Artificial Intelligence Safety and Security License Requirements

Artificial Intelligence (AI) is now broadly capable, which can produce both benefits and dangers. AI will hasten the design and proliferation of bioweapons, cyberweapons, nuclear weapons, progressively more general intelligence, and other threats not yet conceived. Unfortunately, the full capabilities and threats from AIs are unknown even to their creators; each new generation of AIs sees new capabilities emerge that were not predicted, and the most general AIs can be fine-tuned by later users to elicit specific new dangerous capabilities. Today society learns of the threat from an AI only after it has already proliferated. Despite companies' attempts at safety, recently released AIs are capable of rapidly creating novel malware and automated synthesis of chemical weapons. This situation is recognized as untenable by ever more segments of society, including AI developers who are calling for government oversight.

Government involvement with all of AI is neither possible nor useful. However, for the most powerful AIs at the technical frontier, governance is both possible and critical for safety and security. The United States should create oversight of powerful AI development and require safety and security licenses before dangerous AIs are created or proliferated. This can be done in three steps.

1. *AI Hardware*: track and license a critical mass of hardware for making powerful AI
2. *AI Creation*: track and license the creation of powerful AI
3. *AI Proliferation*: license the dissemination of powerful AI

Addressing all three is needed for an effective layered defense. *AI Proliferation* is as easy as anonymously uploading a file to the Internet, and so difficult to oversee without first overseeing *AI Creation*, which is itself difficult to oversee without first monitoring *AI Hardware*. See *Appendix* for a layperson technical explainer.

Needed End State: Computer chips for making powerful AI are tracked, and gathering a critical mass (e.g. >1,000 chips) requires a license. The license certifies the chip holder has sufficient cybersecurity and will monitor and report if the chips are used for a large AI training run (e.g. >10^27 bit operations; just beyond the current frontier, costing >$10M). Large AI training runs require a license certifying the training will use appropriate methods to reduce the chance the resulting AI will generate classified nuclear weapons information, design bioweapons or cyberweapons, or escape as a computer worm. Once an AI is created it goes through a timely, independent evaluation to certify that safety and security requirements were met, yielding a proliferation license. License requirements are lower if the licensee is only providing users interaction with the AI, whose safety can be later improved if needed (e.g. ChatGPT/Bing/Bard/Claude web interfaces) vs. proliferating the entirety of the AI (e.g. open-sourced AIs, as well as Meta's leaked LLaMa AI), which others can permanently modify to be maximally dangerous.

What Probable Success Looks Like: Like the nuclear nonproliferation regime, AI oversight will be imperfect but still critically constrain the proliferation of technologies of mass destruction

while enabling beneficial uses. Tracking and licensing *AI Hardware* will require administrative effort, but involves well-understood computing hardware. In contrast, license reviews for *AI Creation* and *Proliferation* will require evaluating AIs for safety and security, which is a complex problem at the leading edge of research. License reviews must be informed by the latest research, and the inspections themselves will ideally advance the knowledge frontier of safety and security best practices. The right blend of technical acumen for effective reviews will require a close collaboration between the public and private sector.

Private Sector Successes and Gaps: Multiple private actors are starting safety and security reviews. For example, OpenAI published safety test results for GPT-4, some run by independent review organizations; DeepMind, Anthropic and others are establishing similar efforts. Current review techniques include provoking bad behavior from either the unaltered AI, the AI augmented with other tools (e.g. access to the Internet), or the AI fine-tuned for a dangerous purpose with a small amount of additional computation (easily doable by malicious actors after the AI proliferates). These techniques are useful but currently insufficient in scale or scope; successful reviews will include experts in bioweapons, cyberweapons, and nuclear weapons, which requires government involvement. Additionally, safety and security review is currently optional; many AI developers do without them, and all are incentivized to drop safety in the race of competition. Lastly, finding problems is different from fixing problems; known holes are left unplugged because identifying a solution is perceived as too costly for the developer relative to the business advantage the developer would receive, even though the costs to society could be enormous. Many are calling for government oversight to solve these market failures.

