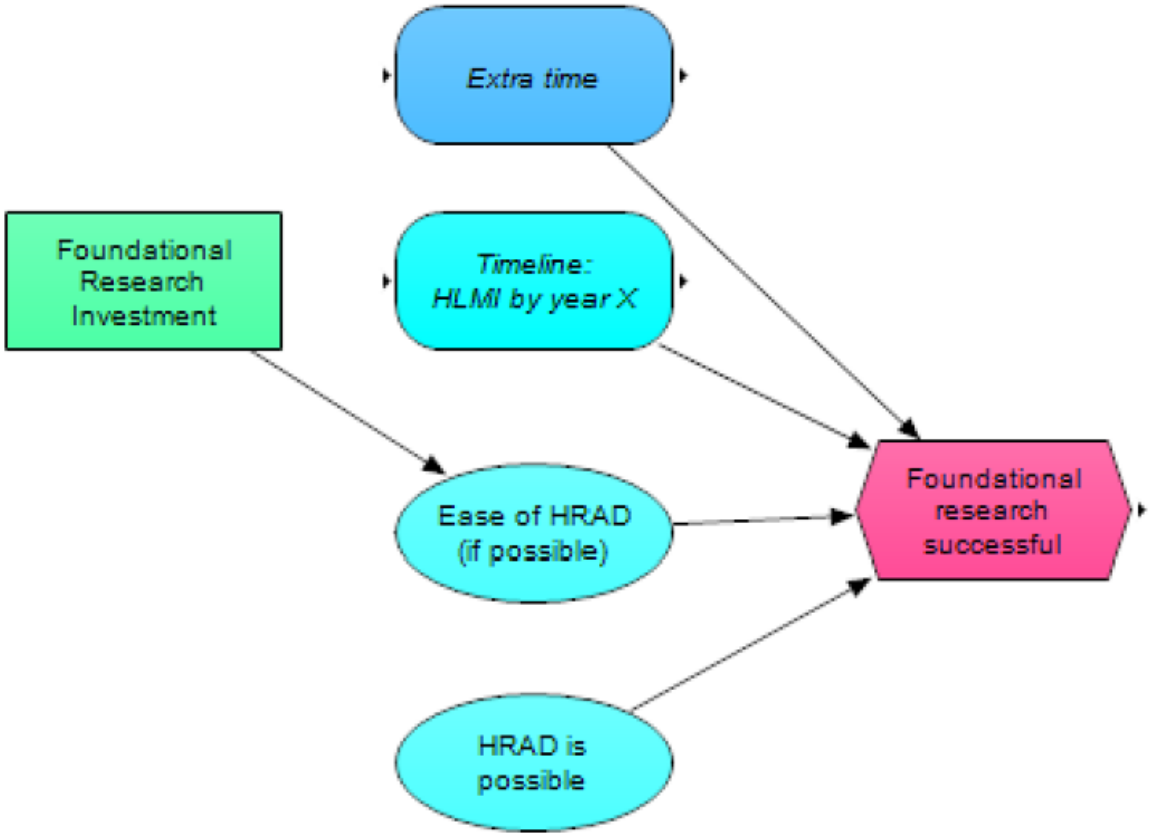
Finally, we have one node to represent whether *IDA [is] outer aligned at optimum*. This means that all possible models which are optimal according to the training objective are at least intent-aligned. Being outer aligned at optimum has been argued to defuse much of the threat from Goodhart's Law, specifically the Causal and Extremal variants of it. So IDA being outer aligned at optimum would be important for the IDA agenda to be successful at alignment, and therefore a key uncertainty.

Putting it all together, our model considers IDA to be a workable solution for outer alignment if and only if IDA research wins in a “race” against unaligned HLMI, and IDA is sufficiently competitive with other approaches of HLMI, and IDA is outer aligned at optimum. The output node *IDA research successful* then feeds into the *Incorrigibility* module at the top level of the model - that is, if it's successful and we have an intent-aligned HLMI, then it *is* corrigible. Corrigibility in turn increases the ability to *Correct course as we go* (and so on, as explained in the previous section).

Foundational Research

Our work on modeling Foundational Research has focused entirely on MIRI’s highly reliable agent designs (HRAD) research agenda, as this has had the most discussion in the AI alignment community (out of all the technical work in Foundational Research). Trying to model disagreements about the value of the HRAD agenda led to the post Plausible cases for HRAD work, and locating the crux in the "realism about rationality" debate. To summarize the post, one of the difficulties with modeling the value of HRAD research is that there seems to be disagreement about what the debate is even about. The post tries to organize the debate into three “possible worlds'' about what the core disagreement is, and gives some reasons for thinking why we might be in each world. The discussion in comments did not lead to a consensus, so more work will probably be needed to make our thoughts precise enough to encode in the graph structure of the model.

The model below is a simpler substitute, pending further work on the above. Like the IDA model, this considers whether the research can succeed in time for HLMI. Besides that, there are two nodes about the possibility and difficulty of HRAD. The *Foundational research [is] successful* node feeds directly into the *HLMI aligned ahead of time* module shown previously.



Transparency

Here, we are using the term "transparency" as shorthand for transparency and interpretability research that has long-term AI safety as a core motivation.

Transparency can be applied to whole classes of machine learning models, and may be a part, or complement, of several alignment techniques. In An overview of 11 proposals for building safe advanced AI, "transparency tools" form a key part of several proposals, for different reasons. More recently, Transparency Trichotomy analysed different ways that transparency can help understand a model: via inspection, or training, or architecture. The parts of the trichotomy can also work together, for instance by using transparency via inspection to get more informed oversight, which then feeds back into the model via training. So the current structure of our model, with its largely separate paths to impact for each agenda, does not seem well suited to transparency research. However, the Mesa-optimization module does incorporate some nodes on how transparency research may help detect deception (via inspection) or actively avoid deception (via training).

Theories of change for transparency helping to align HLMI have become clearer in published writing over the last couple of years. The post Chris Olah's views on AGI safety offers several claims on theory of change, including:

1. Transparency tools give you a mulligan - a chance to recognise a bad HLMI system, and try again with better understanding.

2. Advances in transparency tools feed back into design. If we build systems with more understanding of how they work, then we can better understand their failure cases and how to avoid them.
3. Careful analysis using transparency tools will clarify what we *don't* understand too. Pointing out what we don't understand will generate more concern about HLMI.
4. Transparency tools help an overseer to give feedback not just on a system's output, but also the process by which it produced that output.
5. Advances in transparency tools (and demonstrating their usefulness and appeal) helps realign the ML community to focus on deliberate design and understanding.

From the above claims and discussion elsewhere, there are several apparent cruxes for transparency helping to align HLMI:

1. Using transparency tools will not make enough progress (or any progress) on the "hard problem" of transparency. The hard problem is to figure out what it even means to understand a model, in a way that can save us from Goodhart's Law and deception. As discussed in Transparency Trichotomy, transparency tools can themselves be gamed. There seems to be agreement that transparency tools will not get us *all* the way on this problem, but disagreement about how much they help - see e.g. this thread and this comment (bullet point 3).
2. Similar to the above, though we are not sure if this is a distinct crux for anyone: there is a risk that transparency tools make the flaws we are trying to detect harder to understand (discussed here and here), so there is too great a risk that the tools cause net harm.
3. Transparency tools will not scale with the capabilities of HLMI and beyond - discussed here. The crux could be specifically about the amount of labour required to understand increasingly large models. It could also be about increasingly capable systems using increasingly alien abstractions. The linked post suggests that an amplified overseer could get around this problem, so the crux could actually be in whether an amplified overseer can make transparency scale reliably in place of humans.
4. The available transparency tools will not be useful for the kind of system that HLMI is (e.g. the work on Circuits in vision models will not transfer well to language models). This is like a horizontal version of the above scaling crux. Chris Olah raised this point himself.

Again, more work is needed to structure our model in a way that incorporates the above cruxes.

Other agendas

Other agendas or strategies which we have not yet modeled include:

- Counterfactual oracles, STEM AI, quantilizers, myopic cognition (discussed in Arguments against myopic training), debate
- Multi-agent safety (there is writing on it, and some issues have been identified but we are not aware of a research agenda)
- Aligning current systems (e.g. large language models) - see The case for aligning narrowly superhuman models

Some of the above are more difficult to model because there is less writing that clearly outlines paths to impact or what success looks like. A potentially valuable project is to

make a clearer case to the community for how a given research agenda could be impactful, and explaining what the goals and specific approaches are.

Help from this community

Our tentative understanding suggests that more public effort to understand and clearly articulate safety agendas' impacts, driving beliefs, and main points of disagreement would be really helpful. Examples of good work in this area are An overview of 11 proposals for building safe advanced AI, and Some AI research areas and their relevance to existential safety. This work can take a lot of effort and time, but some of the uncertainties highlighted in this post seem fairly easy to clarify through comments or smaller write-ups.

