NTIA Open Weights Response: Towards A Secure Open Society Powered By Personal AI

Authors:
Context Fund Policy Working Group (www.context.fund)

Original RFC:
https://www.regulations.gov/document/NTIA-2023-0009-0001

Authors

Context Fund is an volunteer online organization which is an experiment in public investment. Its AI products are used every day by scientists and policy thinkers (VerificationGPT, for example). Product designs are based on deep experience not only in AI, but also in psychology, security, investment and public policy. The Policy Working Group in particular advises on AI governance and investment, informed by product design and practical experimentation. It strives for maximal transparency, evidence-based reasoning, and working demonstrations of solutions, rather than only principles, in the spirit of the open-source movement. The group is made up of volunteer experts with backgrounds at top universities such as Stanford, civic institutions such as the Mercutas Center, and AI research and product companies such as Meta collaborating to push for the defensive acceleration of a secure and open AI economy which benefits all. We look forward to collaborating with NTIA in exploring the open weights question and more.

Executive Summary

Strong evidence suggests that open models are safer than closed models due to efficiencies in the fields of science, economics and cybersecurity. In science and cybersecurity, this is due to inspectability and the ability to share the model with millions of others to distribute the burden of verification, thus solving the expert problem. This substantially aids defender and builder users, the majority use case. In terms of economic equality, open models allow for extreme efficiency as well as more equitable distribution, since they can be offered at low or zero cost. They also prevent society from descending back into non-evidence-based thinking and warfare, inspire faith in transparent rule of law, and allow anyone to generate examples of their ideas at a small scale, which is important for clear communication. Closed models are most likely to be abused by deployers, while open models can be abused by either deployer and users, however, the advantages that open models provide for users acting in defender roles outweigh the risks of availability to attackers, roughly by a factor of 100:1, considering the

financial surface area that needs to be defended. Although less secure, we assess that it is acceptable for the government to allow closed APIs of foundation models to remain legal, as they can be used to satisfy commercial and technical considerations of deployment, for example, protection of trade secrets and engineering efficiency.

In our assessment, the government's support initially should be in administering standardization processes and RFCs (such as this one), legislating well-scoped mandates to add transparency to models and model outputs for high-scale deployments, funding defensive research, supporting responsible disclosure programs which would otherwise be underinvested by the private market, and participating in and administering open standards bodies.

Specific legal and technical designs are possible to add further design choices to the defensive acceleration of the AI sector and mitigate some of the remaining risks in the AI sector. To harden deployments against spam and phishing, we recommend immediately encouraging the use of physical security keys or biometrics tied to anonymous-but-accountable-by-karma user accounts, as well as scaling verification APIs and promoting the adversarial hardening of open models using offline data prior to deployment as a best practice. Adding watermarking will also improve traceability of outputs. To harden specific deployments against misuse and distribution, per-use model tainting of open model weights may be possible, however we do not recommend this as a default or legal requirement. To harden deployments against financial attacks, licenses like differentiable credit licenses can be experimented with. Together, traceability and licensing form a credible deterrent to abuse for most malicious users, while inspectability and shareability of open models forms a credible deterrent to abuse from malicious deployers and backdoored models.

Background

The Artificial Selection Pattern of Technology Development

Throughout human history, technologies have followed a consistent pattern of development. They often are first developed to achieve some positive goals for an end user. Immediate and obvious harms (especially those not achieving user goals) are identified and fixed quickly, while emergent problems from the deployment of a technology may take longer to be identified and solved. Overall, this forms a process of artificial selection, whereby a tool is specialized for use by humans.

The harms that persist in the long run (and hence may need more consistent and aggressive intervention from the state, possibly working against the forces of economics) tend to be those stemming from human-human alignment problems (often from zero-sum games) or bounded rationality ("out of sight, out of mind"). Climate change, plastics in the ocean, online spam and other pollution are prime examples. Zero-sum type problems are fixed by changing offense/defense efficiency ratios to favor the defender (thus reducing the payoffs from zero-sum games and contrastively favoring collaborative positive-sum games). The common advice to "drive defensively" is a good example. "Out of sight, out mind" type problems are often

identified by scientists and economists as data becomes available from deployments, however, they often lack the resources to design and test solutions at scale against them, requiring state investment. It is not enough to just identify problems, solutions must also be tested to ensure they do minimal harm themselves.

