The Problem with Forecasting

AI governance has a forecasting problem. When compared with other domains, both experts and forecasters fail to reliably predict AI developments.

Expertise. AI experts disagree widely about AI risk and development trajectories. This is reason enough to distrust any particular expert forecast. Discussing the results of a large survey of AI experts, Grace et al.[2] note that:

> *"these experts are not accurate forecasters across the range of questions we ask. For one thing, on many questions different respondents give very different answers, which limits the number of them who can be close to the truth." (20)*

Since the experts' responses also exhibit large framing effects (variations that depend on the phrasing of the question), "even aggregate answers to any particular question are not an accurate guide to the answer." Therefore, even though AI experts' predictions might be useful in other ways, they "should not be seen as a reliable guide to objective truth [...]." (2)

Forecasting. Forecasters don't fare much better. Best practice when forecasting is to establish a base rate with respect to a reference class (a set of relevantly similar events), but there is no clear reference class for many AI x-risk questions. We can't — to take an obvious example — establish a base rate for AI x-risk by comparing how many times humanity has successfully navigated a transition to a post-TAI world to how many times it has failed.

Forecasting's success for AI risk-relevant questions is accordingly limited. As evidence, consider the 2019 "Forecasting AI Progress" tournament run by Metaculus.[3] An analysis[4] found that "Metaculus narrowly beats chance and performs worse in this tournament than on average across all continuous questions on the site [...]."

In a previous analysis[5] of forecasts for a different set of AI questions, Mühlbacher and Scoblic argued that Metaculus predictions did better than "narrowly" beat chance, but agree that "AI-related questions tend to be

intrinsically harder than many other questions."[6] Metaculus does at least have a worse track record for AI than for other categories.[7]

Forecasting with experts. One way to explain these results would be that AI experts lack training in forecasting methods, and forecasters lack AI expertise.

It was perhaps in response to this dilemma that the Forecasting Research Institute conducted a long-run tournament forecasting existential risks involving both 'superforecasters' and domain experts.[8] However, despite months of debate, the two groups' estimates failed to converge on AI risk-relevant questions. The report observes that:

> *"The most pressing practical question for future work is: why were superforecasters so unmoved by experts' much higher estimates of AI extinction risk, and why were experts so unmoved by the superforecasters' lower estimates? The most puzzling scientific question is: why did rational forecasters, incentivized by the XPT to persuade each other, not converge after months of debate and the exchange of millions of words and thousands of forecasts?" (1)*

Therefore, despite sophisticated efforts to elicit reliable information from experts and forecasters, AI remains a particularly difficult domain in which to predict the future. As a result, AI governance lacks the strategic clarity necessary to evaluate and choose between different intermediate-term options.[9]

Scenario planning. In light of the limitations of forecasting, I argue that AI governance researchers and strategists should explore an alternative and complementary approach: *scenario planning*. This is a core feature of the *AI Clarity* programs's approach at Convergence Analysis. Scenario planning is a group of methods for evaluating strategies in domains defined by uncertainty. The common feature of these methods is that they evaluate strategies across several plausible futures, or "scenarios." Scenario planning can help decision-makers and researchers prepare for several possible futures, rather than banking on a single predicted future.

Scenario planning is a complementary approach to forecasting because it can help identify the right questions to forecast. An ideal strategy would work in every scenario, but it's more likely that any given strategy will work in some scenarios

but not others. In that case, we will have to decide which subset of scenarios is more likely than the other.

In the next section, I introduce and review some methods for generating scenarios. After that, I describe some criteria against which AI scenario planners might evaluate strategies: *threat models* and *theories of victory*.

