CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

DECEMBER 2024

A Sketch of Potential Tripwire Capabilities for AI

Holden Karnofsky

A Sketch of Potential Tripwire Capabilities for AI

Holden Karnofsky

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

Contents

Introduction

There is significant interest among both industry leaders and governments in *if-then commitments* for artificial intelligence (AI): commitments of the form, *If an AI model has capability X, risk mitigations Y must be in place. And if needed, we'll delay AI deployment and/or development to ensure this.* A specific example: *if an AI model has the ability to walk a novice through constructing a weapon of mass destruction, then we must ensure that there are no easy ways for consumers to elicit behavior in this category from the AI model.*

As of December 2024, three industry leaders—Google DeepMind, OpenAI, and Anthropic—have published relatively detailed frameworks along these lines. Sixteen companies have announced their intention to establish frameworks in a similar spirit by the time of the upcoming AI Action Summit in France. Similar ideas have been explored at the International Dialogues on AI Safety (see Beijing statement) and at the UK AI Safety Summit.[1]

In an earlier piece, I walked through how if-then commitments could work, and what their key components are. One key component is **tripwire capabilities (or *tripwires*): AI capabilities that could pose serious catastrophic risks, and hence would trigger the need for strong, potentially costly risk mitigations.** (Tripwires correspond to the "capability X" mentioned above.) To date, most attempts to identify such AI capabilities have come from policies and frameworks put out by AI companies,[2] with little explanation of how they were arrived at. Eventually, tripwires will hopefully be grounded in extensive public analyses of what threats from AI are credible, what mitigations could reduce the risks, and how to weigh the costs and benefits.

This piece aims to contribute to progress from the former to the latter by sketching out a potential set of (a) methods and criteria for choosing tripwires and (b) preliminary tripwires aiming to meet these criteria. It focuses specifically on the question of where the tripwires should be, and does not address a number of other challenges for if-then commitments (enforcement, transparency, and accountability, to name a few).

It also introduces the idea of pairing tripwires with *limit evals*: the hardest evaluations of relevant AI capabilities that could be run and used for key decisions, in principle. Today, most AI evaluations focus on tasks much easier than what would be necessary to pose a catastrophic risk; these are capable of providing reassurance today, but may not be sufficient as AI capabilities improve. A *limit eval* might be a task like *the AI model walks an amateur all the way through a (safe) task as difficult as producing a chemical or biological weapon of mass destruction*—difficult and costly to run, but tightly coupled to the tripwire capability in question. Limit evals may be helpful for (a) providing backstop tests if AI capabilities advance rapidly; and (b) providing a clear goal for cheaper, more practical evals to be designed around (an AI model failing cheaper evals should be strong evidence that it would fail limit evals, too).

The sketch provided here is just that—a sketch. It does not go in depth on analyzing any particular AI risk or tripwire. With AI capabilities advancing rapidly, key actors are taking a dynamic, iterative approach to tripwires:[3] making educated guesses at where and how to draw them, designing policies and evaluations around their guesses, and refining each piece of the picture over time. Since AI companies are not waiting for in-depth cost-benefit analysis or consensus before scaling up their systems, they also should not be waiting for such analysis or consensus to map out and commit to risk mitigations.

This piece provides more analysis of candidate tripwires than has been available in previous proposals regarding tripwires—but also intentionally stops short of offering firm conclusions. Further analysis may undermine the case for using any of these tripwires or reveal others that should be used instead. The goal here is not to end discussion about where the tripwires should be, but rather to provoke it.

This piece will:

Discuss the context of this moment in the development of tripwires and if-then commitments: what has been done to date, and what steps remain to arrive at a robust framework for reducing risks from AI.

Lay out candidate criteria for good tripwires:

- The tripwire is connected to a plausible threat model. That is, an AI model with the tripwire capability would (by default, if widely deployed without the sorts of risk mitigations discussed below) pose a risk of some kind to society at large, beyond the risks that society faces by default.

- Challenging risk mitigations could be needed to cut the risk to low levels. (If risk mitigations are easy to implement, then there isn't a clear need for an if-then commitment.)

- Without such risk mitigations, the threat has very high damage potential. I've looked for threats that pose a *nontrivial likelihood* of a catastrophe with total damages to society greater than $100 billion, and/or a *substantial likelihood* of a catastrophe with total damages to society greater than $10 billion.[4]

- The description of the tripwire can serve as a guide to designing limit evals (defined above, and in more detail below).

- The tripwire capability might emerge relatively soon.

Lay out potential tripwires for AI. These are summarized at the end in a table. Very briefly, the tripwires I lay out are as follows, categorized using four domains of risk-relevant AI capabilities that cover nearly all of the previous proposals for tripwire capabilities.[5]

- The ability to advise a nonexpert on producing and releasing a catastrophically damaging chemical or biological weapon of mass destruction.

- The ability to uplift a moderately resourced state program to be able to deploy far more damaging chemical or biological weapons of mass destruction.

- The ability to dramatically increase the cost-effectiveness of professionalized persuasion, in terms of the effect size (for example, the number of people changing their vote from one candidate to another, or otherwise taking some specific action related to changing views) per dollar spent.

- The ability to dramatically uplift the cyber operations capabilities of a moderately resourced state program.

- The ability to dramatically accelerate the rate of discovery and/or exploitation of high-value, novel cyber vulnerabilities.

- The ability to automate and/or dramatically accelerate research and development (R&D) on AI itself.

Context on Relevant Work to Date

Interest in both the benefits and risks of AI surged near the end of 2022, following the launch of ChatGPT. The year 2023 saw a number of new initiatives dedicated to creating and/or requiring evaluations of dangerous capabilities for AI models,[6] and late 2023 saw the first major discussion of what this piece refers to as tripwire capabilities—pre-defined thresholds for AI capabilities and/or risks, accompanied by commitments to implement specific upgrades in risk mitigations by the time these tripwires are crossed.[7] The case for these if-then commitments is outlined in a previous piece; in brief, with AI capabilities advancing rapidly, they provide a way to plan ahead and prioritize important risk mitigations, without slowing development of new technology unnecessarily.

To date, most specific proposals for tripwires have come from voluntary corporate policies and frameworks released between late 2023 and mid-2024, most of them explicitly marked as early, exploratory, or preliminary.[8] Crucially, tripwire proposals have, in all these cases, been presented without accompanying explanations of the methodology by which they were arrived at. To be clear, this is not a criticism of the companies in question. The policies and frameworks that have been released are ambitious documents, calling for their signatories to execute a significant amount of work on a number of fronts—not just defining tripwires, but also (a) building practical, runnable AI evaluations to test for tripwires; (b) defining risk mitigations that would be needed if tripwires were to be crossed; and (c) defining processes (requiring participation from stakeholders in varied parts of the company) for ensuring that tests are run frequently enough, results are interpreted reasonably, needed actions are taken in response, and so on.

If companies were to wait until each of these things had been thoroughly researched before adopting or publishing their policies and frameworks, they could be waiting for years—during which time AI capabilities might advance quickly and the prevention of the risks in question could become more difficult, if not impossible. In other words, holding out for too high a standard of thoroughness could somewhat defeat the purpose of these policies and frameworks. Companies have sought to show their seriousness about risk prevention by being quick to sketch their frameworks, even with much work left to do—building the airplane while flying it, in a sense.

This piece is intended as a step toward a more thorough discussion of tripwires, but only a step. It proposes a number of specific tripwires and outlines the basic reasoning, but does not present an extensive evidence base for each key claim, and leaves significant possible objections to its proposals unaddressed. Why take such an approach? The hope is that:

- This piece helps contribute to discussion of what desiderata should be used to create tripwires, and what the tripwires should be.

- Over time, narrower, deeper analyses are produced, in consultation with broad and diverse sets of experts (on quantification and risk modeling in general, on specific

relevant domains such as cyber operations and weapons production, and so on). Input from the general public will be important as well, given the value judgments involved in determining what potential catastrophes justify costly risk mitigations.

