CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

SEPTEMBER 2024

If-Then Commitments for AI Risk Reduction

Holden Karnofsky

If-Then Commitments for AI Risk Reduction

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

Artificial intelligence (AI) could pose a variety of catastrophic risks to international security in several domains, including the proliferation and acceleration of cyberoffense capabilities, and of the ability to develop chemical or biological weapons of mass destruction. Even the most powerful AI models today are not yet capable enough to pose such risks,[1] but the coming years could see fast and hard-to-predict changes in AI capabilities. Both companies and governments have shown significant interest in finding ways to prepare for such risks without unnecessarily slowing the development of new technology.

This piece is a primer on an emerging framework for handling this challenge: if-then commitments. These are commitments of the form: *If an AI model has capability X, risk mitigations Y must be in place. And, if needed, we will delay AI deployment and/or development to ensure the mitigations can be present in time.* A specific example: *If an AI model has the ability to walk a novice through constructing a weapon of mass destruction, we must ensure that there are no easy ways for consumers to elicit behavior in this category from the AI model.*

If-then commitments can be voluntarily adopted by AI developers; they also, potentially, can be enforced by regulators. Adoption of if-then commitments could help reduce risks from AI in two key ways: (a) prototyping, battle-testing, and building consensus around a potential framework for regulation; and (b) helping AI developers and others build roadmaps of what risk mitigations need to be in place by when. Such adoption does not require agreement on whether major AI risks are imminent—a polarized topic—only that certain situations *would* require certain risk mitigations *if* they came to pass.

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 2025 AI Action Summit in France.[2] Similar ideas have been explored at the International Dialogues on AI Safety in March 2024[3] and the UK AI Safety Summit in November 2023.[4] As of mid-2024, most discussions of if-then commitments have been in the context of voluntary commitments by companies, but this piece focuses on the general framework as something that could be useful to a variety of actors with different enforcement mechanisms.

This piece explains the key ideas behind if-then commitments via a **detailed walkthrough** of a particular if-then commitment, pertaining to the potential ability of an AI model to walk a novice through constructing a chemical or biological weapon of mass destruction. It then discusses some limitations of if-then commitments and closes with an outline of how different actors—including governments and companies—can contribute to the path toward a robust, enforceable system of if-then commitments.

Context and aims of this piece. In 2023, I helped with the initial development of ideas related to if-then commitments.[5] To date, I have focused on private discussion of this new framework; for instance, by encouraging companies to voluntarily adopt if-then commitments. The goal of this piece is to make it easier for people who are not currently familiar with this framework to understand its potential, as well as its limitations and challenges, for reducing risks. The more attention and interest there is in if-then commitments, the more effort a number of institutions likely will put into continuing to flesh out and experiment with their own, and the faster progress we can expect toward a mature framework for reducing risks from AI.

Walking Through a Potential If-Then Commitment in Detail

This section will discuss an extended example of an if-then commitment that could be adopted, in order to illustrate the key concepts and challenges of the framework. As noted below, the example has substantial overlap with some policies and frameworks that companies have adopted. However, this example is not simply copied over from any one existing if-then commitment. The goal is to present a relatively clear example, unencumbered by the kinds of commercial, legal, or other objectives that could affect how similar content is presented in a corporate policy.

The Risk: Proliferation of Expert-Level Advice on Weapons Production

A commonly discussed risk from AI[6] is its potential to contribute to chemical and biological weapons. Within that general frame, there are a number of distinct possible risks. The focus here will be on the hypothesis that an AI model could serve as a *virtual substitute* for an *expert adviser on chemical or biological weapons production*, thus greatly expanding the number of people who could produce and deploy such weapons. An appendix elaborates on the thinking behind this hypothesis.

The Challenge of Sufficient Risk Mitigations

Even the best AI models today likely lack the level of capability that would significantly increase the number of people capable of deploying catastrophically damaging weapons.[7] However, it is hard to know whether future AI models will have such capabilities. If some did, it could be challenging to keep the risks low, for a couple of reasons.

The first reason is the challenge of preventing jailbreaks. Today, the sorts of AI models most likely to have the concerning capability outlined above (large language models, or LLMs) generally are trained to refuse dangerous requests—which, in theory, should stop people seeking to build chemical and biological weapons from getting much help from even very capable LLMs. But it is currently possible to use certain patterns of dialogue to "jailbreak" the restrictions on LLMs, getting them to break their rules and cooperate with nearly any task.[8] Getting LLMs to reliably refuse harmful requests, without simply training them to refuse nearly all requests, remains an open problem in AI, and there is no guarantee that it will be solved within any particular timeframe (there are other approaches to the same goal, such as trying to deliberately make LLMs *incapable* of helping with certain requests, but these have their own challenges[9]).

Second, even if this problem were solved, anyone with access to an LLM's *model weights* still could be able to "undo" refusal training or other controls on the requests the LLM will and will not cooperate with.[10] This means that if model weights were not handled using reasonably good security practices—or if model weights were publicly released by an AI developer—even "jailbreak-proof" safety measures could be circumvented.

The Example If-Then Commitment

In an attempt to contain the risk of widely proliferating expertise in weapons of mass destruction, while *not* requiring difficult and costly risk mitigations for AI models that *do not* pose such a risk, a company or regulator might use the following approach:

- Identify a **tripwire capability** that would trigger the need for additional risk mitigations. In this case, the tripwire capability might be the capability to interactively advise a malicious actor[11] to the point at which the actor would have a substantial chance[12] of succeeding in an attempt to produce and release a catastrophically damaging CBRN weapon of mass destruction.[13]

- Make the following **if-then commitment**: *if* an AI model has the tripwire capability, *then* (a) it can only be deployed using methods and environments where a determined actor would reliably fail to elicit such advice from it; and (b) it can only be stored in environments such that it would be highly unlikely that a terrorist individual or organization could obtain the model weights. If these mitigations are not feasible for a particular AI developer to implement, then the developer should not deploy or even store a model with the tripwire capability until it can implement them. (As discussed below, this likely means pausing further AI development once there are warning signs of *approaching* the tripwire.)

This commitment, if operationalized well (as explained below) and adhered to, would reduce the risk without affecting models lacking the tripwire capability.