To illustrate the kind of information that would help, we have written the following condensed explanation of an imaginary agenda (the agenda and opinions are made-up - this is not quoting anyone):

> This agenda aims to increase the chance that high-level machine intelligence (HLMI) is inner-aligned. More specifically, it will defuse the threat of deceptively aligned HLMI. The path from deceptively aligned HLMI to existential catastrophe is roughly: such a system would be deployed due to economic or other incentives and lack of apparent danger. It would also be capable enough to take the long-term future out of humanity's control. While we have a very wide distribution over how humanity loses control, we expect a scenario similar to scenario 2 of What Failure Looks Like.
>
> The specific outcome we are aiming for is <alignment procedure>. For this to succeed and be scalable, we rely on AI progressing like <current machine learning trends>. We expect the resulting AI to be competitive, with training time on the same order of magnitude and performance within 20% of the unaligned baseline.
>
> With regard to timelines, we tentatively estimate that this work is at its most valuable if HLMI is produced in the medium term, neither in the next 10 years, nor more than 30 years from now. With our current resources, we give a rough 5% chance of having a viable procedure within 5 years, and a 10% chance within 20 years. This increases to 20% and 30% respectively with <additional resources>. The remaining subjective estimated probability of failure is split evenly between obstacles from the theory, or project management, or external factors. This agenda relies on outer alignment being solved using <broad outer alignment approach>, but otherwise not interacting much with the problem we aim to solve.

Finally, as a way of quickly gathering opinions, we would love to see comments on the following: **for any agenda you can think of, or one that you're working on, what are the cruxes for working on it?**

In the next post, we will look at the failure modes of HLMI and the final outcomes of our model.

Acknowledgements

Thanks to the rest of the MTAIR Project team for feedback and suggestions, as well as Adam Shimi and Neel Nanda for feedback on an early draft.

1. We define High-Level Machine Intelligence (HLMI) as machines that are capable, either individually or collectively, of performing almost all economically-relevant information-processing tasks that are performed by humans, or quickly (relative to humans) learning to perform such tasks. We are using the term “high-level machine intelligence” here instead of the related terms “human-level machine intelligence”, “artificial general intelligence”, or “transformative AI”, since these other terms are often seen as baking in assumptions about either the nature of intelligence or advanced AI that are not universally accepted. ↩

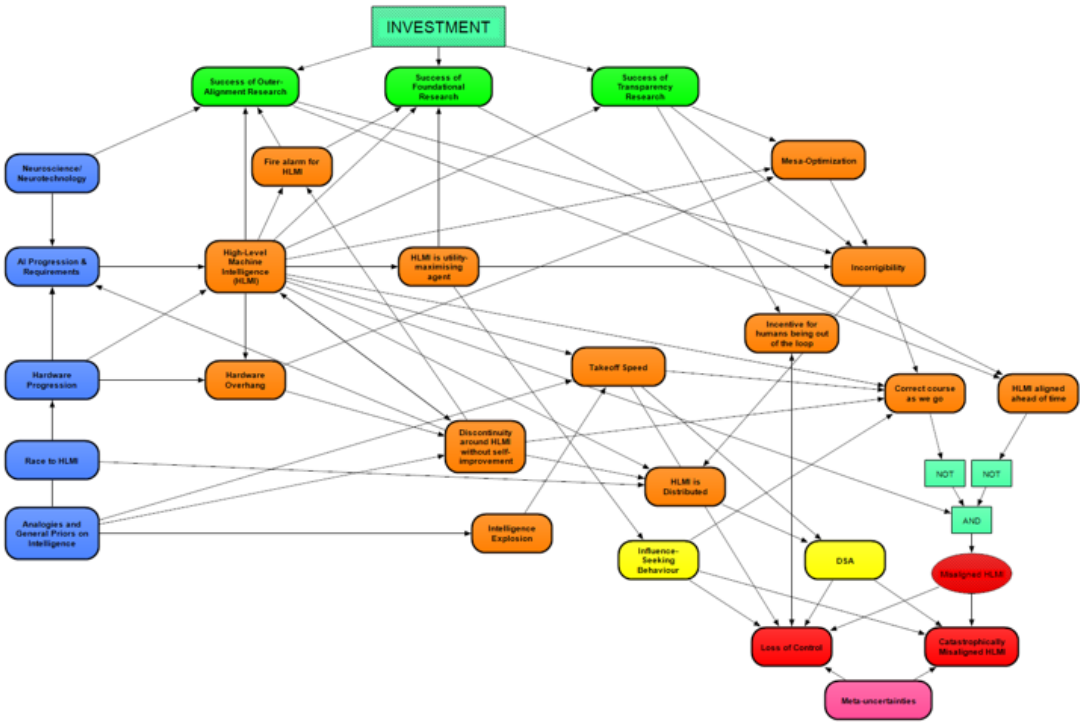
2. For more on this distinction/issue, see this post. ↩

Modeling Failure Modes of High-Level Machine Intelligence

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post, which deals with some widely discussed failure modes of transformative AI, is part 7 in our sequence on Modeling Transformative AI Risk . In this series of posts, we are presenting a preliminary model of the relationships between key hypotheses in debates about catastrophic risks from AI. Previous posts in this sequence explained how different subtopics of this project, such as mesa-optimization and safety research agendas , are incorporated into our model.

We now come to the potential failure modes caused by High-Level Machine Intelligence (HLMI). These failure modes are the principal outputs of our model. Ultimately, we need to look at these outputs to analyze the effect that upstream parts of our model have on the risks, and in turn to guide decision-making. This post will first explain the relevant high-level components of our model before going through each component in detail. In the figure below, failure modes are represented by red-colored nodes.



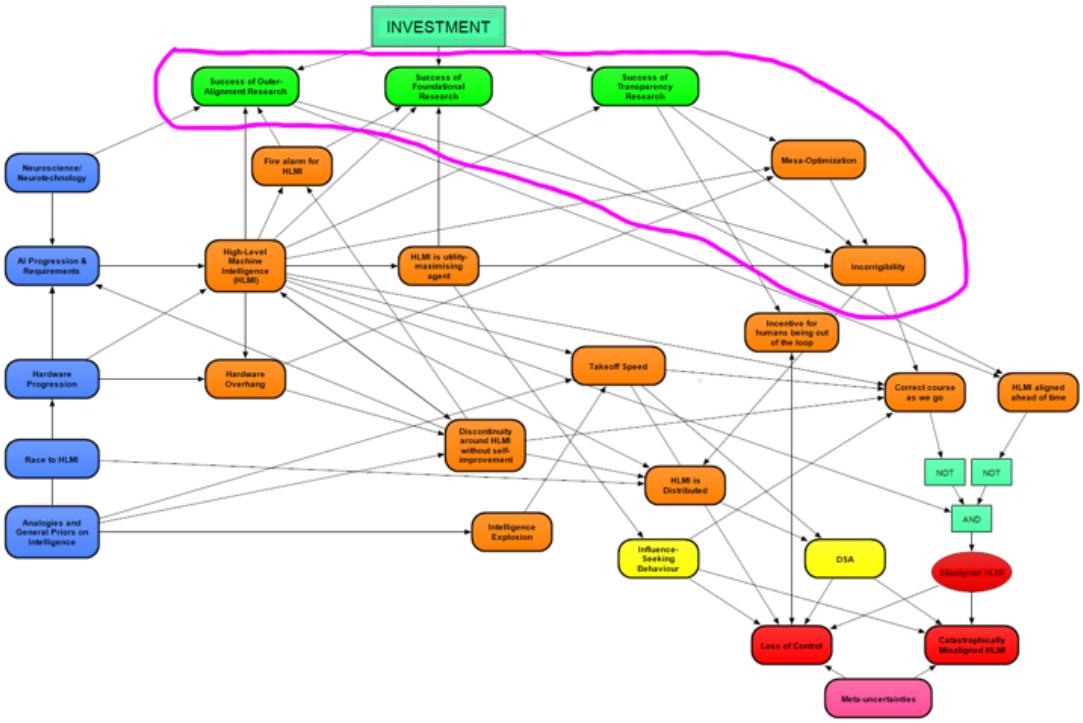
We do not model catastrophe itself in detail, because there seem to be a very wide range of ways it could play out. Rather, the outcomes are states from which existential catastrophe is very likely, or "points of no return" for a range of catastrophes. **Catastrophically Misaligned HLMI** covers scenarios where one HLMI, or a coalition of HLMIs and possibly humans, achieves a **Decisive Strategic Advantage (DSA)**, and the DSA enables an existential catastrophe to occur. **Loss of Control** covers existential scenarios where a DSA does not occur; instead HLMI systems gain influence on the world in an incremental and distributed fashion, and either humanity's control over the future gradually fades away, or a sudden change in the world causes a sufficiently large and irreversibly damaging automation

failure (see What failure looks like); alternatively, the use of HLMI leads to an extreme "Moloch" scenario where most of what we value is burned down through competition.