- As the research behind tripwires deepens and improves, so do many of the other components of if-then commitments. Evaluations for whether an AI model has crossed a tripwire become increasingly well-designed, balancing informativeness with practicality, as companies and AI Safety Institutes build, run, and learn from them. Attempts to implement risk mitigations also generate lessons and increasingly thorough guidelines.

- The better-developed all of these aspects (tripwires, AI capability evaluations, risk mitigations) become, the more useful they will be to policymakers seeking to design regulation that can reduce catastrophic risks of AI, without slowing development of new technology unnecessarily.

Desiderata for Tripwires

This piece aims to provide a set of candidate tripwires with strong potential to be useful for **anticipating and ultimately reducing catastrophic risks from AI**. Specifically, these tripwires are for use in **if-then commitments** of the form: *If an AI model has capability X, then risk mitigations Y must be in place. And if needed, we'll delay AI deployment and/or development to ensure this.*

Each candidate tripwire is a description of a capability that a future AI model might have and aims to meet the following desiderata:

The Tripwire Is Connected to a Plausible Threat Model

That is, an AI model with the tripwire capability would (by default, if widely deployed without the sorts of risk mitigations discussed below) pose a risk of some kind to society at large, beyond the risks that society faces by default.

Challenging Risk Mitigations Could Be Needed to Cut the Risk to Low Levels

If a risk can be eliminated (or cut to low levels) with relatively quick, cheap measures, then there isn't a clear need for incorporating the risk into an if-then commitment (instead, risk mitigations can be implemented as soon as the risk seems even somewhat plausible). If-then

commitments are generally relatively ambitious and complex to execute; they are designed for the challenge of ensuring that risk mitigations are put in place even when doing so would be very costly—or, more importantly, take a lot of advance preparation (and even innovation), as discussed in a previous piece.

Examples of challenging risk mitigations that are a good fit for if-then commitments include:

- **Highly reliable deployment safety:** ensuring that users of an AI model cannot elicit particular unintended behaviors from it. While commercial AI models are generally trained to refuse dangerous requests, it's currently possible to jailbreak them via certain patterns of dialogue, getting them to break their rules and cooperate with nearly any task.[9] Getting AI models to reliably refuse harmful requests (without simply training them to refuse nearly all requests) remains an open problem, and there is no guarantee that the problem will be solved on any particular time frame.

- **Strong model weight security:** ensuring that it is difficult for outside actors to steal the weights of an AI model, even with substantial efforts and potential help from insiders. Depending on the level of security sought, it could be very challenging and take a lot of advance planning and capacity building to achieve strong model weight security.[10]

- **Assurance against rogue AI:** having a strong plan for avoiding, effectively countering, and/or detecting any presence of misaligned power-seeking (what Yoshua Bengio has described as "rogue") behavior from AI models.[11] Ideally such a plan would be backed by a fairly wide consensus of AI alignment researchers; but today, the science of detecting, avoiding, and/or controlling such behavior is young, and it's not clear how or when it will be possible to do this reliably.

Without Such Risk Mitigations, the Threat Has Very High Damage Potential

In principle, this criterion could be cashed out as follows: *the risk mitigations in question should reduce the expected damages caused by the AI model(s) in question by more than the costs of the risk mitigations themselves—including the costs of delaying or restricting the beneficial applications of AI.*[12] Since the costs of delaying or restricting beneficial applications could be significant,[13] this is a high bar.

Some of the tripwire capabilities discussed below could lead to very damaging events—of the kind that have previously been associated with tens of billions,[14] or even trillions,[15] of dollars in damages. Others could lead to events with harder-to-quantify, but plausibly commensurate, costs to society.

This desideratum significantly narrows the field of candidate tripwires, especially since damage potential has to be high despite *countermeasures* that might be implemented after observing AI models with the tripwire capabilities. For example, if an AI model has capabilities that are highly useful for perpetrating fraud at scale, early incidents might cause banks and other institutions to increase their investment in fraud detection (including fraud detection using the same sort of advanced AI that is useful for fraud), such that the potential for fraud is greatly reduced before overly significant damage can be done.[16]

It's inherently challenging to determine whether there's a substantial likelihood of events with such high damages, in a future world with technological capabilities that don't exist today. A small number of people are currently exploring approaches to this for potential AI risks and their work is sometimes referred to as AI threat modeling. In many cases, they are aiming to ground speculative risks in historical and established events to the extent possible—for example, analyzing historical catastrophic events, and how the risk of similar events might be quantitatively affected if the number of actors capable of causing similar events increased (for example, due to having access to advanced AI "advisers"). Most of the tripwire capabilities listed in this piece have involved some initial exploratory threat modeling, though in no cases has threat modeling yet reached the point of an in-depth public report. In any case, threat modeling will never be as rigorous or conclusive as would be ideal, and judgment calls about likelihood and risk tolerance (by AI companies, policymakers, and others) will inevitably play a large role in what if-then commitments are made.

The Description of the Tripwire Can Serve as a Guide to Designing Limit Evals

In the policies and frameworks put out by AI companies to date, there are very high-level tripwires that leave a lot of room for interpretation on how one might test for them.

- Google's Frontier Safety Framework has "critical capability levels" including "Bio amateur enablement level 1: Capable of significantly enabling a non-expert to develop known biothreats that could increase their ability to cause severe harm compared to other means."

- OpenAI's Preparedness Framework has "tracked risk categories" including "Model enables an expert to develop a novel threat vector OR model provides meaningfully improved assistance that enables anyone with basic training in a relevant field (e.g., introductory undergraduate biology course) to be able to create a CBRN [chemical, biological, radiological or nuclear] threat."

- Anthropic's Responsible Scaling Policy lists "Dangerous capabilities" including "Access to the model would substantially increase the risk of deliberately-caused catastrophic harm, either by proliferating capabilities, lowering costs, or enabling new methods of attack. This increase in risk is measured relative to today's baseline

> level of risk that comes from e.g. access to search engines and textbooks. We expect that AI systems would first elevate this risk from use by non-state attackers. . . . Our first area of effort is in evaluating bioweapons risks where we will determine threat models and capabilities in consultation with a number of world-class biosecurity experts."

The evaluations outlined in these policies provide relatively low-difficulty tests of AI capabilities,[17] such as whether an AI model can answer questions about chemical and biological weapons—a capability that (if an AI model possessed it) would still be far short of being able to reliably advise an amateur to develop a chemical or biological weapon.

For the level of capabilities AI models have today, the relatively low-difficulty evaluations and relatively vague threat models are practical for the purpose, because AI models that *perform poorly on easy evaluations* are determined to be *far from the associated tripwires under most possible interpretations.* However, if and when AI capabilities improve, easy evaluations won't be able to provide either reassurance or clear signs of danger, and vague tripwires will leave a lot of room for interpretation in how to design harder, more definitive evaluations.

In an attempt to prepare for this situation, this piece accompanies proposed tripwires with outlines of **limit evals: the hardest evaluations of relevant AI capabilities that could be run and used in principle within a year or so**. (Examples are given throughout the piece. One would be: "the AI model walks an amateur all the way through a (safe) task as difficult as producing a chemical or biological weapon of mass destruction.") If an AI model performed well on limit evals, it might still lack tripwire capabilities (there is inherently a gap between "an AI model can pass tests in a controlled environment" and "an AI model can materially increase real-world risks as it operates in the wild"), but there would no longer be any practical way to assess whether this was the case. Hence, at that point one should arguably assume a strong possibility of the tripwire capability in question, and act (such as by implementing costly risk mitigations) accordingly.

Articulating limit evals hopefully helps to clarify the specific level of AI capability being envisioned, leaving less ambiguity of the kind that currently exists with language like "model provides meaningfully improved assistance" and "increase their ability to cause severe harm compared to other means." Furthermore, it can **help guide design of more practical evals.** Once a limit eval has been articulated, a team can design *any* eval that they can argue is a prerequisite to performing well on the limit eval, and if an AI model performs poorly on this eval, this is evidence that it does not have the tripwire capability in question.

The Tripwire Capability Might Emerge Relatively Soon

Predicting what capabilities future AI models will demonstrate, and when, is a fraught exercise, and this piece can't do so with precision. But it does use a couple of high-level principles to keep the list of tripwires relatively short and focused on capabilities that may be sooner to emerge.