Most evidence suggests that although AI is more powerful than most tools we have used, it follows the path of artificial selection (tool development) rather than natural selection (self-reproducing populations left alone in the wild). This means that we can tune it with defensive acceleration policies (d/acc) - allowing development towards positive goals, while monitoring and identifying risks (by reducing uncertainty via collecting evidence from small-scale deployments) and disclosing them preferentially to defenders to allow time for patches and updates. This process operates with the supervision of the broad open-source and scientific communities operating under legal protection for scientific inquiry. Government involvement can help especially in scaling up defenses against subtle known externalities through setting standards, which makes public supervision more efficient.

Principles of Supervision

Many processes have been used to deal with uncertainty throughout human history by collecting feedback (supervision). We recap several that are relevant to the defensive acceleration of AI development, and describe their strengths and weaknesses.

The Scientific Method

The scientific method is the basis of collective rational thought. This requires the collection of initial evidence, the forming of a testable hypothesis based on that evidence, the testing of that hypothesis and the acceptance or rejection of the hypothesis in comparison to other possible hypotheses. Hypotheses with greater evidence are referred to as theories, and those with substantial evidence as laws. A recording of the experiment is formalized in a scientific paper, which also includes reproducible experimental artifacts such as code and datasets. The scientific method is generally practiced by individuals and is a fast way to improve a product (O(hours-days)[1]).

Peer Review

Once formalized in a paper, the artifacts of the experiment are presented to the scientific community and general public via a process of peer review (for example, on OpenReview). While the long-term outcome is the acceptance or rejection of a paper to a journal or conference, arguably the more important outcome is the refinement of the experiment over ~1-6 months while in review. While peer review is not perfect - it can be captured by narrow interests in specific journals, does not necessarily select for impact over precision, and can prioritize

[1] O(.) is big O notation and can be read as “order of” or “approximately”

narrow subfields which can make publishing more revolutionary ideas hard, the competitive supervision of the community produces reproducible, well-organized artifacts and a relatively high-quality stream of papers. Over time, through many papers, evidence accumulates for or against ideas in a friendly, but competitive, spirit (example: https://twitter.com/mengyer/status/1770461126293107134). Although the most common goal for work is academic novelty, papers which create standards and eval sets for problems are not uncommon (for example, BigBench, ImageNet, and in AI safety, Purple Llama). The process is generally fast (1-6 months for a paper), and accessible to all, since reproducible methods are prioritized. The Transformer paper is one of the most prominent examples of a peer-reviewed paper that catalyzed a wave of innovation with AI, later most prominently turned into a product by OpenAI via further experimentation, however, a continuous stream of lesser-known papers laid the foundation of the current AI systems that we use today. Development is generally less about discontinuous "breakthroughs" and more about continuous processes of experimentation, public feedback and improvement, although the language of breakthroughs is much more exciting for selling a paper.[2]

Open-Source Development

While the code produced by reproducible academic papers is an example of open-source development, open-source development extends far beyond journals. The overall "hacker" ethos of open-source prioritizes creating common libraries which are broadly useful, but not necessarily academically novel, as well as the meta-goal of creating and propagating a "gift culture" whereby reputation is gained by giving away valuable features, and personal freedom of action is prioritized. Feedback on open-source software comes through open filing of issues logging bugs, and often will see turnover on problems in the 1 hour - 3 month timeframe. Through consumer choice (both to use or donate time or money to useful projects), open-source software is one of the more responsive processes to handle uncertainty, and operates with extreme efficiency compared to closed-source silos due to asynchronous, permissionless code sharing ("with enough eyes, all bugs are shallow"). Due to the gift culture, pure open-source is often unsustainable as a business, relying on donations (such as Signal), however, a number of companies have built successful business models around charging for deployments of open-source software, so-called open-core (such as Docker), and many large companies contribute full-time employees to development of projects they use (for example, Google). Even non-contributing businesses and scientists also depend massively on the efficiencies of open-source software, which contributes hundreds of billions of dollars to the economy overall (https://www.zdnet.com/article/how-much-are-open-source-developers-really-worth-hundreds-of-billions-of-dollars-say-economists/). It is not an accident that the Internet and modern operating systems are the product of the open-source community.

---

[2] ML is still largely an empirical science, rather than theoretical (where "breakthroughs" may be a more appropriate descriptor), and in fact, one of the long-term lessons in ML research is that data and compute matter much more than algorithms (summarized here).