Strategic Parameters & Methods

A key tool of analysis in scenario planning are *strategic parameters*. I adopted this term from Matthijs Maas' *Concepts in Advanced AI Governance*.[10] Maas defines strategic parameters as:

> *"[…] features of the world that significantly determine the strategic nature of the advanced AI governance challenge. These parameters serve as highly decision-relevant or even crucial considerations, determining which interventions or solutions are appropriate, necessary, viable, or beneficial to addressing the advanced AI governance challenge; accordingly, different views of these underlying strategic parameters constitute underlying cruxes for different theories of actions and approaches."*

For example, a few commonly-cited strategic parameters are *timelines to TAI*, *takeoff trajectories*, and *difficulty of technical alignment*.

In addition to their content, strategic parameters can vary according to their formal properties. For example, a parameter might be a continuous variable (time to TAI), or it might be a discrete variable (paradigm of TAI).

Strategic parameters might also be more or less fixed. At one extreme, a parameter might capture a rigid feature of the world (for example, the relationship between general intelligence and agency). At the other extreme, a parameter might describe the status of active policy agendas (for example, the existence of international AI safety standards). This might also vary depending on the actor: a rigid parameter for one actor might be a lever for another.

Generating scenarios

The parameters I've mentioned so far are by no means exhaustive. In fact, there is no one correct or appropriate set. What constitutes an appropriate set of parameters depends on the scope and methods of the research in question.

Strategic parameters are sometimes referred to as "variables" or "dimensions" in scenario planning, and are used to generate "scenarios." Scenarios can be defined as combinations of parameters set to particular values (though they often also include narrative descriptions).

One way scenario planning methods are differentiated is by how they use strategic parameters to generate scenarios.

In the simplest case, a scenario planner might consider only one parameter with a few possible values. For example, in his article *Scenario Planning for an A(G)I Future*,[11] Antonin Korinek considered three values across a parameter capturing timelines to AGI. The three scenarios he identifies are 1) "business as usual," 2) AGI in ~20 years, and 3) AGI in ~5 years.

Another example of scenario planning across a single dimension is present in Anthropic's *Core Views on AI Safety*.[12] Anthropic considers the "dimension of uncertainty" describing the difficulty of technical safety across "optimistic", "intermediate", and "pessimistic" scenarios. Anthropic writes that:

> *"Our goal is essentially to develop:*
>
> 1. *better techniques for making AI systems safer,*
> 2. *better ways of identifying how safe or unsafe AI systems are.*
>
> *In optimistic scenarios, (i) will help AI developers to train beneficial systems and (ii) will demonstrate that such systems are safe. In intermediate scenarios, (i) may be how we end up avoiding AI catastrophe and (ii) will be essential for ensuring that the risk posed by advanced AI is low. In pessimistic scenarios, the failure of (i) will be a key indicator that AI safety is insoluble and (ii) will be the thing that makes it possible to convincingly demonstrate this to others."*

A slightly more complicated case would be to consider two parameters. In one common[13] scenario planning method, two extremes along two parameters are combined to create four scenarios, which can be represented as a 2×2 grid.

There's no need to consider only extremes, though: for example, we could combine Korinek and Anthropic's chosen parameters to create 9 scenarios in a 3×3 grid:

	"business as usual"	AGI in ~20 years	AGI in ~5 years
optimistic	Scenario 1	Scenario 2	Scenario 3
intermediate	Scenario 4	Scenario 5	Scenario 6
pessimistic	Scenario 7	Scenario 8	Scenario 9

With only a small number of possible scenarios, it's possible to deductively evaluate strategies across them all. As we continue to add more parameters and values, however, the problem becomes more complicated. An upper limit to straightforward deduction is usually around 4 to 6 parameters, depending on how many possible values each parameter can take.

Examples

Hua and Belfield. Hua and Belfield hit the limit of straightforward deduction in their paper "Effective Enforceability of EU Competition Law Under AI Development Scenarios." They select strategic parameters according to their role in evaluating the efficacy of EU competition law, and divide them into two categories: technical and non-technical.