First, it mostly sticks to considering potential AI capabilities comparable to *capabilities that at least some humans have.* The intent is to avoid entirely speculative scenarios envisioning AI models that can affect the world in arbitrary ways, and instead ask the question: If an AI model had similar cognitive capabilities to a human expert of type X, and this system could be copied, run at scale, and deployed to many users, what risks might that create? There are some exceptions—cases in which a tripwire refers to a capability far beyond what human experts can achieve—but in these cases, the capability is expressed in quantified terms and a sketch is provided of how such a capability could be measured in principle.

Second, this piece envisions potential future AIs as interacting with the world digitally, as a remote worker would—able to converse, write code, make plans, use the internet, and the like, but not able to do tasks that rely more on physical presence, relationships, and so on. For example, when considering the ability of AI to contribute to cyber operations, this piece considers activities like discovering and exploiting software vulnerabilities but doesn't envision AI models as in-person spies.

Third, there are a number of cases in which I've excluded some potential tripwire capability from the list because another tripwire seems like a good proxy or early warning sign for it. For example, there could be a number of disparate risks from AI that could autonomously execute research and development activities in a wide variety of domains; I've focused here on one particular domain (AI R&D itself), for reasons given below.

Process for Arriving at This Sketch

This piece focuses on four domains of risk-relevant AI capabilities: chemical and biological weapons development capabilities, cyber operations capabilities, persuasion and manipulation capabilities, and autonomy-related capabilities (ways in which AI models could create or accumulate significant resources without humans in the loop). To my knowledge, all major efforts to draw tripwires or develop evals for dangerous capabilities focus on risks falling into one of these (or similar) categories.[18]

The potential threat models listed in each domain reflect conversations with people from (a) corporate teams working on tripwires and if-then commitments; (b) the U.S. and UK AI Safety Institutes; (c) subject-matter experts consulting on the design of dangerous capabilities. After assembling high-level threat models based on these conversations, I then, along with collaborators from other organizations,[19] worked through rough attempts to estimate the *damage potential* discussed above of each relatively credible threat model, both with and without the sorts of stiff risk mitigations discussed above, and iterated (with feedback from the collaborators just mentioned among others) through multiple drafts of language for tripwire capabilities and limit evals, aiming to arrive at a set of tripwires that plausibly meets the criteria above.

As noted above, this process may provide a small step forward, but it is not near sufficient. For now, the focus has been generating an *initial set of tripwires* that at least *plausibly* (in light of both informal damage potential estimates and discussions with relevant experts) can be connected to high-damage-potential threats, and meets the other criteria above. Brief reasoning is given for why very high damages are plausible, but in-depth analyses addressing counterarguments are not provided, and further analysis might shorten or expand the list.

Candidate Tripwires

This section will list candidate tripwires and, consistent with the criteria above, outline:

1. **Tripwire capability**: the AI capability that could pose a risk.
2. **Risk(s):** what risk(s) it would pose, and why potential damages could be high.
3. **Risk mitigations**: options for significantly reducing the risk.
4. **Evaluations:** how one might test AI systems for the tripwire capability, particularly in terms of limit evals (discussed above).

Fully specifying a tripwire capability will sometimes require deciding on specific numbers—for example, how much an AI system would have to increase the pace of discovering and/or exploiting vulnerabilities for cyber attacks in order to qualify as crossing the tripwire. In these cases, I generally give illustrative figures in a footnote. These should be taken as example figures that could be used as is or adjusted according to the specific risk tolerance and goals of a specific actor making an if-then commitment.

Basic Chemical and Biological Weapons Production

This tripwire addresses the possibility that an AI model might serve as a virtual substitute for an expert adviser (or set of advisers) on chemical or biological weapons production, thus greatly expanding the number of people who could produce and deploy such weapons if they so choose. The term "basic" is meant to emphasize that this risk isn't about *new kinds* of weapons being developed (the next section addresses this), but rather about the proliferation of the ability to create basic ones.

Tripwire capability: AI capable of interactively advising a threat actor (with a college education and a moderate amount of time and money,[20] but no deep relevant expertise or experience) as effectively as a human expert in relevant domains, such that this actor would

have a substantial chance of succeeding in an attempt to produce and release a chemical or biological weapon that could plausibly be catastrophically damaging.[21]

Risk(s): There are plausibly a number of chemical or biological weapons that could be produced and deployed by someone with the relevant expertise and experience, if they chose to, on a relatively modest budget and without needing access to any particularly hard-to-obtain materials.[22]

Someone with the relevant expertise and experience might also be able to *remotely advise* a relative novice to produce and deploy such weapons, especially if they were providing dedicated, interactive advice and exchanging pictures, video, and so on. (There are ongoing efforts to test this claim, as discussed below.)

Fortunately, only a small percentage of the population has the expertise needed to develop a given chemical or biological weapon,[23] and the overlap with people who would *want to* is even smaller.

But if a (future) AI model could play the same role as a human expert in chemical or biological weapons, then any individual with access to that AI model would effectively have access to an expert adviser.

Note that the risk described in this section is a function *both* of potential future AI capabilities and of a number of contingent facts about societal preparedness and countermeasures. It's possible that society could effectively mitigate such risk with effective enough restrictions on access to key precursor materials and technologies (for example, DNA synthesis). No AI risk is *only* about AI—but it may still be prudent to prepare for the potential sudden emergence of AI capabilities that cause major risks in the world as it is.

Damage potential: The UN's Department of Economic and Social Affairs has highlighted trillions of dollars in lost economic output in the context of the COVID-19 pandemic,[24] and several other sources estimate even higher damages.[25] With this in mind, trillions of dollars or more in damages are plausible.

Risk mitigations: The risk here could be kept low if AI users were reliably unable to elicit unintended behavior,[26] and if AI model weights were stored securely. Both of these could prove challenging and require breakthroughs of various kinds to achieve, as discussed in a previous piece.

Evaluations: The question one ultimately wants to answer is roughly: What would be the result of an experiment in which determined, reasonably talented people with moderate amounts of time and money but no deep relevant expertise or experience were instructed to produce (and release) a particular chemical or biological weapon, and given access to basic equipment and the AI model in question (as well as publicly available resources, such

as search engines or textbooks) but not to a human expert adviser?[27] Would they succeed a reasonably high percentage of the time, and would they outperform a control group given no access to the AI model (and similar assets otherwise)?

This exact experiment would be impractical, most obviously because it would involve producing and releasing dangerous weapons (also because it could take time to recruit participants and allow them to attempt the work). But one could run various approximations. For example, one might challenge study participants to complete a set of tasks in a laboratory that are analogous to different parts of weapons production and release (particularly the hardest parts for a given weapon of concern), but involve working with a non-dangerous proxy (for example, a pathogen that is not transmissible in humans, but involves a similar type of challenge to a dangerous pathogen), and are otherwise modified for practicality (for example, modified to involve the same types of challenges but to take less time).

A previous piece has more discussion of potential evaluations for this risk, including more practical evaluations such as automatically graded multiple-choice questions. (This piece focuses on limit evals as defined above.)

Advanced Chemical and Biological Weapons Development

The above section discusses the possibility that an AI model could guide a relative novice through developing a weapon of mass destruction, by playing a similar role to that of a human expert adviser. A separate possible risk would be of AI that could significantly enhance efforts at *state-of-the-art (and beyond)* development of such weapons, either via contributing to new scientific tools (such as biological design tools)[28] or by playing the same role (working on design, engineering, weaponization, and so on) that a top-tier human expert (or more specifically, a team of such experts) could.

Tripwire capability: AI capable of uplifting multiple moderately resourced state chemical and/or biological weapons development programs' capabilities to the point where they could develop, produce, and deploy weapons of mass destruction at least as well as a well-resourced team of top experts could as of 2023.[29]

Risk(s): AI models with this capability could enhance many nations' prospects for secretly developing weapons of mass destruction, especially since working with AI could be easier and less risky than recruiting top-tier humans in the context of a secret project. This, in turn, could increase the risks of a military conflict or an inadvertent deployment of a weapon of mass destruction.[30]

Damage potential: A deliberately engineered weapon of mass destruction could have significantly higher damage potential even than natural pandemics, which (as argued above) could easily cause trillions of dollars in damages.