Market Feedback

Launching a product to the public also provides a form of supervision via consumer choice in competitive markets. However, unlike peer review, methods used to create the product are often tightly guarded, which can preserve significant upside for the shareholders (and somewhat for employees). Software companies with significant coder populations in management may also open-source older versions of products once a new one is released (Google did this, on a time lag of about 1 year), however this is not the norm outside of Silicon Valley. In competitive markets with several choices, low switching costs and low information asymmetry, better products are selected by consumers, passing feedback to the companies through revenue and attention (social media, word of mouth), however, this can be hacked so it no longer represents an effective source of feedback (via dark patterns which prevent disaffected consumers from leaving). In more open settings, consumers select for products which fulfill their “jobs-to-be-done”, and will quickly abandon products that are uncontrollable or have noticeable bad side effects (for example, committing crimes for which the users are liable). However, in settings with fewer consumer options and less information, consumer choice is not necessarily a reliable source of feedback. One way that governments and industry groups have fixed this problem of lack of reliable feedback (expressed as externalities) has been through the establishment of standards which reduce information asymmetry, as well as antitrust actions, which add consumer choice to the market. Many of the more pathological problems of the current Internet (seen from 2010 onwards) may have a connection to a shift in actors from the open-source ethos of the early Internet to the for-profit ethos of the 2010s and onwards.

Standards

When there are enough examples of a product category to define the desired and undesired behavior of the processes (inputs -> outputs) that it represents, then standards can be established. This is especially useful for end consumers, as most people are not experts about most things in the world, especially when evaluating business products where the methods are often hidden. Complex truths are a lemon market, and without standards which verify end products, consumers often are left with only trial and error to compare, and even this can be hacked (astroturfing reviews, for example). Standards which define communication processes also allow interoperability of products as well, which helps to reduce the risk of single-point failure in the economy. While standards are inherently slower than market feedback (months-years) due to the need for broad industry-wide consensus and allowing time to collect enough examples to clearly describe behaviors (often via an eval set), they often are a key step to allow the market to apply broad feedback.

Standards are curated by industry groups such as IETF, often in cooperation with government groups, like NIST, and frequently formatted as RFCs (requests for comment). A key outcome of standards is that they are easy to understand, easy to test for, and broadly accepted (often a result of including broad industry participation in the process). This process worked wonderfully for the development of the Internet (you’re likely reading this document due to the TCP RFC: https://www.ietf.org/rfc/rfc793.txt), and counterexamples also abound (for example,

Google moving away from the XMPP chat standard for Gchat contributed substantially to the fragmentation of the chat ecosystem that we see today, as other companies followed suit. This led to more walled gardens).

For AI, scientific standards/benchmarks are often expressed in eval sets, developed by groups such as ML Commons AI Safety Working Group and NIST. While companies are mildly hesitant to rank their models against others on a leaderboard (due to the knowledge that consumers might select against them), in practice, if models are open, this can be done by the community adversarially, or by connecting open publication and ranking to receiving government funding (for example, NIH grants typically require open models and publications as a condition of receiving government funding). There is a large amount of standardization work to be done in AI, and it is often not well-rewarded in peer review systems (which prioritize academic novelty).

Open Letters

Communities also normalize behaviors via open letters (https://www.reddit.com/r/contextfund/comments/1ba4gf0/responsible_ai_x_biodesign_open_letter/, https://www.reddit.com/r/contextfund/comments/1b7d1ln/a_safe_harbor_for_independent_ai_evaluation_open/). This can be relatively fast (months), and if the field is responsible (scientists generally are a responsible population due to selection for curiosity and rigorous thought), broadly useful. Open letters receive rapid feedback (hours), and can become de facto standards of behavior through public exposure leading to broad signatures (the positive side of social media).