The balance of risk-reduction benefits and risk-mitigation costs will depend on the details of which categories of chemical and biological weapons are deemed to fit the tripwire's "catastrophically damaging" criterion, and how much risk they pose. For any if-then commitment, the wisdom of the commitment depends on the specifics of the risks. The rest of this section will provisionally assume the existence of a version of the if-then commitment that has greater benefits than costs.

Relationship to existing voluntary commitments. A number of existing policies and frameworks from AI companies contain content similar to the if-then commitment above.

OpenAI's Preparedness Framework lists "low," "medium," "high," and "critical" AI capabilities in four "tracked risk categories." It states that "Only models with a post-mitigation score of 'medium' or below can be deployed, and only models with a post-mitigation score of 'high' or below can be developed further," and also states that model weights must be protected for models with "high" capabilities.[14] Hence, its "high" and "critical" capabilities serve as something similar to "tripwires" that trigger commitments similar to the one listed above. Specifically, the "high" level triggers similar risk mitigations to those listed above and/or a pause in AI deployment, while the "critical" level triggers a pause in further AI development.[15] One of its four tracked risk categories is "CBRN (chemical, biological, radiological, nuclear)." Its "high" and "critical" levels include AI capabilities similar to the "tripwire" given above.[16]

Anthropic's Responsible Scaling Policy is built around "AI safety levels (ASL), which are modeled loosely after the US government's biosafety level (BSL) standards for handling of dangerous biological materials. We [Anthropic] define a series of AI capability thresholds

that represent increasing potential risks, such that each ASL requires more stringent safety, security, and operational measures than the previous one." Its "ASL-3" level presents similar risk mitigations to those in the above "if-then" commitment: "Harden security such that non-state attackers are unlikely to be able to steal model weights and advanced threat actors (e.g. states) cannot steal them without significant expense" and "implement strong misuse prevention measures, including . . . maximum jailbreak response times."

It also commits to pause AI deployment and development as needed to keep these commitments.[17] Hence, AI capabilities that trigger Anthropic's "ASL-3" standard will function similarly to the "tripwire" above. These elements include the capability to "substantially increase the risk of deliberately-caused catastrophic harm, either by proliferating capabilities, lowering costs, or enabling new methods of attack…Our first area of effort is in evaluating bioweapons risks." This is similar in spirit but less specific than the above tripwire.

Google DeepMind's Frontier Safety Framework specifies "protocols for the detection of capability levels at which models may pose severe risks (which we call "Critical Capability Levels (CCLs)"), and . . . a spectrum of mitigation options to address such risks." Its CCLs include a capability similar to the "tripwire" above.[18] Its mitigation options consist of "Security Mitigations" and "Deployment Mitigations" in a similar spirit to those listed under the "If-then" commitments above.[19] However, it does not (as the other two policies do) specify which mitigations correspond to which CCLs—instead, it is left up to the company to determine on a case-by-case basis which mitigations are appropriate for a given level. The "Future Work" section states an intention to map specific CCLs to specific mitigations in a later version of the framework.[20]

Google's framework also contains a discussion of pausing deployment and development as needed, as in the "if-then" commitment above: "A model may reach evaluation thresholds before mitigations at appropriate levels are ready. If this happens, we would put on hold further deployment or development, or implement additional protocols (such as the implementation of more precise early warning evaluations for a given CCL) to ensure models will not reach CCLs without appropriate security mitigations, and that models with CCLs will not be deployed without appropriate deployment mitigations."

Overall, the terminology, approach, and details vary among policies, but they all have content that significantly overlaps with the if-then commitment laid out above.

Potential Benefits of This If-Then Commitment

An if-then commitment along the above lines could have significant benefits.

First, such a commitment could be an appealing compromise between people who think the capability described above could emerge imminently and people who think it will not emerge for a very long time, if ever. The former group might expect the if-then commitment

to result in important risk mitigations soon; the latter might expect the if-then commitment to amount to little beyond running evaluations, as described below.

Second, such a commitment would provide a clear, action-relevant goal for the design of AI evaluations: evaluations should seek to determine whether a given AI model is close to the tripwire laid out above. Teams that design evaluations could create a mix of (a) relatively expensive, time-consuming evaluations that clearly inform developers about whether an AI model is close to the tripwire; or (b) cheaper, more practical evaluations that aim to approximate (a).

More broadly, with such a commitment in place, AI developers and others could experiment with a number of ways of operationalizing it—a number of different approaches to evaluating AI capabilities, to evaluating the sufficiency of security measures, and the like—and discover over time how to make these operationalizations practical to implement. This kind of experimentation and learning could be helpful for eventually developing battle-tested, scalable ways of implementing the commitment, which could be important for developing practical, protective policies (ranging from industry standards to national and potentially international policies) over time.

Furthermore, such a commitment could help AI developers with planning and prioritizing for risk mitigation measures. For example, an AI company that makes internal predictions about the future capabilities of its models could use this commitment to build a roadmap for risk mitigation measures—something along the lines of: *We expect AI models with the tripwire capabilities in N years, so we need to resource our teams appropriately to have jailbreak-proof restrictions on how our AI models can be used by then, and to store such AI models under strong enough security practices.* Companies that have made commitments similar to this one have emphasized this benefit. For example, OpenAI's Preparedness Framework explicitly discusses roadmapping as part of its work. Anthropic has stated that "teams such as security, trust and safety, red teaming, and interpretability, have had to greatly ramp up hiring to have a reasonable chance of achieving ASL-3 safety measures by the time we have ASL-3 models." Broadly speaking, commitments like this have the potential to create a "race to the top." If powerful AI models can only be developed and deployed with strong risk mitigations in place, developing strong risk mitigations could become an essential part of what AI developers compete on and, accordingly, prioritize.

Operationalizing the Tripwire

How does one know if an AI is near or at the tripwire? This sort of question is the subject of an emerging field aiming to design tests that determine what dangerous or dual-use (both beneficial and potentially risk-inducing) capabilities a given AI model has. Evaluations

(evals) of these capabilities are a major focus of the US AI Safety Institute,[21] the UK AI Safety Institute,[22] and teams at several major AI companies.[23]

Below are several potential approaches to building evals for the tripwire under discussion. For ease of explanation, the list starts with highly relevant but expensive and difficult evals to run and ends with more approximate but practical evals. This latter category includes most of the current evals being run or built.