Technical Variables	Non-Technical Variables
Key inputs: the distribution of importance among the input driving AI capabilities development (algorithmic innovation, computational resources, or data). [3 values] Speed of development: "the speed of AI development, measured in terms of the length of time between an arbitrary set of benchmarks" (599) [3 values] Capability: "the tasks and 'work' that can be accomplished by an AI system or collection of systems" (600) [3 values]	Number of actors: the number of actors (corporations or states) developing and deploying AI with comparable levels of capability. [3 values] Nature and relationship: "whether the actor(s) are companies/private actors or states, and whether the relationship between the actors is competitive or cooperative." (602) [6 values]

The total number of possible combinations of Hua and Belfield's parameters is 486, and it would be a herculean task to deductively evaluate a strategy across them all. So, instead of considering combinations of parameters (with the exception of "nature and relationship", which is really two parameters), Hua and Belfield evaluate the efficacy of EU competition law across each parameter separately.

However, one of the key benefits to analyzing parameters together is the possibility of combination effects. For example, consider again the 3×3 grid above. Separately, we might prefer AGI in ~20 years to AGI in ~5 years, and optimistic technical alignment to pessimistic technical alignment. But that doesn't imply that scenario 2 is preferable to scenario 3 —hypothetically, if alignment is easy, then the expected value of TAI might be positive, and we might prefer shorter timelines.[14]

It may be appropriate to consider only a small number of possible scenarios when planning for a particular actor (in Anthropic's case) or when evaluating a particular policy domain (in Hua and Belfield's case). However, larger scopes will likely call for more parameters. The addition of more parameters will in turn require new methods.

Kilian et al. In their paper, "Examining the Differential Risk from High-level Artificial Intelligence and the Question of Control,"[15] Kilian et al. list 14 strategic parameters across four categories:

Technological Evolution, Transitions, and Diffusion	AI Paradigm, Possible Accelerants & Timeline	Geopolitical Race Dynamics, Technical Ecology & Risk	International Governance, Institutional & Technical Control
Capability & Generality: "Overall power of a system to achieve objectives, influence the world, and degree of generalizability across domains." Diffusion: "The distribution of systems at the time of development and the breadth of distribution" Technological Transition: "The rate of change of the system or technology as it increases in capability"	Paradigm: "The paradigm or architecture that can achieve high-level capability and general-purpose functionality." Accelerant: "The technological insight or innovation that accelerates capabilities." Timeline: "The duration of the transition once a sufficient level of capability and generality is reached."	Race dynamics: "The economic and geopolitical dynamics of increased competition" Primary risk class: "The highest impact risks from advanced AI systems" Technical safety risk: "The technical risks from agent systems that could pose a significant danger with advanced systems."	AI Safety: "Technical safety approach to align AI systems and their ability to transfer to more general-purpose advanced systems." Actors: "The entity that leads in the development of transformative advanced AI systems." Region: "The region that leads in advanced AI capabilities or develops the first instantiation." International Governance: "The international governance bodies in place when advanced AI is developed." Corporate Governance: "The degree of coordination on safety standards by AI companies."