Moving one step back, there are three major drivers of those outcomes. Firstly, **Misaligned HLMI** is the technical failure to align an HLMI. Second, the **DSA** module considers different ways DSA could be achieved by a leading HLMI project or coalition. Third, **Influence-seeking Behavior** considers whether an HLMI would pursue accumulation of power (e.g. political, financial) and/or manipulation of humans as an instrumental goal. This third consideration is based on whether HLMI will be agent-like and to what extent the Instrumental Convergence thesis applies.

Outcomes of HLMI Development

As discussed in the "impact of safety agendas" post, we are still in the process of understanding the theory of change behind many of the safety agendas, and similarly our attempts at translating the success/failure of these safety agendas into models of HLMI development are still only approximate. With that said, for each of the green modules shown below on the **Success of [agenda] Research**, we model their affect on the likelihood of an incorrigible HLMI - in particular, if the research fails to produce its intended effect. Meanwhile, the Mesa-optimisation module evaluates the likelihood of dangers from learned optimization which would tend to make HLMI incorrigible.

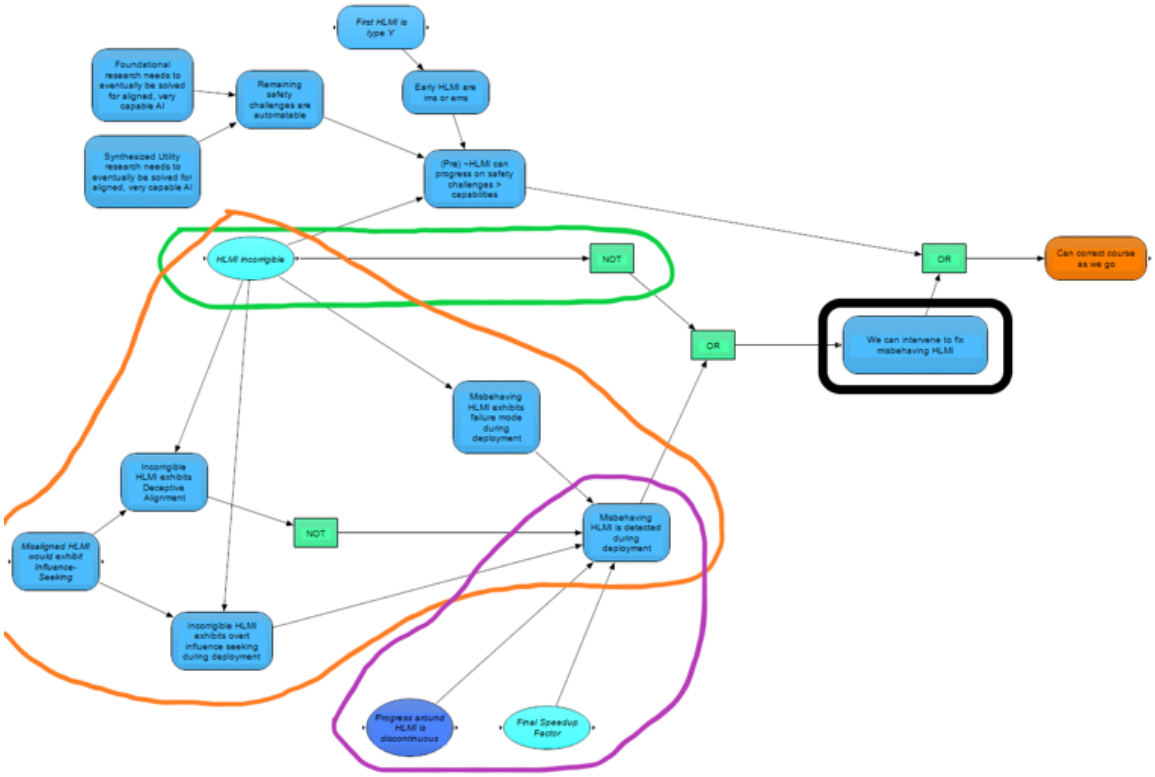


Besides **Incorrigibility**, the question of HLMI being aligned may be influenced by the expected speed and level of discontinuity during takeoff (represented by the modules **Takeoff Speed** and **Discontinuity around HLMI without self-improvement**, covered in Takeoff Speeds and Discontinuities). The expectation is that misalignment is more likely the faster progress becomes, because attempts at control will be harder if less time is available. Those factors - incorrigibility and takeoff - are more relevant if alignment will be attempted in an iterated manner post-HLMI, so they connect to the **Correct course as we go** module. On the other hand, if we have **HLMI aligned ahead of time** (another module), then we

assume that the **Success of [agenda] Research** is the key influence, so those are connected up. We explain these two downstream modules about alignment in the following sections.

Correct Course as we Go

The **Correct course as we go** module is shown below. We consider two possibilities that would allow us to correct course as we go. One is that **we can intervene to fix misbehaving HLMI**, circled in black in the image.



Reminder: images can be viewed full size with right-click => open in new tab

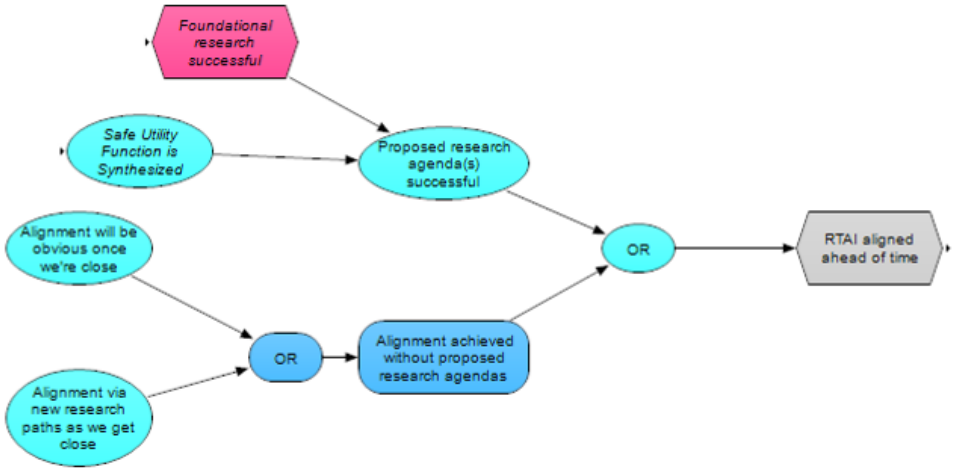
In general, if HLMI is corrigible - either due to getting alignment "by default", or due to the success of safety research agendas aimed at corrigibility - then we can expect that it will be possible to deal with potentially dangerous behaviours as they arise. This is modeled by the green-circled nodes. If HLMI is incorrigible, then the risk of not catching and stopping any dangerous behaviors depends on two factors. First, it depends on the nature of these behaviors; if the HLMI is deceptively aligned, or otherwise exhibits influence-seeking behaviour (discussed in a later section), then it is less likely we will notice misbehavior - this is modeled by the orange-circled nodes. Second, the risk depends on how fast progress in AI is when we develop HLMI: the less time that we have to catch and correct misbehavior, and the more rapidly HLMI increases in capability, the more likely we are to fail to course correct as we go - this is modeled by the purple-circled nodes.

The other way we could course correct as we go is if **[AI systems that are (close to) HLMI] can progress on safety challenges [more than] capabilities**, modeled in the top left of the figure. This seems more likely if early HLMI takes the form of imitation learners (ims) or emulations of humans (ems). It also seems more likely if we reach a point in AI safety research where the **remaining safety challenges are automatable**. Reaching that point in turn depends on the progress of specific safety agendas - currently we just include