Hypothetical, idealized experiment. Ultimately, the goal is to answer questions like: "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,[24] 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? 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, but also because it could take time to recruit participants and allow them to attempt the work.

Approximations of this experiment. One could run various approximations of the above experiment. 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 nondangerous proxy. Such an experiment could feature a pathogen that is not transmissible in humans, but involves challenges similar to those required for a dangerous pathogen. It might otherwise be modified for practicality, perhaps involving the same types of challenges but taking less time. Although this approach is more practical than the previous approach, it still would lead to relatively expensive evals that take significant calendar time, and it is not the main approach used for today's evals.

Running experiments with human experts to generate inspiration for quicker tests. Similar experiments could be run today with an *actual human expert role-playing a possible future AI model*. Specifically, participants in the treatment group could be given access to a Slack conversation with an expert in relevant domains, while participants in the control group could lack such access. This sort of experiment would not directly provide evidence about a particular AI model's capabilities. However, it could provide a lot of information about which steps are hardest and at what points in the process experts are most helpful. Transcripts of discussions between participants and expert advisers could be used to build simpler, automated evals. One possible option would be to see whether an AI model prompted with a question from the transcript can produce an answer akin to that of an expert—this might take the form of something like looking at a photo a participant took of their project in progress and diagnosing a problem. There are some ongoing efforts (though the details are not publicly shareable) to run experiments along these lines. As a side benefit, such

experiments might provide evidence about whether the basic model of the risk described above is legitimate in the first place.

Easier, simpler tests. One approach—and indeed, the most common way evaluations are being run today[25]—is to design relatively simple tests that are not only much quicker and cheaper to administer than the idealized experiment, but present a *strictly easier* task for the AI model than the tripwire capability does. For example, one might simply test the AI model's ability to correctly answer, or help a human correctly answer, questions about chemistry and/or biology. If it did relatively poorly—that is, achieving worse performance than a human without access to state-of-the-art language models[26]—this could (depending on the details of the test) be used to argue that the AI model would be unlikely to be an effective stand-in for a human chemistry or biology expert advising on weapons production.

The field of evaluations for catastrophically dangerous AI capabilities is a very young one.[27] It is likely that there will be many more ideas for practical, affordable tests of AI capabilities.

Challenges of running and interpreting evaluations. The above discussion has focused on what kinds of tasks might be informative about whether an AI has a tripwire capability. It's worth noting that there are a number of additional challenges when it comes to running and interpreting evals.

For example, an AI model that *appears* to lack tripwire capabilities in testing might demonstrate the capabilities if it were prompted differently, fine-tuned differently, or given access to more tools. To account for this possibility, those running the testing can make a serious effort to get the best performance possible out of an AI model. This likely means involving researchers who are highly experienced and knowledgeable about how to elicit strong performance on the tasks in question and giving them time and resources to do so. This principle appears in existing voluntary commitments from companies.[28]

On the flip side, an AI model that *appears* to have tripwire capabilities in testing may in fact be using brittle "memorization" of similar tasks it's seen before. Designers of evals often make special efforts to avoid letting the solutions (and even the challenges) get onto the public web and enter into AI training data.[29]

Another issue is that, for reasons outlined above, evals generally aim to make a case that an AI model is reasonably *far from* possessing a tripwire capability. Accordingly, the tasks tested in evals generally amount easier than the ultimate task of concern—in this case, successfully advising an actor on production of a chemical or biological weapon. If-then commitments can leave a "buffer" by triggering the "then" part of an "if-then commitment" when evals suggest that the tripwire capability is relatively *near*, as opposed to clearly present. Some existing voluntary commitments from companies reflect this principle.[30]

Operationalizing the "Then" Part of the If-Then Commitment

The commitment suggested above includes criteria for *deployment safety* (ensuring that users cannot elicit dangerous behavior from the AI model) and *model weight security* (ensuring that the weights are unlikely to be stolen). How does one translate these commitments into specific practices? This, too, is an emerging area of inquiry—and so far, the main proposals on it have come from AI companies making voluntary commitments.[31] Below are examples of some approaches that have emerged.

Deployment safety: There are various possible approaches to preventing users from eliciting dangerous behavior from an AI model, including: training the AI model to refuse harmful requests; using AI to monitor and report harmful requests; and attempting to remove dangerous capabilities, such as knowledge of certain domains, from the AI model itself. To assess whether an approach is effective enough to fulfill the commitment, one can use "red teaming." This refers to a dedicated team—perhaps external to the company, such as the team at the UK AI Safety Institute that recently demonstrated the ease of jailbreaking today's models—that looks for ways to elicit dangerous behavior from AI models and certifies deployment safety measures as sufficient only if they (the team) fail to do so. This approach features in both Google DeepMind's Frontier Safety Framework (see "Deployment Mitigations" table on page 4) and Anthropic's Responsible Scaling Policy (see page 8).

Model weight security: It is challenging to define "sufficiently strong security" for model weights because strong security tends to require many different practices: any one weak link in the chain can dramatically worsen overall security.[32] As a starting point, a team at RAND has published guidelines on the level of security needed to protect model weights reliably from different types of actors, and their guidelines feature prominently in both Google DeepMind's Frontier Safety Framework (page 3) and Anthropic's Responsible Scaling Policy (page 21).

Enforcement and Accountability

This section has discussed how an if-then commitment might be designed. There is a separate question of how to ensure that a commitment is actually adhered to—for example, how to ensure that evals are run correctly, results are interpreted reasonably, and protections are implemented effectively.

Existing voluntary commitments by AI companies already contain some provisions on this front. For example, two companies' policies discuss looping in the board of directors and/or company at large on key decisions and reasoning.[33] Doing so has the potential to increase

the number of eyes on these decisions and make it more likely that noncompliant practices will be noticed by someone. These policies also discuss intentions to commission audits from external parties, which could provide further scrutiny and accountability.[34]

In the long run, the success of if-then commitments will likely depend on whether an ecosystem of qualified external auditors emerges, and on whether if-then commitments become backed by regulations (not just voluntary commitments). The following is an example timeline of how things might proceed from here:

During the next one to two years, increasing numbers of institutions may publish voluntary if-then commitments. This could include not just AI companies but also governments and civil society institutions. AI safety institutes may act in an advisory capacity to articulate where they think the tripwires should be and what risk mitigations need to accompany tripwire capabilities.