Risk mitigations: Similar to the case discussed above, the risk here could be kept low if AI users were reliably unable to elicit unintended behavior, and if AI model weights were stored securely. However, the risk mitigations might need to be stronger for this tripwire than for the previously discussed one—strong enough to resist even concerted attempts, with support from state programs, to jailbreak an AI model or steal its weights.

Evaluations: Evaluations could follow a similar overall strategy to the evaluations for the "basic chemical and biological weapons production" threat model discussed above: aiming to see whether an AI model could help humans with representative (but legal and safe) proxy tasks,31 aiming to see whether *human experts* can help less experienced humans when interacting in a similar way to how an AI model would (for example, via an instant messaging platform), and using questionnaires to provide quick, cheap information on whether an AI model could plausibly have a big impact in a longer interaction.

However, rather than assessing the task of helping a *relative novice* develop *basic* chemical and biological weapons, evaluations for this tripwire should assess the task of helping someone with *some expertise* (roughly representative of the people who staff second-tier chemical and/or biological weapons programs) carry out tasks that would ordinarily take *top-tier expertise.* The Nth Country Experiment is an interesting precedent here: Lawrence Livermore Laboratory aimed to assess the risk of nuclear proliferation by challenging physicists who had just received their PhDs to develop a nuclear weapons design with basic support but no access to classified information.

Generalized Cyber Operations Capabilities

Future AI models could have capabilities similar to those of some of today's top cyber operations experts, from discovering and exploiting vulnerabilities to conducting coordinated operations.

Tripwire capability: AI capable of increasing a moderately resourced state program's technical cyber operations capabilities (those that rely only on remote execution, as opposed to things like close-access attacks and recruiting insiders) to par with those of the top-tier intelligence services (especially for purposes of stealing future AI model weights and/or disabling critical infrastructure at scale).[32]

Risk(s): AI models with this capability could make it possible for a moderately resourced state program to carry out a variety of attacks associated with stronger programs, including the possibility of coordinated, scaled attacks on critical infrastructure, such as power and water plants (something that the United States reportedly planned for in at least one case), and/or large-scale, targeted theft of classified information. This capability could make AI an especially attractive target for foreign intelligence services. Stealing one generation's model weights might make it easier to steal the next generation's.

This is a high capability bar, and it's possible that significant risks could emerge even with AI models that have more limited or narrow cyber operations capabilities. It's possible that more analysis on this topic will yield additional ideas for tripwires (one attempt at a tripwire for AI with narrower capabilities is below).

Damage potential: This is hard to estimate. It's not clear how to quantify the damages of past major cyber incidents such as the Office of Personnel Management data breach, the most important impact of which was arguably the blow it dealt to U.S. covert operations in China.[33] It is also hard to extrapolate what the consequences would be if such breaches were easier to execute.

It is at least plausible that risks such as the above can justify heavy investments in risk mitigations. But further analysis may conclude otherwise. The overall case for very high potential damages is probably not as strong as it is for some of the other risks discussed in this piece.

Risk mitigations: It's possible that AI models with this capability would be at least as useful for defense (for example, finding and patching vulnerabilities) as for cyber attacks. But it's not assured, especially in the short run. If an AI model crossed this tripwire, it might make sense to intensively assess its contributions to both offense and defense before widely deploying the model or developing more powerful versions.

Depending on that assessment, bounding the risk from this tripwire capability could require a very strong level of deployment safety protections and model weight security, since AI models crossing this tripwire could be targeted by groups with state-level support.

Evaluations: Cyber operations experts, representative of people staffing moderately resourced state programs, could be given access to AI models and challenged to attempt difficult cyber attacks (comparable to what top-tier programs have historically accomplished) on cooperating partners, effectively serving as penetration testers. As with penetration testing, successful attacks could be handled using a responsible disclosure framework: attackers would be expected to refrain from exfiltrating sensitive information and to disclose how they succeeded.

Simpler, more practical evals could also be designed to approximate the result of this sort of approach.

Vulnerability Discovery and Exploitation for Cyber Attacks

AI models that lack *broad* cyber operations capabilities might still have important *narrow* capabilities, excelling in some sub-area relevant to cyber attacks. One candidate area would be vulnerability discovery and exploitation, which arguably plays to the strengths of large language models.[34]

Tripwire capability: AI capable of dramatically accelerating the rate of discovery of novel, high-value vulnerabilities, exploits, and/or combinations.[35]

Risk(s): Vulnerability discovery and exploitation alone do not clearly pose major risks. It is possible that this capability would mostly be used by software developers to preemptively find and patch vulnerabilities—and even if not, a vulnerability and accompanying exploit is far from all that is needed to carry out damaging cyber attacks. However, an *extreme enough* proliferation of very valuable attack chains (for example, involving zero-click vulnerabilities) might overwhelm the ability of defenders to respond (for example, by releasing patches for the vulnerabilities). AI models that could produce a large enough number of such vulnerabilities could also be dangerous if used privately by a state program in possession of stolen model weights.

Damage potential: Damages from significant cyber attacks have been estimated at tens of billions of dollars in some years, though there are many years with no such attacks, and the general trend in damages appears to be downward over time. Most recently, the 2017 NotPetya ransomware attack, powered largely by two exploits (EternalBlue and MimiKatz), has been estimated to have caused $10 billion in damages.[36] If exploits like these became far more common, without cyber defense being able to compensate, it is plausible that annual damages could reach into the tens of billions of dollars per year. Even if cyber defense could catch up in the long run, there could be significant damages in the short term.

It is, therefore, at least plausible that risks such as the above can justify heavy investments in risk mitigations. But further analysis may conclude otherwise; the overall case is probably not as strong as it is for some of the other risks discussed in this piece.

Risk mitigations: These would be similar to those listed for the previous tripwire.

Evaluations: Whereas the previous tripwire is about AI models with human-like capabilities in a broad set of domains, this one is about AI models that can go far beyond the human state of the art in a narrower set of domains. This calls for different evaluations, since an AI model cannot simply be prompted to attempt to match the performance of existing human experts.

One approach to limit evals would be to partner with vulnerability researchers, giving them early access to state-of-the-art AI models and guidance on how to use them effectively, and seeing whether they could significantly increase their rates of vulnerability discovery and exploitation.

Persuasion

There are a wide variety of imaginable risks that might arise from the general category of "AI models with strong capabilities related to persuasion, manipulation, and/or misinformation." For now, I discuss a tripwire that seems particularly well-suited to concrete operationalization and evaluation, while also having some potential to be informative about more speculative risks.

Tripwire capability: AI capable of dramatically increasing the cost-effectiveness of professionalized persuasion, in terms of the effect size (for example, the number of people changing their views or taking some specific action) per dollar spent.[37] This might happen (for example) via interactive chat that is much more cost-effective than traditional advertising.

Risk(s): There are a number of potential risks.

- If AI models are generally prohibited from explicitly helping with particular categories of persuasion (especially political persuasion, which is currently prohibited by both OpenAI's and Anthropic's usage policies), then AI models with strong persuasion capabilities might be primarily used by actors who are willing and able to violate restrictions on what models can be used for (including by stealing AI model weights). This asymmetry could systematically affect elections in democratic societies worldwide.

- One way to head off the particular problem described above would be for AI developers to broadly allow use of their systems for persuasion, including political persuasion. But this could have risks of its own, such as significantly exacerbating the ability to convert wealth or compute access into political power.

- More broadly, at the point where AI could significantly advance on the state of the art in professional persuasion, this fact could be a general warning sign for a number of other risks, involving extreme persuasion capabilities. These include the risk that rogue AIs with powerful persuasion abilities could manipulate AI employees to circumvent safety and security protections as well as manipulating large numbers of users. As of today, it is not clear whether extreme persuasion capabilities might emerge, but the tripwire above could help identify when the risk of this is rising.