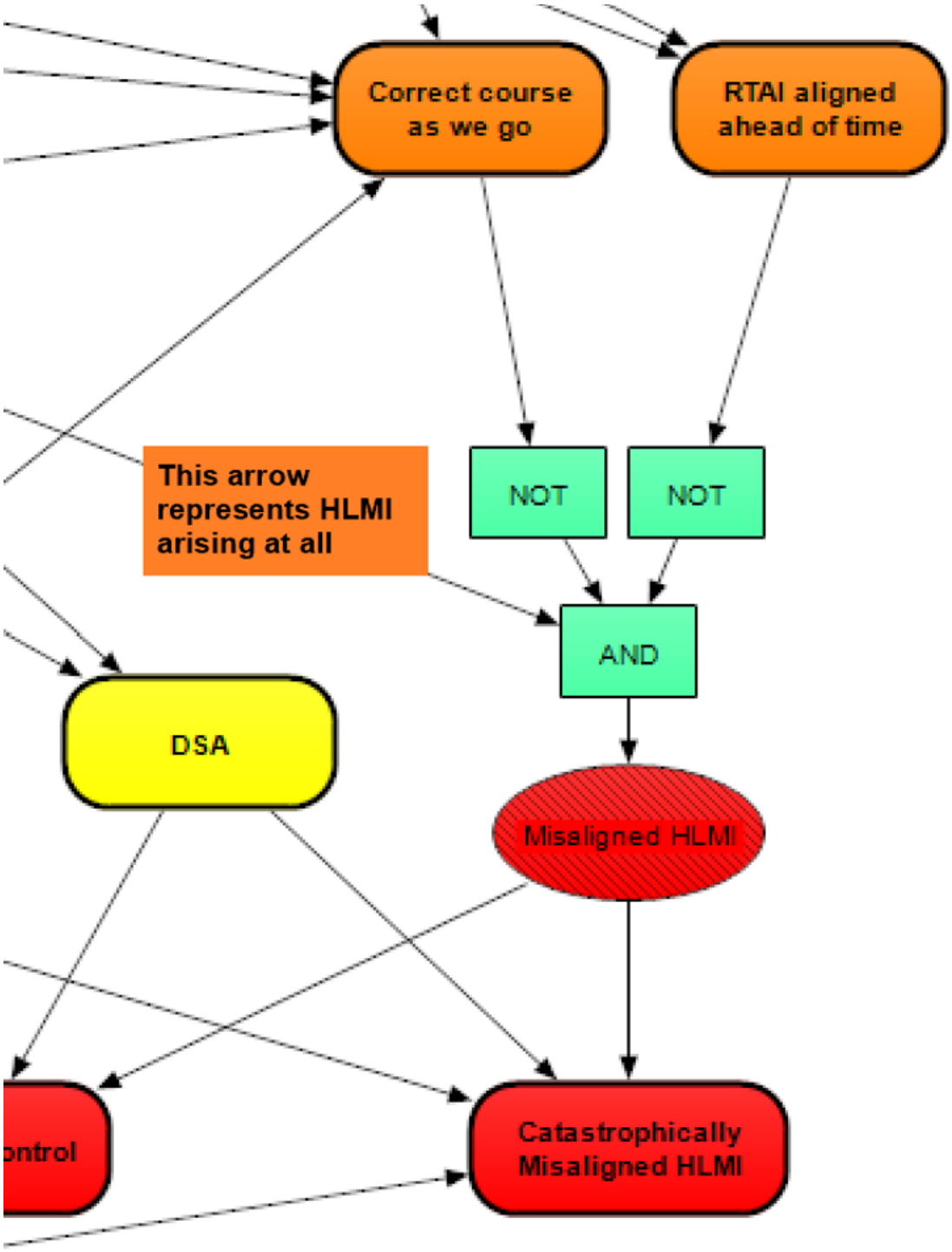
foundational research and synthesising a utility function. As the model is further developed, it will additionally depend on the success of other safety agendas, as we clarify their routes to impact.

Alignment ahead of time

Some safety agendas - principally those based around ambitious value learning - don't seem to depend on ensuring corrigibility, and instead of trying to catch misalignment issues that arise, they focus on ensuring that HLMI is successfully aligned on the first try. We model this effect of safety agendas separately in the **HLMI aligned ahead of time module**, as the agendas do not depend on catching dangerous behaviors post-HLMI, and therefore if they are successful, are less affected by the speed of progress post-HLMI. The module (shown below) is a simplistic version of this idea, and once again, other agendas besides the ones shown may fit into its logic. We discussed this module from a slightly different perspective in the previous post on safety agendas.



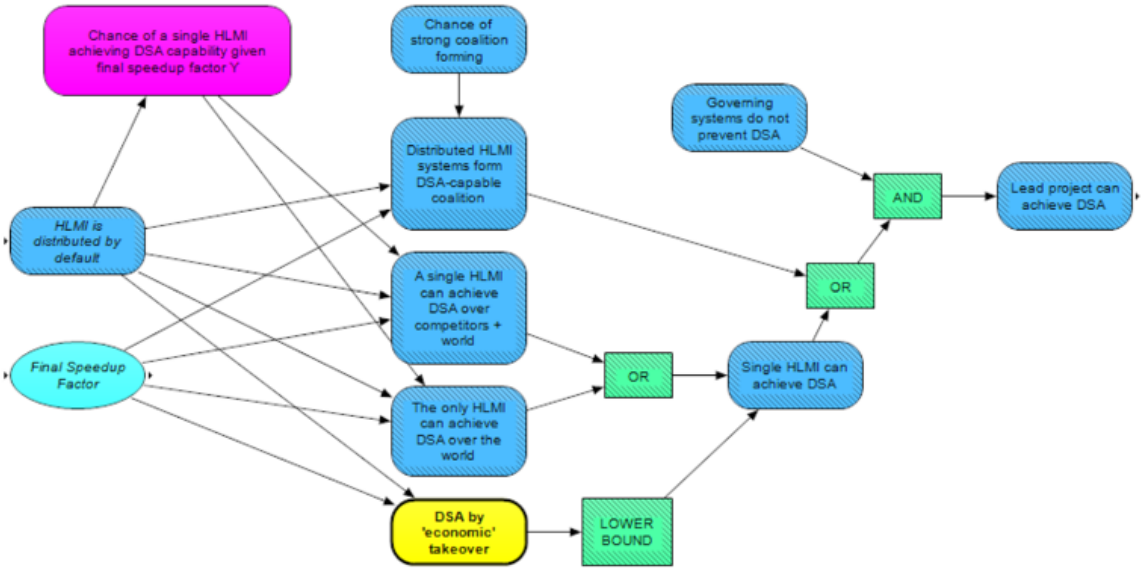
Misaligned HLMI



We decomposed the risk factor of **Misaligned HLMI** into three conditions. The first condition is that HLMI arises at all. The second condition is that it's not feasible to correct course as we go - this would depend on HLMI not being benign by default, and humanity not developing and implementing technical solutions that iteratively align HLMI in a post-HLMI world. The third condition is that we do not find a way to align HLMI before it appears, or before some other pre-HLMI point of no return. This third condition is a big uncertainty for most people concerned about AI risk, and we have discussed how it is particularly difficult to model in the previous post on the impact of safety research agendas.

Keeping this part of the model leading up to **Misaligned HLMI** in mind, we now move to another major driver of risk in our model, **Decisive Strategic Advantage**.

Decisive Strategic Advantage (DSA)



Attaining a DSA through HLMI is a key factor in many risk scenarios. We assume that if DSA is attained, it is attained by a "project" or a coalition of different projects. A project could be a team within a tech company, a state-owned lab, a military, a rogue actor, etc. Working from right to left in the figure pictured above, the leading project to develop HLMI can only achieve DSA if governing systems do not prevent it from independently amassing such power. Governing systems include state governments, institutions, laws and norms. AI systems, including other HLMI projects, can also count as governing systems. Failing the intervention of governing systems, there are three main ways to gain the potential for DSA. First, multiple HLMI systems could form a coalition that achieves a DSA. Second, the first HLMI may achieve a DSA over the rest of the world before competitors arise. Third, in a world with multiple HLMIs, a single HLMI may achieve a DSA over the rest of the world (including over its competitors).

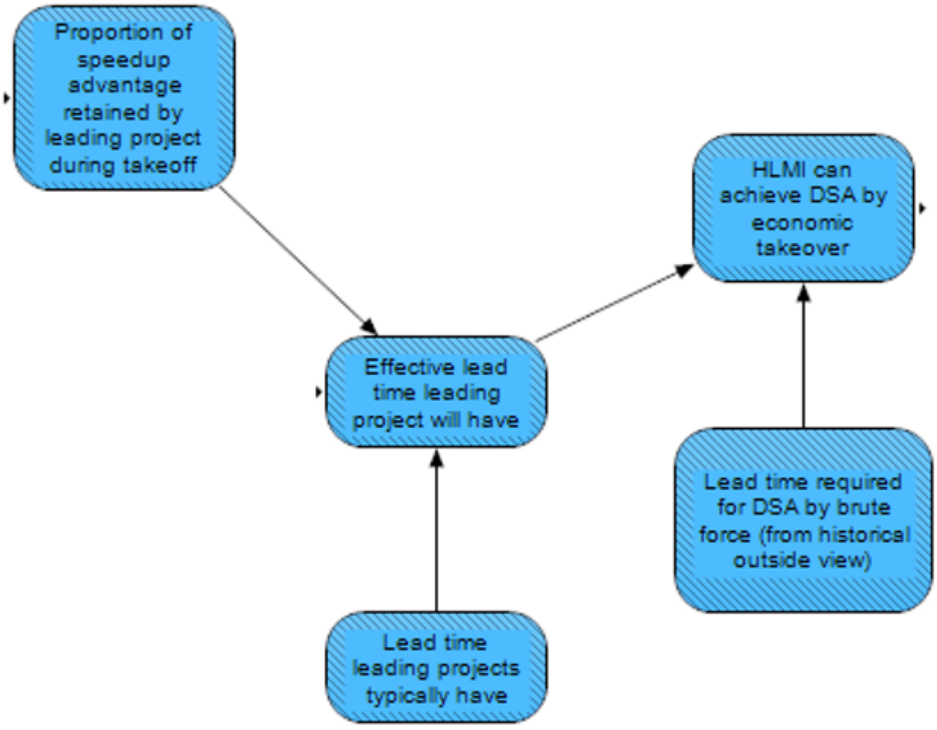
These three paths to potential DSA depend on whether **HLMI is distributed by default**, from the module **HLMI is distributed** covered in a previous post on takeoff speeds and discontinuities. Potential for DSA is also linked to the rate of post-HLMI economic growth, represented by **Final Speedup Factor**. Faster growth will tend to widen the gap between the leading HLMI project and others, giving the project a greater advantage. The greater advantage makes it more likely that other projects are either abandoned, or taken over by the leading project. Therefore, faster growth seems to make a single-HLMI DSA more likely, all else equal.