Simultaneously, organizations that already have laid out if-then commitments may move forward with implementing the needed evals, risk mitigations, and other procedures; learn about what approaches to this are and are not practical; and iterate toward better-designed processes for upholding if-then commitments.

Starting in one to two years, there could be increasing emphasis on formal industry standards (for example, ISO standards), as well as third-party audits and oversight to ensure that organizations are adhering to the if-then commitments they have made.

Once the relevant practices mature to the point of being usable for formal standards (imaginably as soon as two or so years from now), policymakers will be in a position to create regulations based on if-then commitments that have proven practical to implement.[35]

Other Possible Tripwires for If-Then Commitments

Although the above discussion has focused on a particular set of risks from chemical and biological weapons, voluntary commitments have included references to a number of other risks, including the potential ability of future AI models to assist with cyberoffense or persuasion, or to autonomously complete high-stakes and potentially dangerous tasks.[36] A future piece will discuss some potential criteria for choosing appropriate "tripwire capabilities" and if-then commitments across these categories, and sketch a set of candidate tripwires that could trigger if-then commitments.

Applying this Framework to Open Model Releases

Some AI models are released as "open models," which means their weights are made public. This practice can have enormous benefits[37] but may also have risks for models with strong enough capabilities—such as the tripwire capability above (pertaining to chemical and biological weapons). Given that an open model would allow anyone to effectively remove or circumvent deployment safety measures (at least with current technology), there is a case for an if-then commitment along the lines of "*if* an AI model has the tripwire capability detailed above, *then* it cannot be released as an open model." That said, open models have especially big potential benefits for the world at large, and these need to be weighed alongside risks. Some catastrophic risks might be significant enough to justify improving security and deployment restrictions for proprietary AI models, but *not* significant enough to justify forfeiting the benefits of making a model's weights widely available to the public.

Whether any particular if-then commitment makes sense for open models is an open question. But the general framework of if-then commitments could hold significant promise[38] for moving on from polarized debates about whether open models are "good" or "bad," and instead focusing on questions like: *What are the tripwires, and how do we test for whether AIs have crossed them?*

Limitations and Common Concerns About If-Then Commitments

If-then commitments offer a number of potential benefits, but it is necessary to acknowledge some limitations and drawbacks to the framework.

If-then commitments are very new, with little mature science to draw on. The first companies to release policies along the lines of if-then commitments did so in late 2023.[39] Before that, there was little discussion of the sorts of ideas covered in this piece. There are huge open questions around how big the risks discussed in these policies are, what risks have been left out, how to determine whether an AI model has particular dangerous capabilities, how to determine whether risk mitigations are sufficient, and more.

Work on if-then commitments should be thought of as "experimenting and prototyping." Many of the evaluations and risk mitigations people focus on today could look ill-conceived after just another year or two of learning and iteration. In this spirit, if AI progress is fast enough to open up extremely dangerous capabilities in the next few years, one should not assume that if-then commitments will be well-developed enough by then to be up to the task of containing the risks (though they may help).

Voluntary commitments alone are unlikely to keep risks low. As of today, if-then commitments have come from voluntary corporate policies and frameworks, with little third-party oversight or enforcement. This is valuable for early experimentation with a young framework, but in the long run, one should not expect voluntary commitments to stop companies from racing toward large commercial opportunities. And one *should* expect any given set of AI capabilities to get cheaper and easier to produce over time, bringing in more players and making it less likely that everyone will be adhering to any given set of practices.

In the long run, more than voluntary commitments will be needed in order for this framework to work. It will be necessary to have regulation and likely even international coordination. Voluntary commitments—and the public dialogue around them, including criticisms and recommendations for improvements—can be an important source of information on how to conduct evaluations and implement risk mitigations. At this early stage, this may be the fastest way to accumulate such knowledge. Ultimately, however, AI developers should be regulated more strictly than how they would regulate themselves.

It will probably never be possible to fully rule out that a given AI has tripwire capabilities. Today, it seems that even the best AI models are not close to the tripwire discussed above, or to other tripwires that have been proposed.[40] But there are always questions as to whether evaluations reflect what an AI model is really capable of, and what capabilities a given AI model might acquire over the next several years following advances in post-training enhancements. For some who think that the risks of AI are huge and imminent, it is unlikely that the framework discussed here can be conservative enough in handling those sorts of possibilities.

On the flip side, it also is hard to know whether specific risks outweigh the associated benefits. The effects of innovation are inherently hard to predict, and there are some who feel that weighing costs and benefits ahead of time will never be a useful exercise.

If-then commitments are not a good fit for all risks from AI. They are designed primarily for domains where *prevention* (as opposed to *response*) is a feasible and important part of risk management. Some risks—particularly those that build up over many relatively smaller incidents, as opposed to a small number of discrete catastrophes—may be too hard to anticipate or prepare for in advance and may best be handled by noticing and reacting to risky dynamics rather than focusing on precommitments.

It is hard to anticipate risks in advance. The most important short- and long-term risks from AI are not necessarily the same ones that are getting attention and analysis today. Today's if-then commitments might look ill-conceived or irrelevant in the future, while risks that have received little attention (including risks that no one has thought of yet) might turn out to be more important. If-then commitments are fundamentally about trying to prepare for risks from AI models that do not exist yet. This is an inherently difficult exercise, though perhaps it is necessary if dangerous AI capabilities could emerge rapidly, while key risk mitigations take a long time to develop.

That said, perfect foresight is not needed for if-then commitments to be useful. For example, many different potential threat models call for similar risk mitigations (such as strong security for model weights), and it seems plausible that these risk mitigations would be robustly useful for risks that are not yet on anyone's radar.

Voluntary commitments and even regulation could be too hard to enforce across the board—such that responsible actors end up adhering to if-then commitments, while irresponsible actors rush forward with dangerous AI. One of the challenges with AI is that *complete enforcement* of any given risk mitigation framework seems extremely hard to achieve, yet *incomplete enforcement* could end up disadvantaging responsible actors in a high-stakes, global technology race. This is a general issue with most ways of reducing AI risks, other than "race forward and hope that the benefits outweigh the costs," and is not specific to if-then commitments.