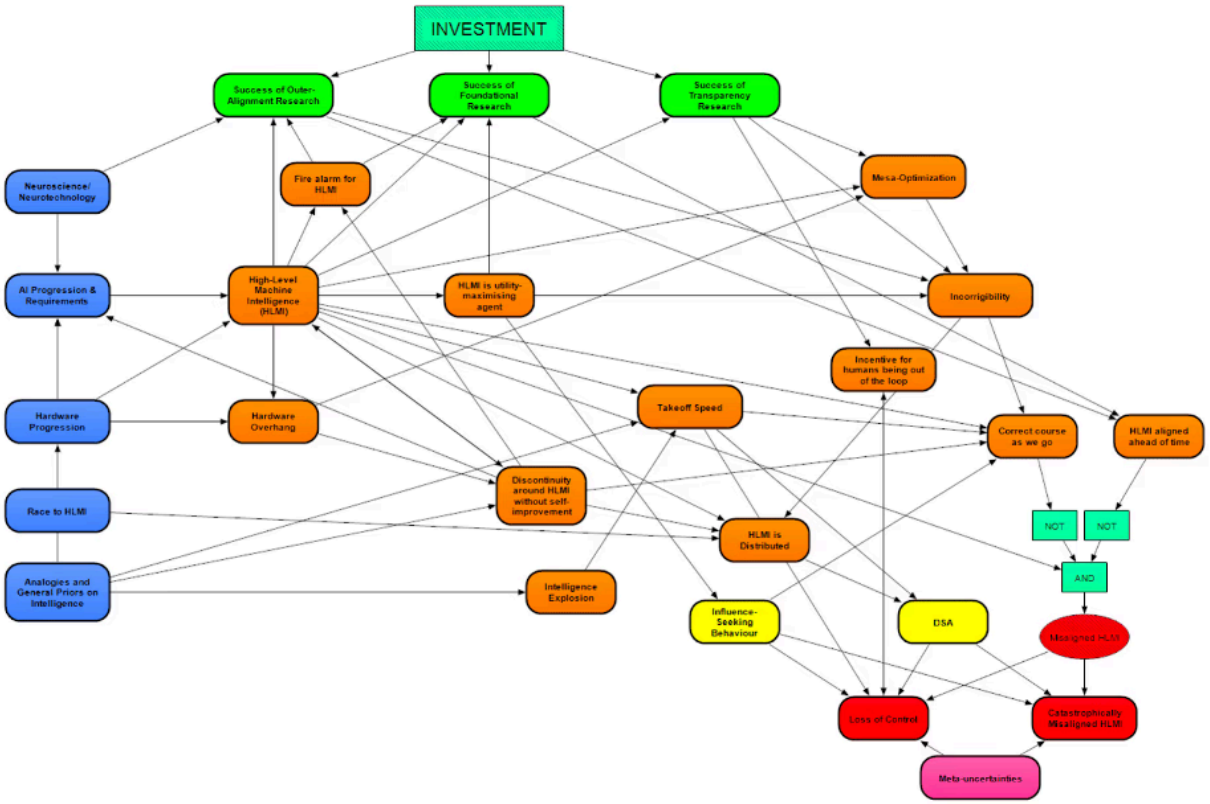
They identify a total of 47 values across these parameters, which can combine to generate 15,116,544 unique scenarios. Clearly, it would be intractable to manually evaluate a strategy across them all. Instead, Kilian et al. survey experts on their forecasts of each dimension, and then use a computational method — General Morphological Analysis (GMA) — to identify four scenarios that best capture

clusters of those forecasts. They name these scenarios “balancing act,” “accelerating change,” “shadow intelligent networks,” and “emergence,” and accompany each with a narrative description.

An advantage of Kilian et al.’s approach is that they are able to extract a deductively-tractable number of scenarios from a large number of parameters. This is possible because their method doesn’t treat each parameter as an independent variable — instead, they assume inferential relationships between parameters such that some combinations of values are plausible, and others not.

However, the data they use to construct inferential relationships are expert forecasts. Therefore, while their four scenarios might accurately describe clusters of expert forecasts, they should only be taken as predictively valuable to the extent that one takes expert forecasts to be predictively valuable.

Clarke et al. Clarke et al.’s Modeling Transformative AI Risks[16] represents a different approach to handling large numbers of parameters. Clarke et al. identify and discuss 26 parameters that, either directly or through each other, influence the probability of an existential catastrophe. Each of these parameters is itself modeled with its own set of sub-parameters.



Clarke et al. share the insight with Kilian et al. that parameters exhibit inferential relationships, but, rather than exploit that insight computationally, they map those relationships graphically.

An advantage of Clarke et al.'s method is that it directly integrates threat models (see the next section). In the previous methods discussed, additional work is necessary to evaluate strategies across generated scenarios with respect to their effect on x-risk. In this method, the effect of a strategy can be represented directly as a change to a particular parameter.