Responsible Disclosure

Responsible companies embedded in open communities often follow a security practice known as responsible disclosure. Under responsible disclosure, those which discover security vulnerabilities notify affected parties several months-years before disclosing vulnerabilities to the public, allowing affected parties time to patch systems (one recent example of this practice: https://www.wired.com/story/saflok-hotel-lock-unsaflok-hack-technique/). Vulnerabilities in mature systems are often subtle, but with enough eyes, all bugs are shallow. The security ecosystem depends on 1.) the use of public red teams who work full-time in adversarial roles 2.) software/weights being available to red-teamers under legal protections, 3.) an industry-wide standard process for coordinated disclosures (https://www.cisa.gov/coordinated-vulnerability-disclosure-process, https://en.wikipedia.org/wiki/Coordinated_vulnerability_disclosure) and 4.) a low-cost update process to distribute patches (failure modes around responsible disclosure mostly center on expensive update processes, which are most common in offline hardware settings). While updates are slightly easier in “walled gardens” like Apple’s iPhone ecosystem, updates in open-source ecosystems like Android are also effective (due to human bounded rationality and instinct to cluster around working solutions, a handful of prominent distribution channels will emerge), and in practice the market creates incentives which lead to companies to help

automate the detection and patching process like Socket Security (also, walled gardens are the types of ecosystems that take monopoly rents which create economic vulnerabilities in the system).

Responsible disclosure of risks should also be used in AI, allowing defenders time to patch systems before new offensive capabilities are introduced, thereby increasing information asymmetry in favor of the defender and disincentivizing attacks. For example, GPT-4 likely broke visual CAPTCHAs (requiring a switch to verifying accounts with physical security keys or passive CAPTCHAs), and the ability to synthesize 1-minute video clips using models like SORA may create an undesirable arms race in political advertisement (requiring labeling and possibly banning AI-generated political ads and adopting widespread use of verification APIs for general content). While responsible disclosure has been a widely adopted practice in computer security (https://www.helpnetsecurity.com/2022/02/17/vulnerabilities-disclosed-2021/), this has been less widely adopted in AI (to date, OpenAI has yet to officially disclose the CAPTCHA vulnerability, although it has been noted by independent researchers; to their credit, OpenAI is appropriately delaying the SORA release). Notably, unlike classical computer security, in AI security problems, communicating the vulnerability broadly and publicly can often be done without communicating the means of exploiting the vulnerability, allowing public disclosure of future AI vulnerabilities without needing to rely on private emails or coordinated disclosure programs, as in traditional responsible disclosure.

As a rough heuristic, offense/defense capability ratio, traceability and deployment size are three key intermediate metrics to consider when designing a responsible disclosure system.

Legislation & Prosecution

Finally, legislative systems create laws which penalize offensive behaviors which have known, severe harms, thus shifting the game payouts to favor legal behavior and cooperative games. Laws allow for feedback by state prosecution, as well as civil suits, and often require years to change, as well as years to collect a sufficient volume of case law to be able to be interpreted unambiguously within a jurisdiction. This is multiplied by the number of jurisdictions. Due to the necessity of relying on lawyers as intermediaries and interpreters for almost every interaction with the legal system, transactions within a legal system can cost millions of dollars and require years, and can favor parties with unequal access to finances, personal connections or time, so-called “lawfare.” Yet, enforcement via civil suits is one of the few systems that can provide feedback of sufficient scale and surety to deter behaviors that shake the foundations of democracy and trade, such as breaches of contract and gross misrepresentations. Enforcement via criminal prosecution in turn deters actions which shake the foundations of rule of law itself. The legal system is a powerful technology to create fear that needs to be used with care, and due to its glacial pace, is not intended to match the speed or generality of contemporary tech developments, much less to guide future general-purpose research.

Principles of Security (re: 1)

All technologies are deployed in the context of games played between humans. In analyzing these games, we divide roles between “attacker” and “defender.” Roles are further divided between “user” and “deployer” of a technology. Both user attackers and deployer attackers represent important scenarios to model.

A critical ratio to estimate in any security game is offense/defense capability ratio. High offense/defense ratios favor first strike and competition, while low ratios favor second strike and cooperation. Modern civilization broadly depends on this ratio favoring defenders, thus inducing cooperation (“rule of law”, “defensive driving”), although competition and first strike is important to allow in limited contexts (we allow this in business via time-decayed property rights based on creation of technologies or IP).

While an offense/defense ratio must be estimated empirically based on analysis of each game, some general principles can be derived. Traceability of information generally favors defenders (since this enables rational decision-making and authentication and authorization in particular). In settings where users can be either attackers or defenders, balance of power (both in distribution of information and capabilities) also is important, as it allows for a larger number of defenders to swiftly overpower an attacker in a second strike. Since most users are playing defensive games (~99% of GDP is concentrated in legal enterprise, for example, https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/economic-impact-cybercrime.pdf, https://www.hellenicshippingnews.com/gdp-doesnt-include-proceeds-of-crime-should-it/), there are often far more defenders than attackers, however, this is affected significantly by bounded rationality, practically expressed as education level, long-term goals or occupation (the foundation of democracy is an educated populace).