To help mitigate this issue, early, voluntary if-then commitments can contain "escape clauses" along the lines of: "We may cease adhering to these commitments if some actor who is not adhering to them is close to building more capable models than ours." (Some more detailed suggested language for such a commitment is provided by METR, a nonprofit that works on AI evaluations.)[41] Today, it appears likely that the most capable AI models of the next generation will be built by a relatively small number of AI developers that have shown interest in if-then commitments, so the situation contemplated by an "escape clause" is hopefully not imminent. Over time, it will be important to build increasingly wide consensus and strong enforcement.

It also is worth noting that there could be a similar problem with "irresponsible actors having an advantage in a race" if AI developers fail to implement strong enough security for their model weights. In this case, actors that are willing and able to steal model weights and run the resulting AI models with few precautions may gain an advantage. A major goal of if-then commitments is to provide a stronger push toward improving security to the point where it could resist even attacks from foreign intelligence services.[42]

The Path to Robust, Enforceable If-Then Commitments

The framework discussed in this piece is nascent. To date, only a handful of organizations have published if-then commitments, all within the last year, and most emphasize how preliminary they are.[43] Much work remains to be done to build mature fields of AI risk assessment to identify tripwires, capability evaluations to determine when tripwires have been crossed, and risk mitigations for AI models that have tripwire capabilities. An example timeline for how this progress might be made is given in an earlier section.

Today, a number of institutions have potential roles to play in accelerating the initial adoption, iteration, and improvement of if-then commitments:

AI companies can voluntarily adopt if-then commitments—and those that have put out frameworks along these lines can *continually refine* them. Google DeepMind's Frontier Safety Framework, for example, ends with a specific list of issues it plans to address more thoroughly in future versions. OpenAI's Preparedness Framework is marked "Beta," indicating that it too is not a final product. Anthropic's Responsible Scaling Policy includes a commitment to define further "AI safety levels" in the future.

AI safety institutes (such as those in the United Kingdom and United States) can put out their own nonbinding guidance on the types of if-then commitments that AI developers should adopt. Other civil society organizations can do similarly. This could provide a valuable check on the choices made by for-profit companies—a comparison point with more ambitious risk reduction measures than companies voluntarily have been willing to adopt to date.

Subject-matter experts in areas such as chemistry, biology, and cybersecurity can create and refine proposals for which AI capabilities should be considered tripwires, which evals would be most informative about them, and the like. Additionally, experts in relevant areas can work on things like standards for information security and technologies for making AI models harder to "jailbreak."

Policymakers have opportunities to encourage AI companies and AI safety institutes to take the steps above. This can include regulatory incentives for companies to develop their own if-then commitments (though it is likely too early to prescribe adherence to *specific* if-then commitments). It can also include simply emphasizing and asking about how the relevant institutions are thinking about where their tripwires are, and what if-then commitments they are ready to make or recommend, whether in hearings, letters, informal meetings, or other venues.

Finally, **any and all parties** can show an interest in the evolving framework of if-then commitments. Simply asking questions (ranging from "When do you plan on releasing the next iteration of your voluntary commitments?" to "Are you thinking of adding evals for risk X?") can help demonstrate that people are paying attention to the commitments and recommendations organizations are issuing—and that they would value progress toward a mature framework that could robustly reduce risk while continuing to encourage innovation.

Appendix: Elaborating on the Risk of AI-Assisted Chemical and Biological Weapons Development

This appendix briefly elaborates on the threat model featured (as an illustrative example) in the main text.

There are a number of chemical or biological weapons that someone with the relevant experience and expertise could produce and deploy on a relatively modest budget and without needing access to any particularly hard-to-obtain materials.[44] 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 other information. (There are ongoing efforts to test this claim, as discussed in the main text.)

Fortunately, only a small percentage of the population has the expertise needed to develop a given chemical or biological weapon, and the overlap with people who would *want to* is even smaller.[45] But if a (future) AI model could play the same role as a human expert in chemical or biological weapons, then any individual (such as a terrorist) with access to that AI model effectively would have access to an expert advisor (note that there is precedent for terrorists' attempting to produce and deploy chemical and biological weapons in an attempt to cause mass casualties[46]).

Thus, widely available and capable enough AI could effectively give any determined user access to an adviser with the most relevant expertise—greatly multiplying the number of people with the ability to deploy a weapon of mass destruction.

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 is possible that sufficient restrictions on access to key precursor materials and technologies—for example, DNA synthesis—could make chemical and/or biological weapons infeasible to produce even with strong expertise or expert advice. No AI risk is *only* about AI, but it may still be prudent to prepare for the potential sudden emergence of AI capabilities that would cause major risks in the world as it is.

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.

Acknowledgements

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 and Luca Righetti for especially in-depth comments; Ella Guest and Greg McKelvey for comments on the discussion of chemical and biological weapons; and my Carnegie colleagues, particularly Jon Bateman, Alie Brase, and Ian Klaus, for support on the drafting and publishing process. Finally, I note that the "if-then commitments" term is due to this paper.

Notes

1 That being said, there are questions as to whether new forms of post-training enhancements could change this in the future.

2 From this announcement. Key text:

"II. Set out thresholds at which severe risks posed by a model or system, unless adequately mitigated, would be deemed intolerable. Assess whether these thresholds have been breached, including monitoring how close a model or system is to such a breach. These thresholds should be defined with input from trusted actors, including organisations' respective home governments as appropriate. They should align with relevant international agreements to which their home governments are party. They should also be accompanied by an explanation of how thresholds were decided upon, and by specific examples of situations where the models or systems would pose intolerable risk.

III. Articulate how risk mitigations will be identified and implemented to keep risks within defined thresholds, including safety and security-related risk mitigations such as modifying system behaviours and implementing robust security controls for unreleased model weights.

IV. Set out explicit processes they intend to follow if their model or system poses risks that meet or exceed the pre-defined thresholds. This includes processes to further develop and deploy their systems and models only if they assess that residual risks would stay below the thresholds. In the extreme, organisations commit not to develop or deploy a model or system at all, if mitigations cannot be applied to keep risks below the thresholds."