However, Clarke et al.'s method is also relatively fragile — it requires a lot of up-front work to be done correctly. It's unlikely that a straightforward inferential diagram accurately captures the dynamics influencing x-risk, which are both unprecedented and complex. For similar reasons, contemporary safety engineering has moved away from cause-and-effect and towards complex systems theory.[17]

Threat Models & Theories of Victory

In the last section, I introduced some methods for generating scenarios. This section discusses some criteria for evaluating a strategy within a given scenario. My focus is on existential risks to humanity's future, which include extinction as well as permanent human disempowerment and dystopia.

Threat models

Threat models are descriptions of and proximal pathways to existential catastrophes. They are sometimes also referred to as "hazards," "failure modes," "existential risks," or "x-risks."

Hendrycks et al. Hendrycks et al.[18] categorize threat models into "malicious use," "AI race," "organizational risks," and "rogue AIs," which they describe as "intentional," "environmental/structural," "accidental," and "internal," respectively. Similarly, Kilian et al. categorize AI risks as "misuse," "structural," "accidents," and "agential."[19]

1. **Intentional.** Intentional threat models involve intentional misuse of TAI to cause catastrophes or "lock-in" dystopic power structures. For example, narrowly-superintelligent systems might enable malicious actors to deploy artificial super-pandemics[20] or cyber attacks against critical infrastructure. They

might also enable authoritarian governments to massively surveil and control populations.

2. **Structural.** Structural threat models involve game-theoretic forces that push society towards collectively catastrophic outcomes. For example, competition between corporations or arms race dynamics between nations might lead to the development and deployment of increasingly dangerous AI systems.
3. **Accidental.** Accidental threat models involve sudden and unintentional catastrophes. AI systems — as well as the organizations and societies in which they are embedded — are complex systems, which can exhibit unpredictable and extreme interactions. For example, automated stock trading precipitated a trillion-dollar "flash crash" in 2010.[21] As AI systems become increasingly powerful and embedded in critical infrastructure, economic activity, and military activity,[22] unforeseeable accidents could have catastrophic consequences.
4. **Agential.** Agential threat models involve the possibility that the goals of agential systems may not be aligned with those of their operators. AI safety literature has argued that agential systems may be power-seeking,[23] and, if sufficiently intelligent, may gain a decisive strategic advantage over humanity. These threat models are sometimes referred to as "rogue AI" or "AI takeover."

This taxonomy likely includes all plausible threat models that describe existential catastrophes from AI. However, it also includes threat models that describe negative but not truly existential outcomes. For example, a massive cyber attack or an accidental "flash crash" would not directly threaten humanity's future. Determining which threat models are both plausible and represent truly existential catastrophes is an open research question.

Most discussions of AI x-risk consider a subset of this taxonomy. For example, Vold and Harris' review of AI x-risk cites the "control problem", "AI arms race", and the "weaponization of AI," which they describe as "accidental", "structural", and "misuses," respectively.[24] Anthropic's *Responsible Scaling Policy* is designed with only "misuse" and "autonomy and replication" in mind.[25] Many other discussions only include agential threat models.

That being said, "mere" catastrophes can't be safely ignored in x-risk scenario planning. On one hand, they might indirectly exacerbate x-risk (for example, AI-enabled misinformation isn't existential in itself, but might erode society's ability to respond well to other threats). On the other hand, they might indirectly

mitigate x-risk (for example, an AI-enabled accident might act as a "warning shot," improving society's response to AI x-risk).

Finally, expected AI x-risk mitigation is not a perfect proxy for expected value. For example, AI might improve humanity's ability to respond to other x-risks, such that the strategy which best mitigates AI x-risk is not the strategy which best mitigates all x-risk. It also doesn't capture the effect that the development of AI moral patienthood might have on expected value.

Theories of victory

Theories of victory are descriptions of and proximal pathways to averting existential catastrophes. They have been defined elsewhere as:

- "complete stories about how humanity successfully navigates the transition to a world with advanced AI,"[26] and
- "the main, high-level plan [...] for how humanity could plausibly manage the development and deployment of transformative AI such that we get long-lasting good outcomes."[27]

Theories of victory may be more useful than threat models for evaluating strategies because it's more straightforward to design a strategy to achieve a particular outcome than it is to avoid a set of outcomes.