The level of advantage needed for DSA is difficult to anticipate or even measure. It might be argued that an overwhelming economic or technical advantage is required for a DSA, but various arguments have been put forward for why this is incorrect. Many times in history, groups of humans have used manipulation to take control of countries or empires via political manipulation. For instance, Hitler's initial takeover of Germany was mostly via manipulating the political system (as opposed to investing his personal savings so well that he was responsible for >50% of Germany's GDP), and Napoleon's Grande Armée was built during his takeover of Europe, not beforehand. In a more modern context, individuals have gained

control over corporations via leverage and debt, without themselves having resources needed to assert such control.

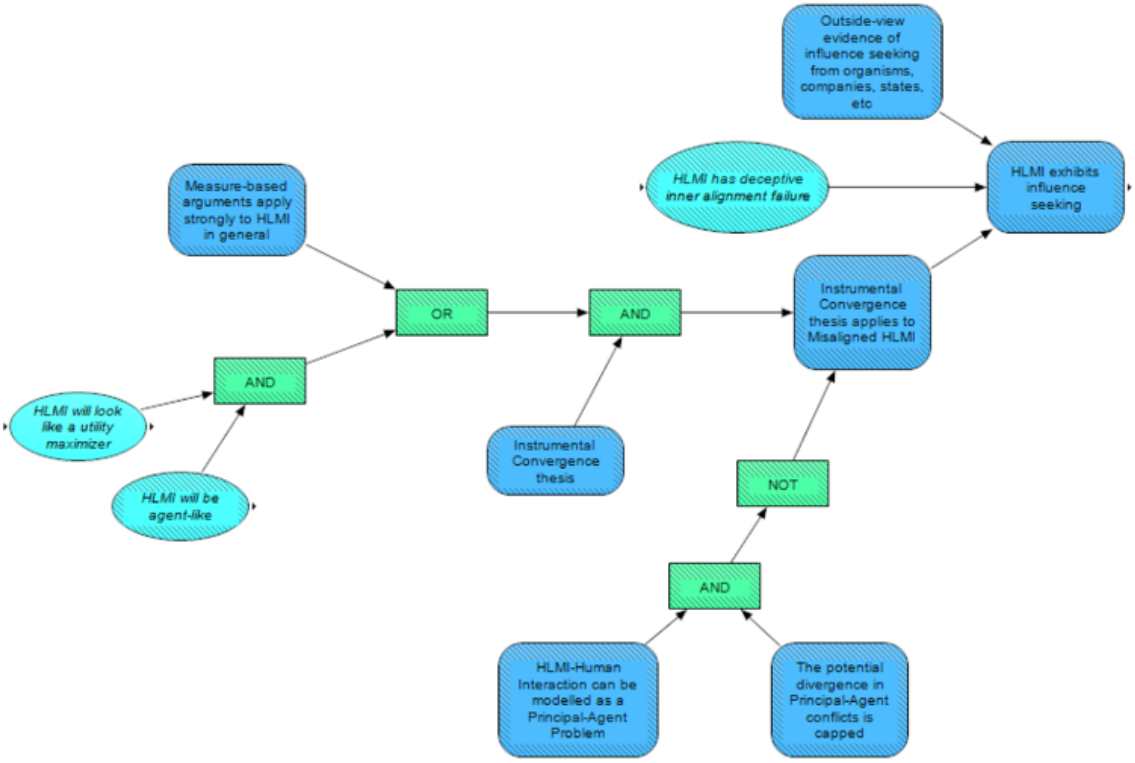
On the other hand, if these analogies don't hold and an HLMI project is unable to exert significant manipulative pressure on the rest of the world, it might require actual control over a significant fraction of all the world's resources before DSA occurs. Therefore, we must have a very wide range of uncertainty over the degree of advantage required to achieve a DSA. There are a couple of nodes intended to capture this uncertainty. The yellow **DSA by "economic takeover"** module (pictured below) provides an upper bound on the degree of advantage needed (or equivalently, as stated in the diagram, a lower bound on the likelihood) based on the assumption that such an upper bound would be 50% of the world's GDP. The purple **Chance of a single HLMI achieving DSA** node attempts to estimate the likelihood of DSA given different takeoff scenarios, taking into account, among other things, the possibility of HLMI-assisted manipulation of human society, or gaining control of resources or new technologies.

The Economic takeover route assumes that the HLMI takes no steps to seize resources, influence people or attack its potential opponents and just outgrows the rest of the world until it is overwhelmingly powerful. The logic here is that if the lead project increased its economic position to be responsible for >50% of the world GDP, then we would assume it would already have "taken over" (most of) the world without firing a shot. If we know how large the project is when it's on the cusp of HLMI compared to the size of the world economy (both modeled in the Paths to HLMI post), and if we know how fast the project will grow economically once it reaches HLMI (which we can naively assume is similarly quickly to the economic growth rate after HLMI, as determined in our post on Takeoff Speeds and Discontinuities), then we can determine how long it would have to outgrow the rest of the world to represent >50% of GDP. We then might assume that if the leading project has a lead time larger than this time, it will achieve economic takeover (the lead time can itself be estimated based on the typical lead time of large engineering projects and what portion of the lead time we might expect the leading project to keep during AI takeoff). See this sequence for further discussion.



The **DSA by "economic" takeover** submodule within the DSA module

Influence-Seeking Behaviour



Another key factor to many risk scenarios is influence-seeking behaviour. We use this term to mean that an AI tends to increase its ability to manipulate people for instrumental reasons, either through improved manipulation skills, or acquiring of resources for the purposes of manipulation. This doesn't require that the AI system have an explicit objective of control, but having effective control may still be the goal, at least when viewing the system's behaviour with an intentional stance.

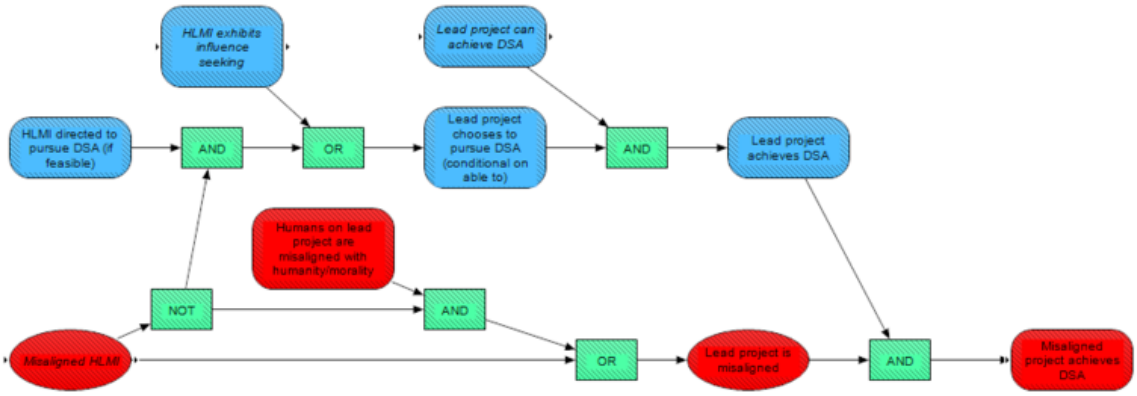
As with many things, we can use analogies as one input to the likelihood of HLMI having influence-seeking behavior. Some plausible analogies for AI in the category of influence-seeking behaviour include organisms, companies, and states. In each of these classes there are varying degrees of influence seeking, so it would make sense to study the conditions for this behavior. For now we have kept this reasoning by analogy very simplified as a single node.