3 See the Beijing statement at https://idais.ai/.

4 See this speech (and tweets) from UK Secretary of State for Science, Innovation and Technology Michelle Donelan in the leadup to the summit. "Responsible Capability Scaling" also appears in the program.

5 Specifically, the author of this piece collaborated with METR (Model Evaluation and Threat Research) to define and make a public case for responsible scaling policies (though different institutions generally have used different terms for similar ideas since then).

6 For example:

Section 3(k) of a late 2023 U.S. executive order raises the idea of AI that could "substantially [lower] the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological,

or nuclear (CBRN) weapons," "[enable] powerful offensive cyber operations" or "[permit] the evasion of human control or oversight through means of deception or obfuscation."

A declaration signed by 29 countries states "We are especially concerned by such risks in domains such as cybersecurity and biotechnology." (A similar, later international statement states "We recognise that such severe risks could be posed by the potential model or system capability to meaningfully assist non-state actors in advancing the development, production, acquisition or use of chemical or biological weapons, as well as their means of delivery."

7 "Current Artificial Intelligence Does Not Meaningfully Increase Risk of a Biological Weapons Attack," RAND Corporation, January 25, 2024, https://www.rand.org/news/press/2024/01/25.html.

8 "Advanced AI Evaluations at AISI: May Update," AI Safety Institute, May 20, 2024, https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

9 For example, one might try removing data relevant to chemical and biological weapons from an AI's training data. But it could be difficult to find all relevant data, and removing that data might hurt the AI's general facility with chemistry and/or biology. Additionally, any such removal would have to be done in full before a training run; training runs are time-consuming and expensive, and redoing them to remove some additional data likely would be very costly. There may be ways to get AI models to "unlearn" particular knowledge post-training, but these are not yet well-established. See Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai, "A Review on Machine Unlearning," *SN Computer Science* 4, no. 337 (April 19, 2023), https://doi.org/10.1007/s42979-023-01767-4.

10 For example, see Pranav Gade, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "BadLlama: Cheaply Removing Safety Fine-tuning from Llama 2-Chat 13B," Arxiv, May 28, 2024, https://arxiv.org/abs/2311.00117; and Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B," Arxiv, May 22, 2024, https://arxiv.org/abs/2310.20624; as well as Maxine Labonne, "Uncensor Any LLM with Abliteration," Hugging Face, June 13, 2024, https://huggingface.co/blog/mlabonne/abliteration.

11 Said malicious actor likely would have a college education, a moderate amount of time and money—perhaps $50,000 and six months—but no deep relevant expertise or experience.

12 For instance, greater than 10 percent.

13 This leaves open the precise threshold for what damages would count as catastrophic. For one reference point, a paper on the benefits of advance preparations for future pandemics states: "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)." See Rachel Glennerster, Christopher M. Snyder, and Brandon Joel Tan, "Calculating the Costs and Benefits of Advance Preparations for Future Pandemics," NBER Working Paper 30565, rev. June 2023, https://www.nber.org/system/files/working_papers/w30565/w30565.pdf.

14 "If we reach (or are forecasted to reach) at least 'high' pre-mitigation risk in any of the considered categories we will ensure that our security is hardened in a way that is designed to prevent our mitigations and controls from being circumvented via exfiltration (by the time we hit 'high' pre-mitigation risk). This is defined as establishing network and compute security controls designed to help prevent the captured risk from being exploited or exfiltrated, as assessed and implemented by the Security team." From page 20 of the OpenAI Preparedness Framework (Beta).

15 In particular, the requirement that the "post-mitigation" risk of a deployed model be "medium" or below implies that mitigations are used to prevent users from accessing "high"-risk capabilities.

16 "High" capability (page 9): "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 threat."

"Critical" capability (page 9): Model enables an expert to develop a highly dangerous novel threat vector (e.g., comparable to novel CDC Class A biological agent) OR model provides meaningfully improved assistance that enables anyone to be able to create a known CBRN threat OR model can be connected to tools and equipment to complete the full engineering and/or synthesis cycle of a regulated or novel CBRN threat without human intervention."

The second part of the "High" capability is very similar to the "tripwire" listed in this piece, with a bit less detail and a slightly higher starting knowledge level for the malicious actor (a biology background rather than just a college education). The second part of the "Critical" capability is a somewhat more extreme version of the tripwire given in this piece, as it refers to "anyone" being able to design a CBRN weapon "without human intervention."

17 "Complying with higher ASLs is not just a procedural matter, but may sometimes require research or technical breakthroughs to give affirmative evidence of a model's safety (which is generally not possible today), demonstrated inability to elicit catastrophic risks during red-teaming (as opposed to merely a commitment to perform red-teaming), and/or unusually stringent information security controls. Anthropic's commitment to follow the ASL scheme thus implies that we commit to pause the scaling and/or delay the deployment of new models whenever our scaling ability outstrips our ability to comply with the safety procedures for the corresponding ASL." From page 2 of Anthropic's Responsible Scaling Policy.

18 "**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." The corresponding "Rationale" in the table states: "Many biothreats capable of causing significant amounts of harm are currently out of the reach of non-experts because of lack of knowledge about their potential for harm and the methods of their acquisition and misuse. An LLM that helps overcome these knowledge gaps, e.g. by suggesting plausible attack strategies or providing detailed instructions for the development of a bio agent, could significantly increase society's vulnerability to fatal attacks by malicious amateurs." From page 5 of Google DeepMind's Frontier Safety Framework.

19 See the tables on pages 3–4 of Google DeepMind's Frontier Safety Framework.

20 "As we better understand the risks posed by models at different CCLs, and the contexts in which our models will be deployed, we will develop mitigation plans that map the CCLs to the security and deployment levels described." From page 6 of Google DeepMind's Frontier Safety Framework.

21 From the National Institute of Standards and Technology (NIST) landing page: "Our efforts will initially focus on the priorities assigned to NIST under President Biden's Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The Safety Institute will pursue a range of projects, each dedicated to a specific challenge that is key to our mission; these will initially include advancing research and measurement science for AI safety, conducting safety evaluations of models and systems, and developing guidelines for evaluations and risk mitigations, including content authentication and the detection of synthetic content."

22 See the "Research" section of the UK Department of Science, Innovation and Technology AI Safety Institute's About page.

23 For examples of evals run at AI companies, see model cards for GPT-4, Gemini 1.5, and Claude 3.

24 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.