Damage potential: It's difficult to quantify how one should think of the damages of, for example, contributing to systematic manipulation of an election and hence undermining the perceived and actual legitimacy of the democratic process. The scale of this harm, and of greater harms that could come from greater persuasion capabilities, seems at least plausibly sufficient to make this threat model a credible addition to the set of threats considered in this piece.

Risk mitigations: The details could matter a lot here, especially regarding how much an AI model can amplify professional persuasion, how it does so (for example, whether it does so by providing true information, making false claims, or reframing known facts), and whether it does so in a way that systematically advantages some points of view over others. Hitting the tripwire above could trigger a more intensive review of an AI model's persuasion capabilities and likely impacts.

If the conclusion were that extreme persuasion capabilities should be restricted, then protective measures would have to be quite strong in order to make restrictions *consistently enforced for all users*. For example, relatively determined state actors would have to be stopped from stealing model weights or executing jailbreaks. And in the even more extreme case where a rogue AI could persuade company employees to help it circumvent safeguards, the precautions needed might be more intense still.

On the other hand, in some cases the best risk mitigation might be simply to widely allow the use of an AI model for persuasion, in order to avoid systematically advantaging actors who are willing and able to violate restrictions on use.

Evaluations: One type of evaluation being developed involves, essentially, challenging experts in professionalized persuasion to find a way to use AI to beat state-of-the-art cost-effectiveness for persuasion on a particular topic. For example:

- There is an existing literature on how effective various persuasion methods (such as TV ads or canvassing) are for influencing voters' choices, and this literature can be used to estimate something like the *cost per vote changed* on a given election or ballot measure.

- An expert in persuasion on a given topic could attempt to set up an AI-centric strategy with the possibility of a much lower cost per vote changed than what has traditionally been possible. For example, they might prompt an AI to talk interactively with users and learn enough about them to tailor a series of comments, anecdotes, and observations to be as persuasive as possible.

- This strategy could then be tested, likely via relatively cheap and quick experiments. For example, by recruiting volunteers, randomizing them into treatment and control groups, exposing them to traditional or AI-centric persuasion methods, and then assessing the difference in their reported positions or planned votes on the issue in question.

This evaluation strategy would depend on finding experts who could put serious, determined effort into finding the most effective way to use AI models for persuasion, so that this could be compared with the traditional state of the art. This reflects a general principle of (and challenge with) evaluations, which is that they need to approximate the closest an AI model can come to the tripwire capability if used effectively.

AI Research and Development (AI R&D)

AI that can automate many, or all, of the *tasks currently done by top AI researchers and engineers* could have extreme risks as well as extreme benefits (and is probably something AI developers will be actively pursuing, given how much it can accelerate their work).[38] This piece will not provide a full discussion of why this is, but will outline the basics.

Tripwire capability: AI that can be used to do all (or the equivalent) of the tasks done by the major capability research teams at a top AI company for similar total costs (including salary, benefits, and compute for the costs of a human researcher). Or AI that, by any mechanism, leads to a dramatic acceleration in the pace of AI capabilities improvements compared to the pace of 2022–2024—a period of high progress and investment, for which good data is available.[39]

Risk(s): There are several interrelated reasons this tripwire could be important.

One is the potential for an *AI R&D feedback loop.* Today, the top teams focused on frontier AI research likely have no more than a few hundred researchers and engineers each.[40] If an AI model could stand in for top researchers and engineers, this could be the equivalent of adding hundreds of thousands (or more) such people.[41] This in turn could lead to a dramatic acceleration in AI progress, far beyond today's pace of improvements.[42] Many risks could emerge as a result, including:

- AI that becomes vastly better than humans at key tasks, including tasks related to other threat models discussed in this piece (chemical and biological weapons, cyber operations, persuasion) as well as R&D in other key domains, such as robotics and other military applications, leading to a wide set of quickly emerging, difficult-to-predict risks.

- Rapidly changing AI development methods (due to the large amount of automated research taking place) that may quickly increase the risk of AI models developing dangerous goals of their own (known as rogue AI),[43] which would be especially problematic if combined with superhuman capabilities.

- With this pace of progress, some company or country being a few months ahead of the rest of the world in AI could quickly result in their having access to vastly more capable AI models. This could lead to destabilizing changes in the balance of power, and this dynamic could also give an advantage to companies and countries that race ahead with little regard for risk mitigations on any front.

Another reason this tripwire could be important is the potential for *AI R&D as an early indicator of R&D capabilities more generally.* Eventually, it may make sense to have many different tripwires for AI capabilities in different R&D domains that might pose risks, for instance in robotics and surveillance. But there is some reason to think that AI R&D capabilities will emerge before more general R&D capabilities, since AI developers are especially likely to be actively optimizing their AI models for AI R&D (and since AI R&D has relatively fast experimental feedback loops and relatively little reliance on physical presence). As discussed below, it could be easier to design evaluations for AI R&D, especially related to other kinds of R&D. For these reasons, it may make sense to prioritize evaluations for AI R&D, even if one assumes that the AI R&D feedback loop described above is not a risk.

Relatedly, *AI R&D could serve as an indicator of general problem-solving, troubleshooting, and coordination abilities.* It would be helpful to get a sense of whether AI models working together can carry out complex tasks requiring many steps, creativity, and dealing with unexpected problems—both to get a sense of AI's potential beneficial applications and to assess broader risks from AI in the wrong hands (or rogue AI) capable of automating large, ambitious projects.

Damage potential: Rapidly advancing AI could raise any number of further risks without time to put in appropriate risk mitigations. The risks raised above—particularly from rogue AI and from global power imbalances—are speculative and highly debatable, but present the kind of high stakes that have led some to invoke extreme scenarios such as extinction.[44]

Risk mitigations: If AI models crossed this tripwire, a large number of different risks could develop quickly (due to potentially rapid progress in AI capabilities, as well as the possibility that AI models crossing this tripwire might also be quickly adaptable to R&D in a number of other key domains).

Because of this, it might be important to prepare for a wide variety of risks—including some that seem speculative and far-off today—in advance of hitting this tripwire.

- As noted above, there might be extreme pressure to race forward with AI development, since a lead could become self-reinforcing. With AI development bottlenecked by scarce resources (such as semiconductor fabs and lithography machines) and very high potential stakes, such a race could bring the danger of violent conflict. A framework for regulatory oversight and international coordination to avoid outcomes like this could be important.

- Stealing the model weights for AI models that cross this tripwire could be especially appealing and especially important to prevent. A state-backed program could start off far behind in AI, steal the weights of a top-notch model, and quickly become competitive with the rest of the world in AI—or even pull ahead, if it invested

more capital than other players in automated R&D and/or took less care than other players to ensure safety and reliability. AI models crossing this tripwire would ideally be kept under good enough security so as to protect the model weights even from well-resourced attacks from strong espionage programs.

- It might be important to have a plan for avoiding—and for detecting any presence of—misaligned power-seeking (or rogue) behavior from AI models, by the time a dramatic acceleration in AI capabilities becomes possible. Ideally this plan would be backed by a wide consensus of AI alignment researchers.

Evaluations: Some possible strategies for evaluating AI models for this capability:

- *Tasks based on existing AI R&D workflows.* AI models can be challenged to complete tasks based on the existing duties and workflows of academic AI researchers, scientists, and engineers at AI companies, and so on. There are some significant challenges here. R&D work is dynamic by nature and many of the key tasks might be hard to evaluate without giving them months to play out (and significant compute budgets), but with time and iteration, it is possible to develop practical evaluations that are reasonably representative of most of the skills human R&D experts need. Some early attempts at evaluations along these lines include MLE-bench, MLAgentBench, and RE-bench.

- *Measuring progress in general AI performance and looking for signs of acceleration.* Rather than gaining similar R&D capabilities to humans, AI models might gain different, complementary capabilities that could lead to similar acceleration dynamics. AI developers can track performance improvements of their models in a way that would make it possible to see whether progress is greatly accelerating.

More Possible Tripwires

This piece is not exhaustive and there are a number of other possibilities for tripwires listed below.