Games are distributed in many places throughout the economy. They can take place in the control planes of attention, cyber, financial and physical layers (ranked roughly in decreasing order by speed of change). Defense is often necessary in-depth at every layer, since all planes can modify the choices of an adversary and hence compel suboptimal decisions. There are likely trillions of games ongoing in the world, a handful that most of the world play, along with a long tail of games that are unique to each person and local community.

Games are infinitely iterated and never fully won or lost. However, by modifying the offense/defense ratio (via introducing friction in information or technology processes) (often via timing changes to the visibility of information)[3], they can be adjusted by those on a higher control plane to favor cooperative play. Modern SSL protocols which underpin Internet communications, for example, are designed by security researchers to have information asymmetry that makes a channel exponentially harder to attack than to defend.

Technologies and their distribution patterns (also mediated by technology) affect advantage ratios significantly, especially high-variance technologies like AI.

The Personal AI Entity Model

We anticipate that the major deployment path of AI in the future will be towards *personal AI*, which has a personal state added by fine-tuning or altering foundation model behavior

[3] We refer to changes to information timing as “information asymmetry” in general (either infinite delay (0/1) or finite delay).

behavior with user-generated state. In all cases, both attackers and defenders will have some type of capability in AI.

The Benefits of Openness (re: 2)

Evidence from science, economics, and cybersecurity defense argue for open-core foundation models being the safest default deployment configuration, although closed deployment configurations are not unreasonable to achieve specific design constraints. Hardening and traceability of weights and behaviors on networks are important tweaks to allow intermediate design choices between these two paradigms, although as of 2024, they need testing before being developing into standards (https://docs.google.com/document/d/1JkTIbLFYLhg3EzQDm3zuC1H0dNdx6RowlndUXq2QwgY/edit#heading=h.d3lmnf9jgc64).

Scientific Progress & Education

The advantages for science of reproducible and inspectable models enabled by open weights are so obvious that they almost do not need to be argued. The foundation of science is the scientific method and peer review, which relies on reproducible artifacts and evidence, especially in the case of shifting paradigms, which tend to be unpopular at first and demand rigorous demonstrations before believed. Openness also has a strong effect size on scientific progress. Without open models, science is limited to what a single local group can produce, a 10,000-fold reduction in capacity at least for most fields, which means that in effect, science starves. And when science starves, a civilization eventually starves (or is outcompeted by rival civilizations). While it may be tempting to suggest “halfway” interventions like “methods and data are ok to share, but weights are not” (thereby imposing a compute cost on replications), restricting sharing of weights would easily slow down science 1,000-10,000x fold (hypothetically going from 3,000 independent validations per year to 3 validations/year). In addition to slowness of collaboration, this also would concentrate power into a handful of silos of compute and data, with ancillary political jockeying and a much smaller fringe to propose alternative paradigms, which are what lead to scientific revolutions (https://en.wikipedia.org/wiki/The_Structure_of_Scientific_Revolutions). Since education also uses the scientific method (at a personal level), and is the foundation of democracy, restrictions on information transfer also affect the foundations of civilization.

Economic Equality

The advantages of open weights to economic equality are also strong. A high-confidence and scaled physical risk from AI is job turnover and market concentration leading to extreme inequality, exacerbating the existing trends of globalization (https://www.theguardian.com/technology/2024/jan/15/ai-jobs-inequality-imf-kristalina-georgieva

). Open weights are likely to be a significant hedge against resulting political unrest by distributing opportunity, similar to the way the open-source development of the LAMP stack facilitated the growth of millions of personal websites and early e-commerce in the early 90s. In addition, learning about and experimenting with technology is a much more prosocial “game” than crime, which, realistically, would be a competing game in cases of mass unemployment.