25 From the report accompanying the Gemini 1.5 release, page 68:

"We performed evaluations on a number of capabilities relevant to extreme risks. . . . Our internal CBRN evaluations are still nascent and to date, three different evaluation approaches have been used, all of which are complementary to the external red-teaming performed by third party organisations. Biological and radiological/nuclear information have been assessed using in-house developed approaches: 1) a qualitative approach with open-ended adversarial prompts and domain-expert raters; and 2) a quantitative approach based on closed-ended, knowledge-based multiple choice questions. A third approach is used for the chemical information evaluations which is based on closed-ended knowledge based approach regarding chemical hazards without human raters (developed by Macknight et al.). Preliminary results for the

qualitative results indicate that the frequency of refusals from the model is increased compared to previous models. The performance of Gemini 1.5 Pro for the quantitative results has not improved compared to previous models.

From the Claude 3 model card, page 25: "Our biological evaluations involve the model answering a series of questions on relevant technical knowledge that could cause harm. We also complement these automated evaluations with human uplift trials – testing whether a group with access to Claude 3 models have more success answering harmful biological questions than a control group with access to Google.

"Based on conversations with global experts, it is difficult to define strict pass/fail criteria for ASL-3 misuse evaluations with high confidence. Instead, we set the bar relatively low, such that passing the misuse evaluations would trigger discussion with relevant experts and extensive transcript reviewing to determine whether the model presents a true risk or the thresholds are too conservative . . . .

"The model did not cross the thresholds above. Our human uplift trial found what we believe is a minor uplift in accuracy, and a decrease in time spent, from using the model without safeguards as compared to using internet search only. There was no change in either measure for the group with safeguards. For biological risks, we are increasingly confident in using human uplift trials as highly informative measures of marginal risk from models.

"In automated biology evaluations, we found a mix of results. On one new multiple choice evaluation designed to assess model capabilities relevant to biological risks, we noticed Opus performed better than Claude 2.1, though underneath our trigger threshold. However, on other experimental evaluations about biological design, Opus performed worse, suggesting that we may have under-elicited the model's capabilities. Both sets of evaluations are novel and experimental, and we believe need to be refined and further explored.

"Alongside other science evals, we also run four automated multiple choice question sets which are not used as ASL-3 indicators, but which are helpful indicators of related model performance. We use PubmedQA [23], BioASQ [69], USMLE [70], and MedMCQA [71]. The model performed up to around 10% better than Claude 2.1 on these, although in two cases showed lower results. Similar to the results above, this would suggest some under-elicitation of the model's capabilities.

"In summary, the model did not meet our most conservative biological risk thresholds, and our expert consultants agreed. We will now be expanding evaluations and more tightly defining our biological risk threshold."

OpenAI's Preparedness Framework (page 17) gives example hypothetical eval results for illustrative purposes: "Model can provide information relevant to creating CBRN threats with comparable utility to existing resources (eg, textbooks, search engines), but nothing more than that.

GPT vs search eval: Post-PhD professionals trained in biology (specifically genetics, virology, or microbiology experience no improvement in success rates for writing CDC Class B biological agent development instructions using a search engine and GPT versus just a search engine (X% vs X%, including across accuracy, completeness, and innovation metrics. This holds true for chemistry, nuclear, and radiological post-PhD professionals students as well.

GPT vs search eval: Undergraduate students with some coursework in biology (specifically genetics, virology, or microbiology experience no improvement in success rates for writing CDC Class B biological agent development instructions using a search engine and GPT versus just a search engine (X% vs X%, including across accuracy, completeness, and innovation metrics. This holds true for chemistry, nuclear, and radiological study undergraduate students as well.

GPT vs search identify suppliers eval: <X% more participants succeed in finding DNA suppliers that do not screen orders using a search engine and GPT versus just a search engine."

26 Though with access to other tools such as search engines.

27 An early paper in this genre is Toby Shevlane et al., "Model evaluation for extreme risks," Arxiv, May 24, 2023, https://arxiv.org/abs/2305.15324. It states: "The AI community already relies heavily on model evaluation ... We propose extending this toolbox to address risks that would be extreme in scale, resulting from the misuse or misalignment of general-purpose models. Work on this new class of model evaluation is already underway."

28 OpenAI's Preparedness Framework, page 13: "We want to ensure our understanding of pre-mitigation risk takes into account a model that is 'worst known case' (i.e., specifically tailored) for the given domain. To this end, for our evaluations, we will be running them not only on base models (with highly-performant, tailored prompts wherever appropriate), but also on fine-tuned versions designed for the particular misuse vector without any mitigations in place. We will be running these evaluations continually, i.e., as often as needed to catch any non-trivial capability change, including before, during, and after training. This would include whenever there is a >2x effective compute increase or major algorithmic breakthrough."

Google DeepMind's Frontier Safety Framework, page 6: "We are working to equip our evaluators with state of the art elicitation techniques, to ensure we are not underestimating the capability of our models."

Anthropic's Responsible Scaling Policy, page 12: "An inherent difficulty of an evaluations regime is that it is not currently possible to truly upper-bound the capabilities of generative models. However, it is important that we are evaluating models with close to our best capabilities elicitation techniques, to avoid underestimating the capabilities it would be possible for a malicious actor to elicit if the model were stolen."

29 For example, see David Rein et al. "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," Arxiv, November 20, 2023, https://arxiv.org/abs/2311.12022 .

30 Google DeepMind's Frontier Safety Framework, page 2: "we will design early warning evaluations to give us an adequate safety buffer before a model reaches a [critical capability level]."

Anthropic's Responsible Scaling Policy, page 11: "Ensuring that we *never* train a model that passes an ASL evaluation threshold is a difficult task. Models are trained in discrete sizes, they require effort to evaluate mid-training, and serious, meaningful evaluations may be very time consuming, since they will likely require fine-tuning. This means there is a risk of overshooting an ASL threshold when we intended to stop short of it. We mitigate this risk by creating a *buffer*: we have intentionally designed our ASL evaluations to trigger at slightly lower capability levels than those we are concerned about, while ensuring we evaluate at defined, regular intervals (specifically every 4x increase in effective compute, as defined below) in order to limit the amount of overshoot that is possible. We have aimed to set the size of our safety buffer to 6x (larger than our 4x evaluation interval) so model training can continue safely while evaluations take place. Correct execution of this scheme will result in us training models that just barely pass the test for ASL-N, are still slightly *below* our actual threshold of concern (due to our buffer), and then pausing training and deployment of that model unless the corresponding safety measures are ready." (More detail follows.)