Foreign Adversaries: Both China and Russia will be limited in making powerful, general-purpose AIs themselves cost-effectively because of recent export controls on *AI Hardware*. However, they can both try to abuse or steal American AI capabilities. Russian cybercriminals are bypassing ChatGPT's safety and security systems to create malware. Chinese intelligence must find tempting the prospect of simply stealing the entirety of the AI from OpenAI's servers. Many AI companies are startups and know they lack the expertise to confront attacks by nation states. Government safety and security tracking and licensing can mitigate these threats by defining and requiring sufficient cybersecurity methods to protect from theft and sufficient safety methods to protect from misuse.

Maintaining Innovation: Only by making powerful AI trustworthy can it be deployed into many high-stakes uses in the economy and national security. As such, the future of the AI industry hinges on whether powerful AIs are made safe and secure, which is why increasingly many firms are calling for government oversight to eliminate the temptation to compromise safety and security. Thankfully, oversight is only needed for the most powerful, dangerous AIs, which are made by a small number of well-resourced actors. For those few AI developers that are covered, safety and security reviews will be an incentive for innovation: by keeping license requirements adaptable and allowing AI creators to come up with new ways to achieve safety, oversight will create innovative pressure to develop new methods for making AIs demonstrably safe and secure, which will create many positive spillovers and enable broader use of powerful AI.

Options to Reach the End State: There are three options to achieve AI oversight. Each option is imperfect, so all three should be done in concert to cover gaps. All options should be taken immediately; full implementation will likely take two years, while dangerous AI continues to advance and proliferate. Action today is needed to cover the threats of tomorrow; the threats of today are already lost.

	AI Hardware		AI Creation		AI liferation	d		ent
n of Funding	✓	✓	✓	✓	✓	Global		s
ies	✓	✗	~	~	~			rs
New ies	ally ✓	ally ✓	ally ✓	ally ✓	deally ✓	All, US		e created

Condition of Federal Funding: Require compliance with an AI oversight system for those receiving federal contracts and grants, as well as their customers. This is the only current authority to create oversight over all of *AI Hardware, Creation* and *Proliferation*. Federal funding is a frequently-used policy lever, such as the Common Rule stipulating how recipients can perform biomedical research on human subjects, not just for their federally-funded research, but *all* research they perform. For AI oversight, requirements could apply both to federal funding recipients *and* their customers. The U.S. government is a major customer of a few companies that provide computing hardware and cloud services, who are in turn essential suppliers for AI developers; thus, the new oversight requirements created by federal contracts to computing vendors would propagate to many AI developers, though not all.

New requirements would likely require notice and comment rulemaking and an interagency process to coordinate a common requirement across agencies and update existing contracts. This will likely take at least 2 years, but the timeline could be shortened through starting with the largest government customers of DOE, DOD and DOC and leveraging several existing processes like NTIA's recent Request for Comment on AI Accountability or CISA's effort on AI acquisition standards.

Existing Authorities: Several existing authorities could plausibly be used to cover parts of AI oversight, though not robustly.

1. Tracking *AI Hardware* and *AI Creation*: 50 USC 4555 gives the President authorities to obtain information from industry to support the national defense, which includes protecting critical infrastructure and combating terrorism. These authorities can be used to require the private sector to track and report powerful *AI Hardware* and perhaps also *AI Creation.* This would be a novel, though credible, use of the authorities.

2. Licensing *AI Creation* and *AI Proliferation*: The Atomic Energy Act makes data on nuclear weapons classified until declassified, which likely includes AIs that can produce such restricted data. AI developers should be alerted that their AIs may be classified, which should strongly incentivize them to participate in *AI Creation* and *Proliferation* license reviews. Additionally, export controls can be created to require a license for *AI Proliferation* for an AI above a threshold level of power; export license review would then include safety and security review. Controlling foreign sales of software is common, but controlling open source software like AI is rare, and unfortunately required to address proliferation risk. However, both nuclear- and export-controls-based oversight on *AI Proliferation* is difficult to enforce unless *AI Creation* is also overseen, since anonymously uploading to the internet even today's large AIs is technically trivial.