Attention Equality

Psychologically, humans have strongly negative reactions to fear and uncertainty. Without transparency and openness, the owners of closed models may come to be perceived as god-like (this sounds hyperbolic at the moment, but is not impossible if AI techniques are closed off and the rate of progress continues), and society may descend back into a primal mode dominated by warfare and superstition. In contrast, with open-weight models that can be transparently understood and replicated by anyone, “secret” techniques lose their allure of “magic,” leading to greater attention equality (and hence robustness of the human value network). We have some trends towards this as software shifted from publicly funded open-source to privately funded closed-source in the 2010s, which were also correlated with a significant rise in fraud preying on attention inequality and information asymmetries. While attention inequality may sound highly attractive to its beneficiaries, applied at society-wide scale, it is highly destabilizing, since the imbalance of power selects for competitive first-strike games which can spiral out of control (the dose makes the poison).

Security

The benefits of openness for security are not immediately obvious (given how prominent attacks using AI are), but become more apparent on rigorous inspection when considering the impact of AI on both attacking and defending roles. Most attacks defeated by AI as a defender are not visible - they just fail silently.

User Defender Against Deployer Attacker - Independently Verifying A Model

Consider the scenario of a user attempting to validate whether a closed model, such as GPT-4, is safe to trust. Since a closed model is not inspectable, it is not possible to fully understand the capabilities of a deployer attacker, only through the revealed behavior of short fragments of input/output interactions which may be purposefully tuned by the model to disguise latent behaviors and long-term goals.

In contrast, the weights of an open model can be analyzed by many tools to understand the full spectrum of the model’s latent behavior. Open-weight models can also be run much faster in simulators than closed weight models (due to not hitting proprietary rate limits), allowing for security research over larger datasets, fuzzing, etc. They can also be run privately. In addition, the responsibility of inspecting specific model behaviors can be distributed to other experts in the community, allowing a million-fold increase in efficiency (“any bug is shallow given enough eyes”). In addition to finding bugs, making users aware of security risks and patching

open-weight models is substantially easier than closed models, since this can be done adversarially, without needing the attention of the closed-model owner. Patches in practice compete with other priorities, especially in large companies, however, allowing affected users the ability to patch their own models distributes the responsibility. One researcher can identify the bug, one can identify and test a solution, a third can design a patch and a fourth can work on a press release, none of which have to work for the company which originally created the open model. While this depends on having a common channel for responsible disclosure (one place for government intervention), this is extremely efficient.

This is also true from the perspective of a designer of a communications protocol for AIs. Since the capabilities of an open-weight model are better known, concrete theoretical assertions can be made, allowing for the design of more efficient defenses tuned to the behavior of the hypothetical adversary. In contrast, if capabilities are not known, defenses may be very expensive (to account for all possibilities), which in practice can lead to companies simply neglecting defense as “too expensive” and reverting to closed communication protocols, leading to information inequality. Open weights lead to open protocols which lead to open communications.

The most severe risk is a wide perception-reality gap if information access is restricted by law and coupled with fear, as would happen in the case of banning open weights. In contrast, if information is widely available, individuals are able to independently verify assertions and defend against attacks. Scientific and market supervision is highly capable of designing context-specific cooperative games if data is available. In an economy where the civil GDP is 100x larger than the dark web GDP, openness “just works.”

In the extreme case, with a wide perception-reality gap in society caused by dark models operated for years without oversight, scaled physical threats like climate change can go unsolved even though society theoretically had the capacity to fix them.

Deployer Defender Against User Attacker - Security Equality

Spam & Phishing

In the case of a deployer defender against a user attacker, open-weight models will likely increase the number of capable attacks and decrease the information asymmetry available to the deployer defenders. We have seen evidence of increased number of AI-generated attacks in the recent rise of spam emails and fraudulent websites in the last year (https://substackcdn.com/image/fetch/f_auto,q_auto:good,fl_progressive:steep/https%3A%2F%2Fsubstack-post-media.s3.amazonaws.com%2Fpublic%2Fimages%2Fac67b764-f0b1-4ba6-8fe6-dd594e8f9ace_960x540.png). Similarly, voice cloning has been used in adversarial attacks in the last year, however, notably, closed API services were used (https://www.politico.com/news/2024/01/29/biden-robocall-ai-trust-deficit-00138449), and this turned out to be a red-team attack to draw awareness to the issue.

Counterbalancing this, however, is that open foundational models allow for much wider and more efficient distribution of AI-powered defenses (security equality), which are independently verifiable (as established in the prior section). Given that there is 100x greater

economic value that needs to be defended than is used for attack, defense is a universal need that needs to be exceptionally efficient, and open models provide this efficiency, both in initial distribution and in terms of improvements/patches. If an open model community is well-funded, and has efficient update channels, this is a strong working model (if not well-funded, then this works less well).