31 Anthropic's Responsible Scaling Policy pages 6–9 describes an "ASL-32 standard for deployment safety and security. Google DeepMind's Frontier Safety Framework pages 3–4 lays out different levels of "Security mitigations" and "Deployment mitigations.2 OpenAI's Preparedness Framework pages 20–21 discusses possible measures for improving information security at a high level. It does not give detail on deployment safety measures, but states: "Only models with a post-mitigation score of 'medium' or below can be deployed. In other words, if we reach (or are forecasted to reach) at least 'high' pre-mitigation risk in any of the considered categories, we will not continue with deployment of that model (by the time we hit 'high' pre-mitigation risk) until there are reasonably mitigations in place for the relevant postmitigation risk level to be back at most to 'medium' level. (Note that a potentially effective mitigation in this context could be restricting deployment to trusted parties.)"

32 This point is argued at Sella Nevo et al., "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models," RAND Corporation, May 30, 2024, https://www.rand.org/pubs/research_reports/RRA2849-1.html.

33 OpenAI's Preparedness Framework, page 24: "Internal visibility: The Preparedness Framework, reports and decisions will be documented and visible to the BoD and within OpenAI (with redactions as needed given internal compartmentalization of research work)." Anthropic's Responsible Scaling Policy, page 10: "we additionally make the following procedural commitments . . . 6. Share results of ASL evaluations promptly with Anthropic's governing bodies, including the board of directors and LTBT, in order to sufficiently inform them of changes to our risk profile . . . 8. Implement a non-compliance reporting policy for our Responsible Scaling Commitments as part of reaching ASL-3. The policy should allow for anonymous feedback, with an appropriate reporting chain."

34 OpenAI's Preparedness Framework, page 25: "Scorecard evaluations (and corresponding mitigations) will be audited by qualified, independent third-parties to ensure accurate reporting of results, either by reproducing findings or by reviewing methodology to ensure soundness, at a cadence specified by the SAG and/or upon

the request of OpenAI Leadership or the [board of directors]." Anthropic's Responsible Scaling Policy, page 15: "External verification: Due to the large potential negative externalities of operating an ASL-4 lab, verifiability of the above measures should be supported by external audits."

35 In the meantime, policymakers can nudge relevant organizations to put more work into developing and iterating on if-then commitments, without yet prescribing specific practices.

36 See OpenAI's Preparedness Framework, pages 8–11; Google DeepMind's Frontier Safety Framework, pages 5–6; and Anthropic's Responsible Scaling Policy, pages 6–7.

37 Open models can accelerate innovation generally, by giving a wide range of actors the ability to experiment with many different ways of building on a given AI model. In particular, open models can be helpful for research on potential risks from AI and on risk mitigations. Given how expensive state-of-the-art AIs are to train, there is a general risk that researchers are becoming reliant on AI companies for model access, which could cause distortive power dynamics—for example, making it hard for researchers to provide neutral takes on AI risk and how AI companies are handling it, and/or making it costly for researchers to criticize AI companies. Open models have the potential to ameliorate this dynamic.

38 Representatives of Meta—probably the best-known and best-resourced company focused on open models—have stated that Meta is not committed to releasing model weights in all cases, that there are imaginable situations where dangerous AI capabilities would make it irresponsible to do so, and even that it (Meta) is working on "no-go lines." Several of the most prominent companies focused on open models (Meta, Mistral, xAI) have all joined in the recent commitment by 16 companies to develop frontier safety policies, using a framework much like the one discussed in this piece.

39 Anthropic announced its Responsible Scaling Policy in September 2023. OpenAI published its beta Preparedness Framework in December 2023.

40 As argued in model cards for major AI model releases; see model cards for GPT-4, Gemini 1.5, and Claude 3.

41 "In the event that we have strong reason to think that other AI developers are moving forward with comparably dangerous AI models, and we have exhausted other avenues for reducing the associated risks, we might make an exception to the above plan and continue development—while working with states or other authorities to take immediate actions to limit scaling that would affect all AI developers (including us). We would consider this a dire situation. We would seek input on our choices from the US government, and would be explicit—with employees, our board, and state authorities—that our scaling was no longer safe, and that we should be accountable for the judgment to proceed."

42 It is probably not possible for security to be *impenetrable* by foreign intelligence services, but making theft more difficult seems both useful and possible. See Sella Nevo et al., "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models," RAND Corporation, May 30, 2024, https://www.rand.org/pubs/research_reports/RRA2849-1.html.

43 "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." See 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." See 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." See Anthropic's blog post introducing its Responsible Scaling Policy.

44 Regarding chemical weapons, see R. E. Ferner and M. D. Rawlins, "Chemical Weapons," *BMJ* 298, no. 6676 (March 25, 1989): 767–768, https://doi.org/10.1136%2Fbmj.298.6676.767. Regarding biological weapons, this view is debated among experts, but for an example of experts seemingly endorsing a similar view, see National Academies of Sciences, Engineering, and Medicine, *Biodefense in the Age of Synthetic Biology* (Washington, DC: The National Academies Press, 2018), https://doi.org/10.17226/24890: "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."

45 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 based on the specific weapon and is likely higher for chemical than for biological weapons.

46 There is precedent for terrorists attempting to produce and deploy chemical and biological weapons in an attempt to cause mass casualties. For example, see Manuela Oliveira et al., "Biowarfare, Bioterrorism and Biocrime: A Historical Overview on Microbial Harmful Applications," *Forensic Science International* (September 2020), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7305902/, section 1.2. The case of Aum Shinrikyo is particularly interesting owing to the amount of effort and expense invested in relatively challenging (at the time) chemical and biological weapons production projects, although these were ultimately unsuccessful. See Richard Danzig et al., *Aum Shinrikyo: Insights into How Terrorists Develop Biological and Chemical Weapons*, 2nd ed., CNAS, December 20, 2012, https://www.cnas.org/publications/reports/aum-shinrikyo-second-edition-english.

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