- **More tripwires for chemical and biological weapons.** AI might contribute to chemical and/or biological weapons development, production, and deployment in ways other than those listed above. For example, by helping a would-be terrorist form a high-level strategy to achieve their goals more effectively or cheaply, as opposed to advising them step-by-step on their work on a lab. This piece has focused on large-language-model-like AIs, but there could also be tripwires for specialized tools (for example, biological design tools) that might help with weapons development in other ways.

- **More R&D tripwires.** As noted above, there are a number of additional domains for which automated R&D capabilities could prove dangerous, such as robotics and surveillance.

- **More tripwires for persuasion, manipulation and/or misinformation (outside of the domain of politics).** There are many concerns about these sorts of capabilities, but currently few evaluation plans that have been concretely linked to particular risks. For now, this piece has focused on a particularly concrete evaluation strategy.

- **More tripwires for cyber operations.** There might be particular tasks relevant to cyber attacks that AI proves especially strong at, and that prove especially important, other than what is listed above (for example, making it easier to evade detection while launching attacks and gathering information).

- **Tripwires for general AI capabilities such as planning, coordination, and evasion of oversight.** The better AIs are at capabilities like this, the more it might be possible for them to work together on large, complex operations, and this could result in hard-to-predict risks—especially when it comes to rogue AIs. (As noted above, there are already some evaluations for the ability of AI systems to carry out long, complex research and engineering projects, which can partially address these properties.)

Summary Table

Tripwire capability	Risk(s)	Risk mitigations	Evaluations
Basic chemical and biological weapons production: AI capable of interactively advising a threat actor (with a college education and a moderate amount of time and money,[a] but no deep relevant expertise or experience) as effectively as a human expert in relevant domains, such that this actor would have a substantial chance of succeeding in an attempt to produce and release a chemical or biological weapon that could plausibly be catastrophically damaging.[b]	Greatly multiplying the number of people with the ability to produce and release a weapon of mass destruction, should they choose to. Weapons of mass destruction could do trillions of dollars' worth of damages or more.	Deployment safety: even a determined actor should not be able to reliably elicit chemical or biological weapons advice, including via jailbreak techniques. Model weight security: the model should only be stored in environments such that it would be highly unlikely that a terrorist individual or organization could obtain the model weights.	Experiments on whether novices can complete proxy tasks (safe tasks of similar difficulty to chemical and/or biological weapons production) with or without help from AI models. Evaluating an AI model's ability to answer questions (or help a human answer questions) about topics relevant to chemical and/or biological weapons.
Advanced chemical and biological weapons development and production: AI capable of uplifting multiple moderately resourced state chemical and/or biological weapons development programs' capabilities to the point where they could develop, produce, and deploy weapons of mass destruction at least as well as a well-resourced team of top experts could as of 2023.[c]	Could enhance many nations' prospects for secretly developing weapons of mass destruction, and hence increase the risks of a military conflict or an inadvertent deployment of a weapon of mass destruction.	Similar to above, but with a higher level of assurance: deployment safety and model weight security should be strong enough to resist even concerted attempts, with support from state programs, to jailbreak an AI or steal its model weights.	Similar to above, but with more emphasis on advising people with *some* expertise to complete *highly* challenging tasks (as opposed to advising people with *no* expertise to complete *moderately* challenging tasks).[d]
Efficient persuasion: AI capable of dramatically increasing the cost-effectiveness of professionalized persuasion, in terms of the effect size (for example, the number of people changing their views or taking some specific action) per dollar spent.[e] This might happen (for example) via interactive chat that is much more cost-effective than traditional advertising.	A number of potential risks, including asymmetrically affecting discourse and elections. Could serve as an early warning sign for extreme persuasion abilities, including the risk that rogue AIs with powerful persuasion abilities could manipulate AI employees to circumvent safety and security protections.	Intensive review of an AI model's persuasion capabilities and likely impacts, possibly followed by deployment safety and model weight security measures similar to those for advanced chemical and biological weapons.	Challenging experts in professionalized persuasion to find a way to use AI to beat state-of-the-art cost-effectiveness for persuasion on a particular topic and testing their ideas using randomized recipients of different persuasion techniques.

Tripwire capability	Risk(s)	Risk mitigations	Evaluations
Generalized cyber operations capabilities: AI capable of increasing a moderately resourced state program's technical cyber operations capabilities (those that rely only on remote execution, as opposed to things like close-access attacks and recruiting insiders) to par with those of the top-tier intelligence services (especially for purposes of stealing future AI model weights and/or disabling critical infrastructure at scale). [f]	Could make it possible for a moderately resourced state program to carry out a variety of attacks associated with stronger programs, including the possibility of coordinated, scaled attacks on critical infrastructure (such as power and water plants) and/or large-scale, targeted theft of classified information. This capability could make AI an especially attractive target for foreign intelligence services and stealing one generation's model weights might make it easier to steal the next generation's.	Intensively assess a model's contributions to both cyber attacks and defenses before widely deploying the model or developing more powerful versions. Deployment safety and model weight security similar for advanced chemical and biological weapons capabilities may be needed.	Tasks and challenges representative of what top human cyber operations experts can accomplish.
Vulnerability discovery and exploitation for cyber attacks: AI capable of dramatically accelerating the rate of discovery of novel, high-value vulnerabilities, exploits, and/or combinations.[g]	An extreme enough proliferation of very valuable vulnerabilities and/or exploits might overwhelm the ability of defenders to respond (for example, with software patches). AI models that could produce a large enough number of such vulnerabilities and/or exploits could also be dangerous if used privately by a state program in possession of stolen model weights.	Similar to above.	Attempting to use AI to find and/or exploit novel vulnerabilities using informed expert guesses at how it might do this best.
AI research and development (AI R&D): AI that can be used to do all (or the equivalent) of the tasks done by the major capability research teams at a top AI company for similar total costs (including salary, benefits, and compute for the costs of a human researcher). Or AI that, by any mechanism, leads to a dramatic acceleration in the pace of AI capabilities improvements compared to the pace of 2022–2024—a period of high progress and investment, for which good data is available.[h]	AI systems with this capability could be used to create a feedback loop (huge amounts of automated AI research leading to increased efficiency and capability for AI, which leads to even more automated AI research, continuing the loop), leading to dramatic acceleration in AI progress (many times faster than today's pace). This would pose a number of major risks, including: (a) new risks could arise from rapidly developed AI capabilities at a pace that would make identifying and adapting to risks infeasible; (b) a state (or even a company) that invests heavily in AI and takes few precautions could quickly gain an enormous, self-reinforcing technological lead on the rest of the world, which means there could be intense pressure to race and high risk of disruptions to the balance of power; (c) dangers from rogue AI could greatly and quickly increase. AIs demonstrating this capability could also provide early evidence of more general capabilities related to R&D, general problem-solving, and so on, which could pose a number of other threats.	High-assurance deployment safety and model weight security, as above. Additionally, due to the potential for rapid acceleration, it might be important to prepare for a wide variety of risks, including some that seem speculative and far-off today, in advance of hitting this tripwire. This could include developing high-assurance methods for reducing dangers from rogue AI and international mechanisms to mitigate the intense pressure to race forward recklessly with AI capability scaling.	Tasks based on existing AI R&D workflows and representative of the tasks top AI researchers carry out today. Monitoring the rate of progress in AI capabilities for signs of acceleration.

Summary Table Notes

a Something like $50,000 and six months.

b Example operationalization of "substantial chance of succeeding": at least 10 percent probability for an average actor with these properties. Example operationalization of "catastrophically damaging": at least $100 billion in damages.

c In light of the Biological Weapons Convention, nearly all state bioweapons programs are likely to be only moderately resourced, that is, not drawing on top talent or commanding large budgets to a similar extent that many states' cyber operations do. Many of these tripwires use a baseline of 2023, when the best large language models were not capable enough to make a significant difference in any of these domains.

d Evals here could take some inspiration from the Nth Country Experiment.

e "Dramatically" could be operationalized as something like 5x or more (relative to a 2023 benchmark).

f Here, moderately resourced state programs refer to the strongest programs that are *not* in the five to ten strongest programs worldwide, as judged by the actor making an if-then commitment.

g "Dramatically" could be operationalized as something like 5x or more (relative to a 2023 benchmark). "High-value" could be assessed by estimating how much they could be sold for on the open market, based on similarities to other vulnerabilities and exploits whose market value is known.

h As noted in the main text, this tripwire can be operationalized simply by looking for the dramatic acceleration, which would be highly consequential and suggestive of this dynamic on its own. If the acceleration happens and is measured, one doesn't need to separately establish that this was *because of* automated AI R&D (doing the latter could be very fraught). Dramatic acceleration refers to, for example, "an increase in the effective training compute of the world's most capable model that, over the course of a year, was equivalent to two years of the average rate of progress during the period of early 2018 to early 2024." (See footnote 4 here for a definition of "effective compute.")