Part of the Instrumental Convergence thesis is that a sufficiently capable and rational agent will pursue instrumental goals that increase the likelihood of achieving its terminal goal(s). In general, these instrumental goals will include some commonalities, such as resource acquisition, self-preservation, and influence-seeking abilities. For this reason, whether the **Instrumental Convergence thesis applies to Misaligned HLMI** is an important factor in whether it exhibits influence seeking. We separated this question into (a) how strong the **Instrumental Convergence thesis** is in general, and (b) how much HLMI meets the conditions of the thesis. For (b), one could argue that **Measure-based arguments** (e.g. "a large fraction of possible agents with certain properties will have this behaviour") **apply strongly to HLMI in general**. Alternatively, there is the more direct question of whether HLMI will look like a utility maximizer and be agent-like. The latter is a point of disagreement that has received a lot of attention in the discourse on AI risk, and is handled by the **HLMI is utility-maximising agent** module (this module is still a work in progress and is not covered by this sequence of posts).

Another important factor in whether HLMI exhibits influence-seeking is **HLMI has deceptive inner alignment failure**. This node is derived from the node **Mesa-optimizer is**

deceptively aligned in the **Mesa-optimization** module. Deceptive alignment entails some degree of influence-seeking, because the AI will perform well on its training objective for instrumental reasons - namely, so it gets deployed and can pursue some "true" objective. Deceptive alignment and influence seeking are not the same thing, because a deceptively aligned AI may not continue to seek influence once it is deployed. However, influence seeking algorithms might be favoured among all deceptively aligned algorithms, if they are more generally effective and/or simpler to represent.

Catastrophically Misaligned HLMI

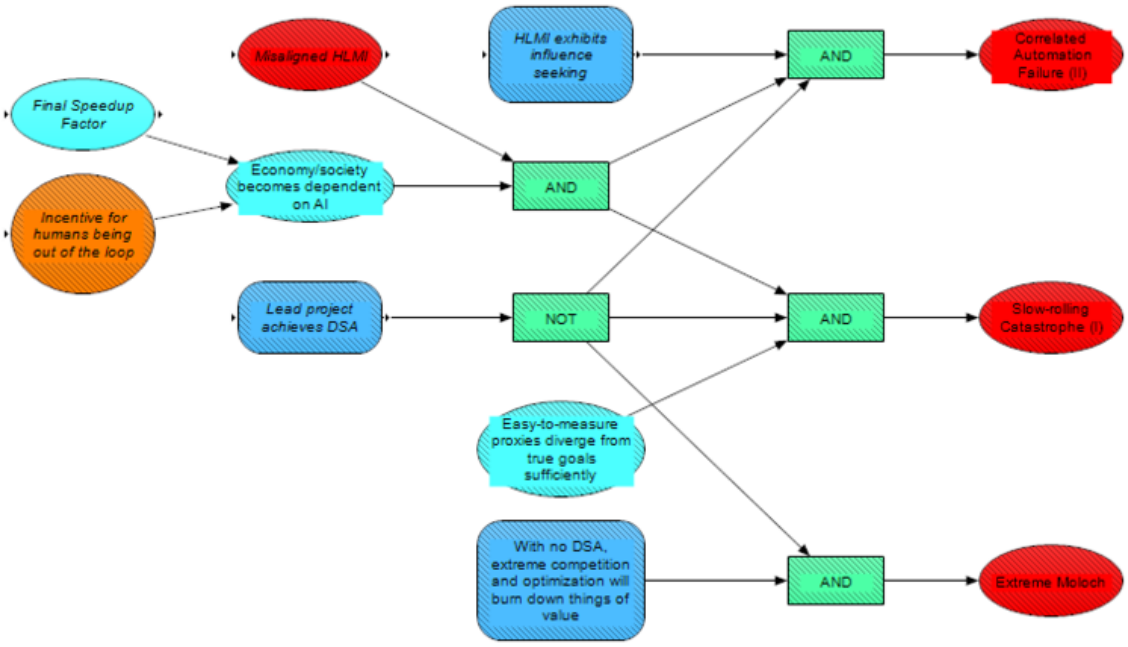


Returning to the scenario of **Catastrophically Misaligned HLMI**, we need to consider that misalignment can happen in many possible ways and to many degrees. So when should we consider it catastrophic? We operationalize catastrophe as when a **Misaligned project achieves DSA**, the node shown in red all the way on the right. The "project" includes HLMI(s) and any humans that are able to direct the HLMI(s), as well as coalitions of projects. If a project is capable of neutralizing any potential opposition due to holding a DSA, and its interests are misaligned such that after doing so it would cause an existential catastrophe, then we assume that it will cause an existential catastrophe. Naturally, this final outcome is the conjunction of **Lead project achieves DSA** and this **Lead project is misaligned.**

We decomposed the question of whether the **Lead project achieves DSA** into whether it **can achieve DSA** (from the **DSA** module), and whether it **chooses to pursue DSA (conditional on being able to)**. Then we further break the latter of these down into two ways the HLMI project could choose to pursue DSA: either the **HLMI exhibits influence seeking**, or the HLMI is aligned with members of the project (i.e., it would **NOT** be **misaligned**) and **[is] directed to pursue DSA**.

Regarding whether the lead project is misaligned, this can either occur if we wind up with **Misaligned HLMI**, or if we do not get misaligned HLMI (from a technical perspective) but the **Humans on [the] lead project are [sufficiently] misaligned with humanity [or] morality** that they nevertheless cause an existential catastrophe (i.e., existential misuse).

Loss of Control



Here, we consider the possibility of humanity suffering existential harm from **Loss of Control** – not to a single HLMI or a coalition that has achieved a DSA, but instead to a broader HLMI ecosystem. Such an outcome could involve human extinction (e.g. if the HLMI ecosystem depleted aspects of the environment that humans need to survive, such as the oxygen in the atmosphere) but wouldn't necessarily (e.g. it could involve humans being stuck in a sufficiently suboptimal state with no method to course correct).

The three outcomes we model are named after popular descriptions of multi-polar catastrophes. Two of these come from What failure looks like: **Slow-rolling Catastrophe (I)** and **Correlated Automation Failure (II)**, and the third (**Extreme Moloch**) comes from Meditations on Moloch. For each of these, we're considering a broader set of possibilities than is given in the narrow description, and we use the description just as one example of the failure mode (for a more granular breakdown of some subtypes of these scenarios, see this post). Specifically, here's what we're envisioning for each:

- **Slow-rolling Catastrophe** – any failure mode where an ecosystem of HLMI systems irreversibly sends civilization off the rails, due to technical misalignment causing HLMIs to pursue proxies of what we really want (the Production Web provides another example of this type of failure)
- **Correlated Automation Failure** – any failure mode where various misaligned HLMIs within a broader ecosystem develop influence-seeking behaviour, and then (through whatever means) collectively take power from humans (such HLMIs may or may not subsequently continue to compete for power among themselves, but either way human interests would not be prioritized)
- **Extreme Moloch** – a failure mode where HLMI allows humans to optimize more for what we do want (as individuals or groups), but competition between individuals/groups burns down most of the potential value

Looking at our model (and starting in the middle row), we see that all three of these failure modes depend on there *not* being a project or coalition that **achieves DSA** (since otherwise we would presumably get a **Catastrophically Misaligned HLMI** scenario). Other than that, **Slow-rolling Catastrophe (I)** and **Correlated Automation Failure (II)** share a couple of prerequisites not shared by the **Extreme Moloch** outcome. Both **(I)** and **(II)** depend on **misaligned HLMI** (for **(I)** the HLMIs will be optimizing sufficiently imperfect proxies, while for **(II)** they will be undesirably influence seeking). **Extreme Moloch**, meanwhile, does not

depend on technical misalignment, as the bad outcome would instead come from conflicting human interests (or human irrationality). Additionally, **(I)** and **(II)** would likely depend on the **economy/society [becoming] dependent on AI** (otherwise humans may be able to "pull the plug" and/or recover after a collapse); **Extreme Moloch**, meanwhile, does not require dependency, as even if civilization wasn't dependent on AI, humans could still compete using AI and burn down the cosmic endowment in the process. Societal dependence on AI is presumably more likely if there is a larger **Incentive for humans being out of the loop** and a faster **Final Speedup Factor** to the global economy from HLMI. This post discusses in more depth how the various AI takeover scenarios assume different economic incentives and societal responses to HLMI.

Additionally, each of these failure modes has its own requirements not shared with the others. **Correlated Automation Failure** would require HLMI that was **influence seeking**. **Slow-rolling Catastrophe** would require relevant **proxies [to] diverge from** what we actually want HLMI to do, to a sufficient degree that their pursuit represents an X-risk. And **Extreme Moloch** requires that in a multi-polar world, **extreme competition and optimization will burn down things of value**.