Create New Authorities: New, clear, explicit authorities could be created for tracking and licensing *AI Hardware*, *Creation* and *Proliferation* within the U.S. for the purpose of safety and security oversight, with criminal penalties for non-compliance.

Designated Agency: All three options can designate a single agency with the responsibility and authority for the actual tracking and licensing of *AI Hardware*, *Creation* and *Proliferation*. A single agency on point would enable building the critical mass of expertise needed to effectively integrate the technical and administrative elements of AI oversight operations; this expertise and oversight operations can be stood up while the new oversight requirements are being created. The agency may need to be formally delegated 50 USC 4555 authorities by the President to obtain information on *AI Hardware* and *Creation*, but all other *Existing Authorities* and the *Condition of Federal Funding* could be handled by the interagency.

International Context: The U.K.'s new AI policy of "life cycle accountability" aligns with the aims stated here. The E.U.'s AI governance efforts are now looking to target the threats from the general-purpose AIs described here. Even allied and partnered countries may not appreciate their own AI developers being captured under the *Condition of Federal Funding*, but the U.S. could offer to route such developers to their domestic governments for oversight, creating a governance hook the U.S. can offer to those other countries. China already has AI regulations, and may follow the U.S. in establishing oversight of dangerous AI; if not, U.S. control over AI chip exports can be used to require Chinese AI efforts comply with safety procedures. The U.S. is in a good position to lead.

Appendix

Technical Explainer of AI Life Cycle and Mechanics for Oversight

Today's AIs are large sets of equations connected to each other, described as a set of parameters. An AI starts as incapable, with a random set of parameters, and then it is trained on data that covers the tasks the AI should learn; today's most powerful AIs are trained on data from large swaths of the Internet, which contains adequate information to learn many, unanticipated tasks. During training, the AI's parameters change and its capabilities improve. An AI can be trained all at once or in segments, including first training the AI for one purpose and then later training it further to fine tune it for another purpose.

1. *AI Hardware* (and Data): tracking and licensing a critical mass of hardware for making powerful AI

Training an AI to have great capabilities typically requires great quantities of computer chips and great amounts of data. Computer chips for training AI are a physical, rivalrous good that can be tracked and licensed, and should be the primary focus for oversight. Data, in contrast, is typically public (e.g. free Internet content) or synthesizable, and data can be copied, making it a far weaker point for oversight. There are a few types of exquisite data that are likely accelerants for specific threats (e.g. bioweapon design), which should be identified and addressed as a separate effort.

Cost-effective AI training requires chips that are built for purpose, and training the most powerful AIs requires thousands of such chips. These same chips can be used for other purposes that only require one or two at a time, such as video gaming. Chips typically need to be located near each other and communicate over fast connections; training an AI with chips communicating over the Internet is possible, but brings increased cost that goes up with the capability of the AI created. Similarly, other chips not specialized for AI can also be used for training AI, but doing so brings increased cost. Thus, oversight over *AI Hardware* can be accomplished by tracking only chips above a specific capability threshold and requiring a license to bring together a critical mass of them. Today there are only a few relevant chip product lines on the market (e.g. the Nvidia H100) and only a few actors with the critical mass of chips need to train powerful AIs (e.g. over 1,000 chips). This is the same category of chips the U.S. is now controlling from proliferation to China and Russia.

Oversight of the hardware for AI is necessary because without it oversight at later phases can be evaded through undeclared computing hardware. However, merely tracking and licensing computing hardware is insufficient, because the danger is not the hardware itself, but what is done with it.