Hobbhahn et al. There is less published work developing and categorizing theories of victory than for threat models. Below, I summarize and discuss two attempts. In the first attempt, Hobbhahn et al.[28] describe six "good TAI transition scenarios[29]":

1. **"Alignment is much easier than expected."** This scenario considers the possibility that technical AI safety is able to easily control the behavior of TAI. It is more of a hope than a plan of action. It should also be noted that even this optimistic scenario is only a theory of victory over agential threat models.
2. **"The combination of many technical and governance strategies works."** In this scenario, while no one strategy is sufficient, a combination of technical safety and governance interventions successfully complement each other, or combine to form a "defense in depth."
3. **"Accident and regulation."** In this scenario (modeled on nuclear power), a catastrophic but not existential accident spurs public outcry and strict

regulation of the development and deployment of AI systems.

4. **"Alignment by chance."** Like the first scenario, this scenario is optimistic. Unlike the first scenario, however, it assumes that the first TAI system developed is aligned by chance, even though the general problem of technical safety remains unsolved. This system is then used to solve technical safety, and effective governance ensures that no unaligned TAI systems are deployed.
5. **"US-China driven global cooperation."** In this scenario, the two leaders in AI — the US and China — cooperate to establish effective international governance of the development and deployment of AI systems. This includes both agreements setting international safety standards as well as institutionalized enforcement of those standards.
6. **"Apollo Project for AI."** The final scenario assumes that the resources necessary to develop TAI are beyond those available to corporations. TAI development is taken up by a few national governments, with one nation (perhaps the US) at a clear lead. This clear lead eliminates an arms race dynamic, and the leading nation is able to sufficiently invest in technical safety, or coordinate with other nations to end TAI development.

Räuker and Aird. Räuker and Aird[30] taxonomize theories of victory into five categories, which largely overlap Hobbhahn et al.'s scenarios:

1. **"A multilateral international monitoring & enforcement regime emerges and prevents unsafe AI development/deployment."** This theory of victory corresponds to Hobbhahn et al.'s second scenario.
2. **"The US (likely alongside allies) uses geopolitical influence to prevent unsafe AI development/deployment."** This theory of victory assumes that the US is able to unilaterally regulate AI. For example, the US might be able to successfully control the global semiconductor supply chain.
3. **"Leading corporate labs come to prioritize safety more, coordinate among themselves, and implement and advocate for various risk-reducing actions."** This theory of victory partially overlaps with Hobbhahn et al.'s second scenario. It assumes that leading AI labs have the potential to effectively self-regulate.
4. **"A single lab develops "minimal aligned AGI" and uses it to end the acute risk period."** This theory of victory corresponds to Hobbhahn et al.'s first and fourth scenarios. In an extreme form, this theory is sometimes referred to as the first aligned TAI system performing a "pivotal act."
5. **"Humanity pursues a diverse range of risk-reduction methods, ensures key institutions and norms are adaptable and competent, and 'muddles**

through.'" This theory of victory corresponds to Hobbhahn et al.'s second scenario.

Some of these theories of victory are more plausible than others. Additionally, some are action-guiding across scenarios (for example, developing strong international regulation), and others are only possible in specific scenarios (for example, technical safety is easy).

There is much more work to be done developing theories of victory. For example, theories of victory should likely be developed to meet certain conditions, such as 1) plausibility, 2) action-guidingness, and 3) sustainability.

By "sustainability," I mean that a theory of victory should ideally not reduce AI x-risk per year to a constant, low level, but instead continue to reduce AI x-risk over time. In the former case, "expected time to failure"[31] would remain constant, and total risk over a long enough time period would inevitably reach unacceptable levels. (For example, a 1% chance of an existential catastrophe per year implies an approximately 63% chance over 100 years.) If instead x-risk continued to decrease each year, then expected time to failure would increase with time, and total risk might approach a limit.