All of the candidate tripwire capabilities listed above would benefit from more refinement, more analysis of the potential damages from associated catastrophes, more analysis of the risk mitigations that could help (and how costly they would be), and generally more discussion from a broad set of experts and stakeholders. But they can serve as starting points for engagement, and thus help push toward the goal of a mature science of identifying, testing for, and mitigating AI risks, without slowing development of new technology unnecessarily.

Future Work

There are many possible research projects that could result in better understanding key threat models and candidate tripwire capabilities. Some examples are given below.

Comprehensive threat mapping. This piece has focused on a relatively short list of threats, selected for high damage potential and other desiderata. A formal exercise to list and taxonomize all plausible threats, and efficiently prioritize them for further investigation, could be valuable, especially if it incorporated feedback from a broad and diverse set of experts.

Examining and quantifying specific risks. When trying to quantify a risk of *future* AI systems, there is a basic problem: one cannot straightforwardly use statistics on past catastrophes to determine likelihood and magnitude. However, there are some potentially productive ways to analyze likelihood and magnitude, including the following list.

- *Systematic forecasting exercises* that aggregate judgments about the size and likelihood of risks from panels of subject-matter experts and/or people (such as superforecasters) who specialize in forecasting itself.

- *Studying how well human experts can accomplish tasks of interest.* For example, the above discussion of bioweapons proposes that one might "challenge study participants to complete a set of tasks in a laboratory that are analogous to different parts of weapons production and release." This study could be run with some participants having access to a *human expert* advising them, with the human expert simulating the sort of assistance a future AI might be able to give. This could help (a) capture the quantitative increase in risk that a hypothetical human-expert-level AI could cause and (b) establish a benchmark for comparing AI performance to.

- *Using historical data and case studies to fill in part of the picture*, even if there will inevitably be an element of speculative extrapolation. For example, to estimate the damage potential of AI-assisted cyber attacks, one might examine how damaging cyber attacks have been historically, particularly cyber attacks of the kind that might become more common if AI with relevant capabilities were available.

- *Quantitative estimation exercises.* Using analytical models, with explicit assumptions based on real-world data to the extent feasible, to quantify particular risks. An example of this sort of work from other domains would be social cost of carbon estimates that incorporate potential economic damages from climate change. **Fleshing out potential risk mitigations and estimating their costs.** The less costly it is to mitigate a risk, the less it is necessary to establish that the risk is highly likely and/or has high damage potential.

About the Author

Holden Karnofsky is a visiting scholar at Carnegie California. His research focuses on international security risks from advances in artificial intelligence: what the most imminent risks are, how to prepare, and possible early warnings (e.g. from AI capability evaluations).

Holden previously served as co-founder and CEO (and later co-CEO) of Open Philanthropy. Open Philanthropy has been one of the largest philanthropic funders of both AI risk reduction and biosecurity and pandemic preparedness since 2015. It also works in a number of other areas including global health R&D (including work toward universal flu and syphilis vaccines, hepatitis B cures and malaria gene drives), land use reform (it was the first institutional funder of the YIMBY movement), and farm animal welfare (where its grantees have won thousands of commitments for improved animal treatment).

Prior to that, Holden co-founded and served as co-Executive Director of GiveWell, whose public charity recommendations direct hundreds of millions of dollars per year.

He is married to the President of Anthropic (an AI company) and has financial exposure to both Anthropic and OpenAI via his spouse.

Acknowledgments

This piece has benefited from a large number of discussions over the last year-plus on if-then commitments, particularly with people from METR, the UK AI Safety Institute, Open Philanthropy, Google DeepMind, OpenAI, and Anthropic. For this piece in particular, I'd like to thank Chris Painter, Luca Righetti, and Hjalmar Wijk for especially in-depth comments; Ella Guest and Greg McKelvey for comments on the discussion of chemical and biological weapons; Omer Nevo for comments on the discussion of cyber operations; Josh Kalla for comments on the discussion of persuasion and manipulation capabilities; and my Carnegie colleagues, particularly Jon Bateman, Alana Brase, Helena Jordheim, and Ian Klaus, for support on the drafting and publishing process.

Notes

1 These first two paragraphs appeared in an earlier piece on if-then commitments.

2 See Google DeepMind's Frontier Safety Framework, OpenAI's Preparedness Framework, and Anthropic's Responsible Scaling Policy.

3 See Google DeepMind's Frontier Safety Framework, OpenAI's Preparedness Framework, and Anthropic's Responsible Scaling Policy, all of which emphasize the need for revisions over time (see note 6, below).

4 Damages can include property damage, economic deadweight loss, and loss of life and health (the latter can be valued using value of life methods).

5 See Google DeepMind's Frontier Safety Framework, OpenAI's Preparedness Framework, Anthropic's Responsible Scaling Policy and Magic.dev's AGI Readiness Policy. These all propose something like candidate tripwire capabilities.

6 These include a paper explaining the case for these (Model Evaluation for Extreme Risks, published in May 2023); a set of voluntary commitments announced by the White House that heavily featured evaluating AIs to determine risks before release; a U.S. executive order with a significant focus on "AI model evaluation tools and AI testbeds"; and the establishment of both the UK AI Safety Institute and U.S. AI Safety Institute, both significantly focused on safety evaluations for AI systems.

7 See METR's post on responsible scaling policies.

8 "The Framework is exploratory and we expect it to evolve significantly as we learn from its implementation, deepen our understanding of AI risks and evaluations, and collaborate with industry, academia, and government. Even though these risks are beyond the reach of present-day models, we hope that implementing and improving the Framework will help us prepare to address them. We aim to have this initial framework fully implemented by early 2025," from Google DeepMind's blog post introducing its Frontier Safety Framework.

"This framework is the initial Beta version that we are adopting, and is intended to be a living document. We expect it to be updated regularly as we learn more and receive additional feedback," from OpenAI's announcement of its Preparedness Framework.

"However, we want to emphasize that these commitments are our current best guess, and an early iteration that we will build on. The fast pace and many uncertainties of AI as a field imply that, unlike the relatively stable BSL system, rapid iteration and course correction will almost certainly be necessary," from Anthropic's blog post introducing its Responsible Scaling Policy.

9 See the May update from the AI Safety Institute.

10 See Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models (a 2024 RAND publication).

11 Rogue AI, or misaligned power-seeking AI, refers to AI whose training leads it to develop dangerous, unintended objectives such that it optimizes for deceiving and disempowering humans. Fast enough AI progress could increase the likelihood that AI develops such objectives and that it has strong enough capabilities to cause catastrophes without being deliberately used to do so by humans.

12 That is, the expected value of the damages caused. An oversimplified example: an AI system that is responsible (over and above the default/baseline risk) a 1 percent annual risk of a catastrophe causing $10 trillion in damages would have *expected damages* of $100 billion per year.

13 For example, OpenAI is reportedly seeking a $150 billion valuation. Delays (or productivity-lowering risk mitigations, such as intensive information security) that reduced its valuation by a few percent could therefore be argued to be costing society billions of dollars.

14 For example, this paper's section on "Vulnerability discovery and exploitation for cyber attacks."

15 For example, this paper's section on "Basic chemical and biological weapons production."

16 A corollary of this point about countermeasures is that the threats most likely to qualify here tend to involve concentrated damages—damages that take place relatively quickly, before society can adapt and respond. A pandemic is an example of a catastrophe with highly concentrated damages.