2. *AI Creation*: tracking and licensing the creation of powerful AI

With computing hardware accounted for, it is possible to track the creation of powerful AIs. Creating today's most generally capable AIs requires several thousand state-of-the-art chips

operating for over a month. Owners of a critical mass of chips can be required to report any computing jobs above a threshold and to receive a license before running a computation above a higher threshold (e.g. 10^27 bit operations, just beyond the current frontier, costing >$10M). Today such oversight would affect only a small number of chip owners and chip users. Receiving a license to create an AI would be contingent on the creation process meeting safety and security requirements.

It is necessary to license powerful AI creation and not just proliferation because, after creation, proliferation is as simple as copying a file. AIs can be stolen through cyberattacks by criminals or countries, or even leaked by insiders; Meta's most recent powerful AI, LLaMa, was leaked and is now proliferating, being fine-tuned for various purposes. AIs could eventually escape on their own, similar to a biological virus escaping a research lab or a computer worm spreading from computer to computer. OpenAI and others actively assess the possibility of an AI escaping and find that AIs can now perform some of the necessary steps. As such, one of the license requirements for creating a powerful AI should be adequate cybersecurity and "sandboxing" of powerful AIs until safety and security checks have been completed.

Notably, oversight of the creation of powerful AIs through oversight of large-scale computation will not cover the case where someone acquires a powerful AI and then fine tunes it for further purposes (e.g. fine-tuning a general purpose language model to create cyberweapons or fine-tuning a general purpose biology model for designing bioweapons). Fine tuning typically needs much less computing power, and so can be done by lone individuals with little computing hardware. This further illustrates the need to fully assess the safety and security implications of an AI before it is created and proliferated.

3. *AI Proliferation*: licensing the dissemination of powerful AI

Oversight at *AI Hardware* and *AI Creation* will address the majority of safety and security risks. However, AI is a rapidly advancing technology, with AIs being created the likes of which have never been attempted before. As such, *AI Creation* safety and security reviews will at times be imperfect, and the resulting AI must be evaluated to make sure it is actually safe and secure before *AI Proliferation.*

Additionally, a determined actor could try to evade oversight in *AI Hardware* and *AI Creation* by flying below the radar, training an AI in segments that are below the license thresholds on computation or computer hardware. This would bring logistical and monetary costs, but is conceivable, even though the only purpose to doing so would be to evade safety and security reviews. As such, there must be an oversight mechanism over all powerful AIs, regardless of how they were created, to review their safety and security and determine acceptable levels of deployment and dissemination that do not undermine national security.

AI Proliferation license review should only be required for AIs passing a simple, well-defined capability threshold. The review process itself can then be simple or complex depending on the AI. One simple tripwire for *Proliferation* license review would be the overall size of the AI (e.g.

the number of parameters). Today there are on the order of a few dozen AIs with over 100 billion parameters (several hundred gigabytes).

Publishing an AI's parameters would only require a license the first time, after which proliferation is functionally unstoppable. As discussed above, a proliferated AI can be modified by others, which could constitute a new AI requiring a license. Such a license requirement would be unenforceable practically, as lone individuals could make the modified AI with small amounts of computation and proliferate it anonymously; this is why license review is critical at the steps of *AI Creation* and initial proliferation. Still, a license requirement could be attempted for AI modifications that pass some bar of changes to the AI's capabilities, or, more enforceably, total computation used (which returns to oversight on *AI Creation*).

Keeping an AI's parameters private but giving users the ability to interact with the AI (as done by several companies today) should also require license review but likely bring simpler safety and security requirements.

Adaptability of Oversight

AI technology will continue to evolve, and so must AI oversight. As computer hardware or AI algorithms change the thresholds or tripwires for license reviews must change. Additionally the assessments of what constitutes an unsafe or unsecure AI will change as different types of AI are shown to be reliably safe or unsafe and as society deploys mitigations for specific threats that an AI might pose. As such, the implementation of government oversight must be adaptable.