That being said, sustainability might trade off against tractability, especially in the short term. In that case, It might be best to pursue a more tractable theory of victory in the short term to "buy time" to develop and pursue a sustainable theory of victory in the longer term.

Conclusion

Scenario planning is a promising complementary approach to forecasting in AI governance. I recommend that AI governance researchers use scenario planning methods both to evaluate strategies as well as identify the right questions to forecast.

Different scenario planning methods are appropriate for different scopes. However, the methods for smaller scopes — such as scenarios involving short timelines to TAI — are more straightforward, and likely more appropriate for initial research.

There also remains much work to be done selecting and developing threat models and theories of victory. I'm particularly excited about work designing plausible,

action-guiding, and sustainable theories of victory.

Acknowledgements: Thank you to David Kristoffersson, Zershaaneh Qureshi, Justin Bullock, Elliot Mckernon, Deric Cheng, and Justin Shovelain for feedback.

NOTES

1. ^ https://forum.effectivealtruism.org/posts/M2SBwctwC6vBqAmZW/a-personal-take-on-longtermist-ai-governance#Key_bottlenecks
2. ^ https://arxiv.org/abs/2401.02843
3. ^ https://www.metaculus.com/tournament/ai-progress/
4. ^ https://forum.effectivealtruism.org/posts/x5Re9EKwGvAjZSmeb/takeaways-from-the-metaculus-ai-progress-tournament
5. ^ https://www.metaculus.com/notebooks/16708/exploring-metaculuss-ai-track-record/
6. ^ I'm not taking a stand on the definition of "narrow". The predictions on binary questions had a Brier score of 0.207. For reference, uniformly predicting "50%" on binary questions would yield a Brier score of 0.25, and omniscience would yield a Brier score of 0.
7. ^ https://www.metaculus.com/questions/track-record/
8. ^ https://forecastingresearch.org/xpt
9. ^ https://forum.effectivealtruism.org/posts/M2SBwctwC6vBqAmZW/a-personal-take-on-longtermist-ai-governance#Key_bottlenecks
10. ^ https://www.legalpriorities.org/research/advanced-ai-gov-concepts
11. ^ https://www.imf.org/en/Publications/fandd/issues/2023/12/Scenario-Planning-for-an-AGI-future-Anton-korinek
12. ^ https://www.anthropic.com/news/core-views-on-ai-safety
13. ^ https://core.ac.uk/download/pdf/288287532.pdf
14. ^ To emphasize: this is a hypothetical argument solely meant to illustrate the possibility of combination effects.
15. ^ https://arxiv.org/abs/2211.03157
16. ^ https://arxiv.org/abs/2206.09360
17. ^ https://doi.org/10.7551/mitpress/8179.001.0001
18. ^ https://arxiv.org/abs/2306.12001
19. ^ https://arxiv.org/abs/2211.03157v3

20. ^ https://www.rand.org/pubs/research_reports/RRA2977-1.html
21. ^ https://en.wikipedia.org/wiki/2010_flash_crash
22. ^ https://www.cnas.org/publications/reports/technology-roulette
23. ^ https://arxiv.org/abs/2206.13353
24. ^ https://doi.org/10.1093/oxfordhb/9780198857815.013.36
25. ^ https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf
26. ^ https://forum.effectivealtruism.org/posts/ydpo7LcJWhrr2GJrx/the-longtermist-ai-governance-landscape-a-basic-overview
27. ^ https://rethinkpriorities.org/publications/survey-on-intermediate-goals-in-ai-governance
28. ^ https://forum.effectivealtruism.org/posts/AuRBKFnjABa6c6GzC/what-success-looks-like
29. ^ Note that Hobbhahn et al are using "scenario" how I use"theory of victory" — and not in the technical sense as a combination of values across strategic parameters.
30. ^ https://rethinkpriorities.org/publications/survey-on-intermediate-goals-in-ai-governance
31. ^ https://en.wikipedia.org/wiki/Mean_time_between_failures