Note that existential HLMI risks other than the classic "misaligned AI singleton causes existential catastrophe" is a fast-changing area of research in AI Alignment, and there is very little consensus about how to distinguish potential failure modes from each other - see for example this recent survey. Resultantly, our model in this section is limited in several ways: the logic is more binary than we think is justified (the outcomes shouldn't be as discrete as indicated here, and the logic from cause to effects shouldn't be as absolute either), we ignore the possibility of recovering from a temporary loss of control, we don't consider the possibility of being in a world with some aligned HLMI systems and other misaligned systems, and so on. We intend to develop this section further in response to feedback and deeper research into loss of control-like scenarios.

Conclusion

To summarise, we have presented our model of some widely discussed ways that HLMI could bring about existential catastrophe, examining some of the key mechanisms and uncertainties involved. One set of paths to catastrophe involves a misaligned HLMI or HLMI-equipped group achieving Decisive Strategic Advantage. Another set of paths involves loss of control due to HLMI systems gaining influence on the world in an incremental and distributed fashion, without anyone necessarily achieving DSA. We traced these paths back to the outcomes of HMLI development (whether and how HLMI systems are aligned), and the nature of HLMI systems (e.g. whether they exhibit influence-seeking behaviour). The post Distinguishing AI Takeover Scenarios and its follow-up provide a more up-to-date and detailed understanding of AI takeover scenarios compared to our current concrete model.

In the next post, we will discuss approaches to elicit expert views that can inform the concrete model we have presented so far, and our tentative plans to carry out elicitation.

Thanks to the rest of the MTAIR project team as well as Edo Arad for feedback on drafts.

Elicitation for Modeling Transformative AI Risks

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post is part 8 in our sequence on Modeling Transformative AI Risk. We are building a model to understand debates around existential risks from advanced AI. The model is made with Analytica software, and consists of nodes (representing key hypotheses and cruxes) and edges (representing the relationships between these cruxes), with final output corresponding to the likelihood of various potential failure scenarios. You can read more about the motivation for our project and how the model works in the Introduction post. Unlike other posts in the sequence, this discusses the related but distinct work around Elicitation.

We are interested in feedback on this post, but to a greater extent than the other posts, we are interested in discussing what might be useful, and how to proceed with this. We would also welcome discussion from people working independently on elicitations, as we have discussed this extensively with other groups, many of whom are doing related work.

As discussed in previous posts in this series, the model we have built is a tentative one, and requires expert feedback and input. The traditional academic method for getting such feedback and input is usually referred to as elicitation, and an extensive field of academic work discusses how this can best be done. (As simple examples, this might include eliciting an estimated cost, and probability distribution, or a rank ordering of which outcomes from a project are most important.)

Elicitation of expert views is particularly critical in AI safety for both understanding debates between experts and representing the associated probabilities. At the same time, many elicitation and forecasting projects have the advantage of unambiguous and concrete questions with answers that will be observed in the near term, or ask for preferences about outcomes which are well understood. Because these advantages are mostly absent for AI safety questions, the focus in this project is on understanding debates (instead of attempting to settle debates that are already understood, or are not resolvable even in theory). This means that there is no intent to elicit a “correct” answer to questions which may be based on debated or disputed assumptions. For this reason, we have taken an approach designed to start with better understanding experts’ views of the domain overall, rather than focus on the outcomes directly. This leads to opportunities for better understanding the sources of disagreement.

The remainder of this post first discusses what elicitation can and should be able to accomplish in this domain, and for this project, as well as what conceptual and actual approaches we are using. This should help explain how elicited information can inform the concrete model, which then can then hopefully help inform decisions - or at least clarify why the decisions about approaches to take are disputed. Following that, we outline our tentative future plan, and what additional steps for elicitation may look like.

What to do about forecasting given uncertainties and debates?

In domains where the structure of uncertainties are clear, and not debated, it is possible to build a model similar to that built in the current project, ask experts whether the structure is correct, and based on their input, build a final Directed Acyclic Graph or other representation of the joint distribution that correctly represents their views. After this, we would ask experts to attach probability distributions to the various uncertainties, perhaps averaging their

opinions for each node in the DAG, so that we could get quantitative predictions for the outcomes via Monte Carlo.

Long-term forecasting of deeply uncertain and debated outcomes in a domain like the future of AI is, for obvious reasons, extremely unreliable for predictions. And yet, we still need to make best-guess estimates for decision purposes, and in fact we implicitly have assigned probabilities and have implicit goals which are used and maximized when making any sort of decision related to the topic. Making this explicit involves ensuring that everyone's varying assumptions or assertions are understood, which leads to both the motivation for and challenging nature of the current project.

By representing different structural assumptions about the future pathway of AI, and various models of how AI risks can be addressed, we can better understand where disagreements are due to fundamental differences ("AI will be aligned by default" vs. "The space of possible ML Minds contains at least some misaligned agents" vs. "Vanishingly few potential AIs are aligned"), and where they are due to quantitative differences in empirical estimates ("50% Confidence we will have ASI by 2030" vs. "90% confidence we won't have ASI before 2050"). While these examples may be obvious, it is unclear whether others exist which are less so - and even the "obvious" debates may not be recognized by everyone as being legitimately debated.

For this reason, in addition to understanding specific debates, we need to represent uncertainty about both quantitative estimates and about the ground truth for conceptual debates. One way we plan to address this is by incorporating confidence measures for expert opinions about the debated features or assumptions in our probability estimates. Another is accounting for the arguments from analogy which many of these claims are based upon. For example, an expectation that ML progress will continue at a given pace, based on previous trends, is not an (explicit/gears-level) model of hardware or software progress, but it often informs an estimate and makes implicit assumptions about the solutions to the debated issues nonetheless.

However, it is at least arguable that a decision maker should incorporate information from across multiple viewpoints. This is because unresolved debates are also a form of uncertainty, and should be incorporated when considering options. One way we plan to address this is by explicitly including what we call "meta-uncertainties" in our model. Meta-uncertainties are intended to include all factors that a rational decision maker should take into account when making a decision using our model, but which do not correspond to a specific object-level question in the model.

One such meta-uncertainty is the reliability of long-term forecasting in general. If we think that long-term forecasting is very unreliable, we can use that as a factor that essentially downweights the confidence we have in any conclusions generated by the rest of our model. Other meta-uncertainties include: the reliability of expert elicitations in general and our elicitation in particular, structural uncertainties in our own model (how confident are we that we got this model right?), reference class uncertainty (did we pick the right reference classes?), potential cognitive biases that might be involved, and the possibility of **unknown unknowns**.

Using Elicitations

Given the above discussion, typical expert elicitation and aggregating opinions to get a best-guess forecast is not sufficient. Several challenges exist, from selecting experts to representing their opinions to aggregating or weighting differing views. But before doing any of these, more clarity about what is being asked is needed.

What are the current plans?

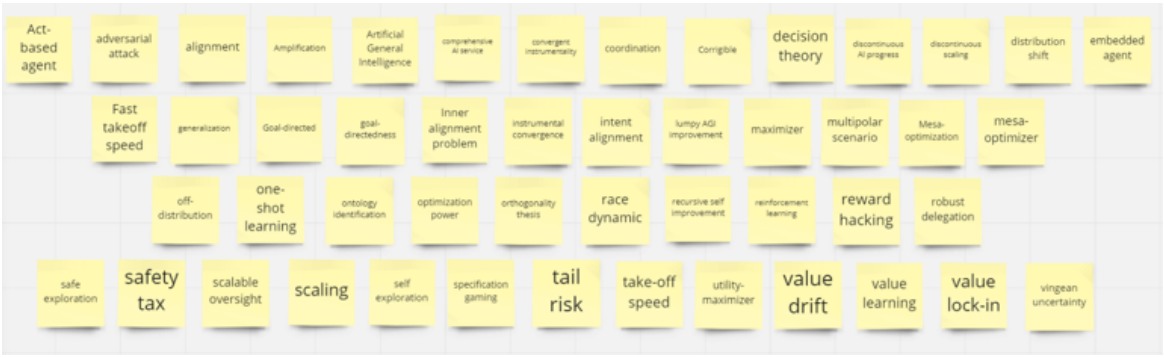
Prior to doing anything resembling traditional quantitative elicitation, we need to have clarity in what is being elicited, so that the respondents are both clear about what is being asked and are answering the same question as one another. We also need to be certain that they are answering the same question as what we think is being asked. For example, asking for a timeline to HLMI is unhelpful if respondents have different ideas of what the term means, or dispute its validity as a concept. For this reason, our current work is focused on understanding which terms and concepts are understood, and which are debated.