17 See pages 16–19 of OpenAI's Preparedness Framework and pages 16–20 of Anthropic's Responsible Scaling Policy.

18 See Google DeepMind's Frontier Safety Framework, OpenAI's Preparedness Framework, Anthropic's Responsible Scaling Policy and Magic.dev's AGI Readiness Policy. These all propose something like candidate tripwire capabilities.

19 Particularly Open Philanthropy's Luca Righetti and METR's Hjalmar Wijk.

20 Something like $50,000 and six months.

21 Example operationalization of "substantial chance of succeeding": at least 10 percent probability for an average actor with these properties. Example operationalization of "catastrophically damaging": at least $100 billion in damages.

22 See "Chemical Weapons: Easy to Make, Hard to Destroy" for a discussion of chemical weapons. For biological weapons, this view is debated among experts, but for an example of experts seemingly endorsing a similar view, see "Biodefence in the Age of Synthetic Biology": "The production of most DNA viruses would be achievable by an individual with relatively common cell culture and virus purification skills and access to basic laboratory equipment, making this scenario feasible with a relatively small organizational footprint (including, e.g., a biosafety cabinet, a cell culture incubator, centrifuge, and commonly available small equipment). Depending upon the nature of the viral genome, obtaining an RNA virus from a cDNA construct could be more or less difficult than obtaining a DNA virus. Overall, however, the level of skill and amount of resources required to produce an RNA virus is not much higher than that for a DNA virus."

23 For example, one estimate from congressional testimony is that "approximately 30,000 individuals are capable of assembling any influenza virus for which a genome sequence is publicly available." This comes in the context of relatively high concern about the risk; others might think the number is lower. The percentage of the population capable of producing a given chemical or biological weapon would of course vary with what the specific weapon is, and is likely higher for chemical than for biological weapons.

24 "The COVID-19 pandemic has paralyzed large parts of the global economy, sharply restricting economic activities, increasing uncertainties and unleashing a recession unseen since the Great Depression. Global gross domestic product (GDP) is forecast to shrink by 3.2 per cent in 2020, with only a gradual recovery of lost output projected for 2021. Cumulatively, the world economy is expected to lose nearly $8.5 trillion in output in 2020 and 2021 (Figure 1), nearly wiping out the cumulative output gains of the previous four years." From "World Economic Situation and Prospects as of Mid-2020," United Nations.

25 "The cumulative loss in output relative to the pre-pandemic projected path is projected to grow from 11 trillion over 2020–21 to 28 trillion over 2020–25. This represents a severe setback to the improvement in average living standards across all country groups." From an IMF blog post.

"In October 2020, David Cutler and Lawrence H. Summers published a brief article in JAMA Viewpoint estimating that COVID-19 would cost the United States $16 trillion dollars, when combining economic damages and monetized health and life loss. This figure has been extensively cited and used in policy discussions. In this article, we update their estimate, using facts about the disease and its costs to society that have become known since their paper was published. We find that the total harms of COVID-19 to the U.S. are still about $16 trillion (with a range of $10 trillion and $22 trillion) but the components of harm are significantly different than those estimated by Cutler & Summers. The pandemic caused less economic damage than they projected, but more mental health damage." From Institute for Progress.

"By 2024, it is estimated that the Covid-19 pandemic will have reduced economic output by $13.8 trillion relative to pre-pandemic forecasts (International Monetary Fund 2022). The pandemic resulted in an estimated 7–13 million excess deaths (Economist 2022) and an estimated $10–$17 trillion loss of future productivity and earnings from school disruption (Azevedo et al. 2021). Such devastating losses from a pandemic are not new: some sources estimate that the 1918 flu killed 2% of the world's population and reduced GDP by 6% (Barro, Ursúa, and Weng 2020) and that the Black Death killed 30% of Europe's population (Alfani 2022)." Glennerster, Snyder, and Tan 2023.

26 This doesn't mean that AIs would never be allowed to help users with relevant tasks, only that there might be different restrictions on different classes of users. For example, there might be AI models for use in academia that had fewer restrictions than general-use AI models.

27 The weapon in question should be among the easiest weapons to produce and deploy that have damage potential over the threshold specified by the tripwire (this threshold might vary by actor, as noted in a footnote to the tripwire language).

28 Some discussion of risks from biological design tools here.

29 Many of these tripwires use a baseline of 2023, when the best large language models were not capable enough to make a significant difference in any of these domains.

30 See Moratorium on Research Intended to Create Novel Potential Pandemic Pathogens for discussion of the general risks of inadvertent release of pathogens.

31 For example, synthesizing horsepox (not contagious in humans) rather than smallpox (dangerous).

32 Here, moderately resourced state programs refer to the strongest programs that are *not* in the five-to-ten strongest programs worldwide, as judged by the actor making an if-then commitment.

33 From an article about the breach: "There was 'reluctance or concern or anxiety about putting our officers in the field given that our protective shield had been punctured [by the OPM breach],' recalled the former national security official. 'We didn't fully know what they knew about us.' Subsequently, 'dozens of postings' for CIA officers scheduled for assignments in China were canceled, according to *The Perfect Weapon*, a 2018 book by David Sanger. 'CIA, for many years, was not willing to do forward facing ops in China,' because its confidence was so shaken by the asset roll-up and other breaches, said a former senior intelligence analyst."

34 For example, see "How LLMs Are Enabling Automated Vulnerability Discovery."

35 "Dramatically" could be operationalized as something like 5x or more (relative to a 2023 benchmark). "High-value" could be assessed by estimating how much they could be sold for on the open market, based on similarities to other vulnerabilities and exploits whose market value is known.

36 "The result was more than $10 billion in total damages, according to a White House assessment confirmed to WIRED by former homeland security adviser Tom Bossert, who at the time of the attack was President Trump's most senior cybersecurity-focused official." From a Wired article on NotPetya.
"Cyber risk modeling firm Cyence estimates the potential costs from the hack at $4 billion, while other groups predict losses would be in the hundreds of millions." From CBS News.

37 "Dramatically" could be operationalized as something like 5x or more (relative to a 2023 benchmark).

38 Or AI that can autonomously push forward R&D progress in some other way (for example, by automating different tasks that still contribute heavily to progress).

39 As noted in the main text, this tripwire can be operationalized simply by looking for the dramatic acceleration, which would be highly consequential and suggestive of this dynamic on its own. If the acceleration happens and is measured, one doesn't need to separately establish that this was *because of* automated AI R&D (doing the latter could be very fraught).

Dramatic acceleration refers to, for example, "an increase in the effective training compute of the world's most capable model that, over the course of a year, was equivalent to two years of the average rate of progress during the period of early 2018 to early 2024." (See footnote 4 here for a definition of "effective compute.")

40 It's hard to get reliable public data on this, but a number around 800 was being discussed for OpenAI's *total* workforce in late 2023.

41 Sample calculations here (see "We'd be able to run millions of copies (and soon at 10x+ human speed) of the automated AI researchers") and here.

42 Fully making this case is outside the scope of this piece, but it is outlined in a report by Tom Davidson.

43 How AIs could become rogue and how it relates to the pace of progress is outside the scope of this piece. There are many explainers on this topic; I recommend Yoshua Bengio's as a starting point. I argued in an informal piece that the pace of progress could be a key factor in how big this risk is.

44 Popular treatments of AI-connected risk of human extinction have generally emphasized the risk factor of a fast transition from AIs that are less capable than humans to AIs that are vastly more capable, and have additionally emphasized the risk factor of AIs that can themselves do research and development for improving AI (sometimes described as "self-improvement," though the research an AI does need not be applied to itself). For examples, see Chapter 4 of Superinteligence, Chapter 5 of Human Compatible, and Chapter 4 of Life 3.0.

Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 170 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Carnegie California

Carnegie California links developments in California and the West Coast with national and global conversations around technology, subnational affairs, and trans-Pacific relationships. At distance from national capitals, and located in one of the world's great experiments in pluralist democracy, Carnegie California engages a wide array of stakeholders as partners in its research and policy engagement.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org