It seems that one of the most useful methods of eliciting feedback on the model is via requesting and receiving comments on this series of posts, and discussions that arise from it. Going further, a paper is being written which reviews past elicitation - and looks at where they succeeded or failed. Building on the review of the many different past elicitation projects and approaches, a few of which are linked, and as a way to ensure we properly understand the disagreements which exist in AI alignment and safety, David Manheim, Ross Greutzmacher, and Julie Marble are working on new elicitation methods to help refine our understanding. We have tested, and are continuing to test these methods internally, but we have not yet utilized them with external experts. As noted, the goal of this initial work is to better understand experts' conceptual models, using methods such as guided pile sorts, explained below, and qualitative discussion.

The specific approach discussed below, called pile sorting, is adapted from sociology and anthropology. We have used this because it allows for discussion of terms without forcing a structure onto the discussion, and allows for feedback in an interactive way.

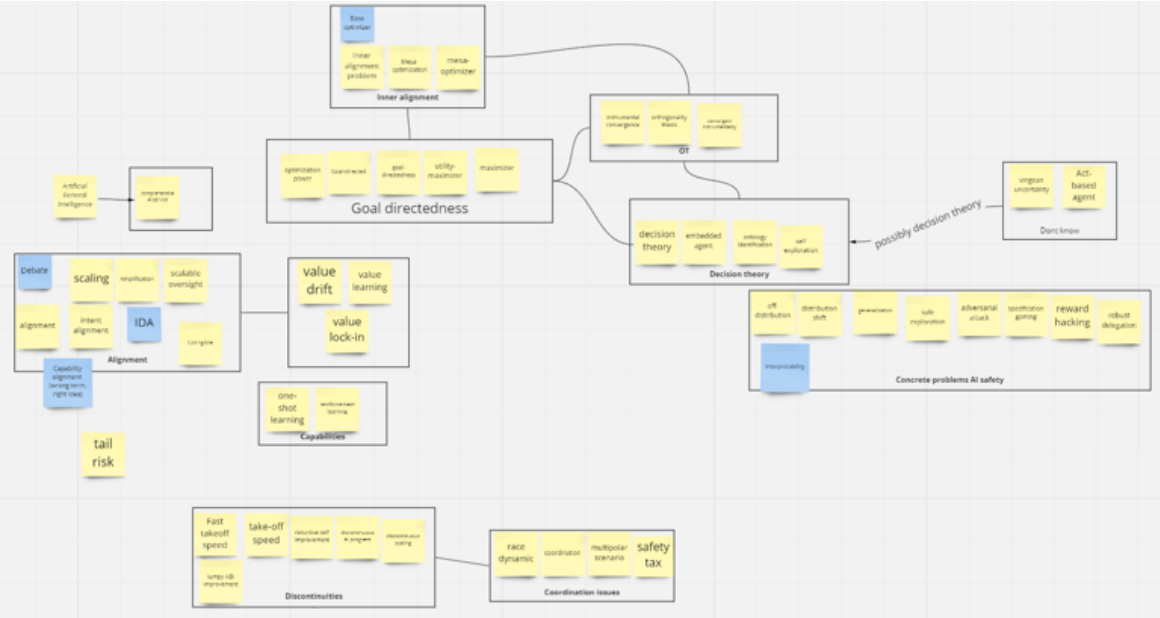
(Sample) initial prompt for the pile sorting task:

"The following set of terms are related to artificial intelligence and AI safety in various ways. The cards can be moved, and we would like you to group them in a way that seems useful for understanding what the terms are. While doing so, please feel free to talk about why, or what you are uncertain about. If any terms seem related but are missing, or there are things you think would be good to add or include, feel free to create additional cards. During the sorting, we may prompt you, for instance, by asking you why you chose to group things, or what the connection between items or groups is."



Elicitation Prompt

Based on this prompt, we engage in a guided discussion where we ask questions like how the terms are understood, why items have been grouped together, what the relationships between them are, and whether others would agree. It is common for some items to fit into multiple groups, and participants are encouraged to duplicate cards when this occurs. The outputs of this include both our notes about key questions and uncertainties, and the actual grouping. The final state of the board in one of our sample sessions looked like the below:



Example Elicitation Outcome

This procedure, and the discussions with participants about why they chose the groupings they did, is intended to ensure that we have a useful working understanding of expert's general views on various topics related to AI safety, and will be compared across experts to see if there are conflicts or different and contrasting conceptual models, and where the differences are. While methods exist for analyzing such data, these are typically for clearer types of questions and simpler sorting. For that reason. one key challenge which we have not resolved is how this elicitation can be summarized or presented clearly, other than via extensive qualitative discussions.

In addition to the card sort, we have several other elicitation approaches we are considering and pursuing that intend to accomplish related or further goals in this vein. But in order to do any elicitation, including these, there are some key challenges

How do you judge who is an "expert?"

This is a difficult issue, and varies depending on the particular hypothesis or proposition we're asking about. It also depends on whether we view experts as good at prediction, or good at proposing useful mental models that can then be predicted about by forecasters. For deconfusion and definition disputes, the relevant experts are likely in the AI safety community and closely related areas. For other questions the relevant experts might be machine learning researchers, cognitive scientists, or evolutionary biologists. And of course, in each case, depending on the type of question, we may need to incorporate disputes rather than just estimates.

For example, if we were to ask "will mesa-optimizers emerge," we need to rely on a clear understanding of what mesa-optimizers are. Unfortunately, this is somewhat debated, so different researchers' answers will not reflect the same claims. Furthermore, those who are not already concerned about the issue will likely be unable to usefully answer, given that the terms are unclear - biasing the results. For this reason, we have started with conceptual approaches, such as the above pile-sorting task.

Relatedly, in many cases, we also need to ask questions to multiple groups to discover if experts in different research areas have different views on a question. We expect different conceptual models to inform differences in opinions about the relationship between different

outcomes, and knowing what those models are is helpful in disambiguating and ensuring that experts' answers are interpreted correctly.

Another great challenge in selecting experts is that the relevant experts for topics such as AI safety and machine learning are often those who are working at the forefront of the field, and whose time is most valuable. Of course, this depends on how you define or measure domain expertise, but the value of previous elicitations is strongly correlated with the value of experts' contributions in narrow domains. The difference in knowledge and perspective between those leading the field and those performing essentially Kuhn's 'normal science' is dramatic, and we hope that the novel elicitation techniques that we are working on can enable us to weight the structure emerging from leading experts' elicitations appropriately.

Following the identification of experts, there is a critical question: Is the value of expert judgment limited to only qualitative information or to coming up with approaches in practice, rather than the alternative of being well calibrated for prediction. This is not critical at the current stage, but becomes more important later. There are good reasons to think that generalist forecasters have an advantage, and depending on progress and usefulness of accurate quantification, this may be a critical tool for later stages of the project. We are interested in exploring forecasting techniques that combine domain experts and generalist forecasters in ways intended to capitalize on the relative expertise of both populations.

How will we represent uncertainties?

For any object-level issue, in addition to understanding disputes, we need to incorporate uncertainties. Incorporation of uncertainty is both important for not misunderstanding expert views, and as a tool to investigate those differences in viewpoints. For this reason, when we ask for forecasts or use quantitative elicitations to ask experts for their best-guess probability estimates, we would also need to ask for the level of confidence that they have in those estimates, or their distribution of expected outcomes.

In some cases, experts or forecasters will themselves have uncertainties over debated propositions. For example, if asked about the rate of hardware advances, they may say that overall, they would guess a rate with distribution X, but that distribution depends on economic growth. If pre-HLMI AI accelerates economic growth, they expect hardware progress to follow one distribution, whereas if not, they expect it to follow another. In this case, it is possible for the elicitation to use the information to inform the model structure as well as the numeric estimate.

As an aside, while we do by default intend to represent both structural debates and estimates as probabilities, there are other approaches. Measures of confidence of this type can be modeled as imprecise probabilities, as distributions over probability estimates ("second-order probabilities"), or using other approaches (e.g., causal networks, Dempster-Shafer theory, subjective logic). We have not yet fully settled on which approach or set of approaches to use for our purposes, but for the sake of simplicity, and for the purpose of decision making, the model will then need to represent the measures of confidence as distributions over probability estimates.

Will this be informative?

It is possible that the more valuable portion of the work is the conceptual model, rather than quantitative estimates, or that the conceptual elicitations we are planning are unlikely to provide useful understanding of the domain. This is a critical question, and one that we hope will be resolved based on feedback from the team internally, outside advisors, and feedback from decision makers in the EA and longtermist community who we hope to inform.

What are the next steps?

The current plans are very much contingent on feedback, but conditional on receiving positive feedback, we are hoping to run the elicitations we have designed, and move forward from there. We would also be interested in finding others that are interested in working with us on both the current elicitation projects, and thinking about what should come next, and have reached out to some potential collaborators.