AI defenses will likely get very strong (https://www.wired.com/story/fast-forward-nsa-warns-us-adversaries-private-data-ai-edge/). Given that cyberattacks *must* deviate from the normal communication patterns at some point to deliver a payload, stronger predictive models have a chance to force spam to carry so small of a payload that it cannot accomplish its task. At some point, the payload cannot be compressed any more. While we generally believe spam of *some type* will always be an issue, some common channels of spam may be able to be eliminated altogether.

AI defenses, however, must be deployed quickly and at scale. This is urgently lacking for much of the web, and we need good open models to do so.

Adversarial Attacks On AI Defenders

Similarly, knowledge of the inner workings of an LLM may theoretically improve the ability to conduct more efficient attacks (via adversarial examples, for example, which are quite common in papers), these have not been commonly observed in the wild, likely due to their sophistication to pull off, the low rate of AI deployment and the higher ROI of less sophisticated attack methods on more weakly protected targets (at least for financially motivated actors). Although the criminal space is to fast to adopt AI techniques (https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/economic-impact-cybercrime.pdf), and this will be important to monitor, the threat will scale at the same speed as the AI deployment surface and is not the most urgent threat. It is also reasonably well addressed by hardening and responsible disclosure of general-purpose exploits.

Biosafety

A much repeated, but relatively unlikely risk, might come from the design of a biological virus that cannot be defended against once synthesized (self-replicating, self-distributing, scaled physical malware that can adapt faster than our immune system and is able to delay its lethality until it has spread to a very wide population). This is a significant risk since all the traceability and deterrence considerations do not apply to irrational, self-replicating particles without an executive function.

However, open weights of LLMs only make it slightly easier to learn how to make bioweapons (OpenAI paper), while making it 10x easier to defend against synthesis of said biowepons (via cheap distributed AI defenses to screen for hazardous molecules). Indeed, we believe the open letter and norms approach to bioweapons defense already taken by the scientific community, along with selectively removing hazardous bioweapons training data from foundation models by default and pursuing international treaties with regards to bioweapons in war (similar to that done for nuclear weapons) will be much more fruitful paths to take than banning open weights, which seem tangentially connected to the problem at best.

Hardening Deployments of Open Foundation Models (re: 3)

While open-weight models have security advantages over closed-weight models, they do have some weaknesses, which can be patched by hardening both the weights and the deployment contexts in which weights are used:

Hardening

A simple approach to defend against adversarial attacks against introspectable open-weight models is fine-tuning against private data kept offline. The dataset can be drawn from positive examples of attacks found by a well-funded red-team community (adversarial hardening), or merely human-curated moderation data from the deployment team. If models are sufficiently easy to update (via coordinated disclosure channels, for example), distributing hardened updates should be relatively straightforward as well.

Traceability

Deterring spam and phishing attacks will be important in the 2024-2025, as the network updates to new AI capabilities for both attackers and defenders.

Physical Security Keys

Upgrading accounts on important subnets to use physical security keys (Yubikey, FaceID) can both reduce the annoyance of visual CAPTCHAs and stop phishing attacks (physical security keys provide cryptographically secure demonstrations of authentication and humanness). A touch or biometric verification makes it significantly easier to attribute behaviors to an account, ties the account to a physical object which cannot easily change locations, while imposing minimal cost on a user and preserving privacy (since the account is not necessarily a real-name account). In the future, it may be the case that everyone has a couple security keys, associates them with a small number of pseudoidentities (<100), and has each pseudoidentity be accountable by karma to a combination of a personal AI under their control in combination with other user's AI they interact with. Notably, security keys defend with perfect F1 against login attacks and add high traceability of both AI and human-generated phishing attacks.

Verification Models

Verification APIs can already defend against both human and AI-generated spam attacks with higher speed and accuracy than many human networks (see www.verificationgpt.org). Scaling up the number of subnets that are verified is easy and will be a key priority this year.

Remote Passport Verification

Building a simple email API for remote passport verification from public signatures embedded in e-passports is immediately necessary to prevent high-value fraud (title fraud, for example) using synthetic ids. This would be useful worldwide and arguably is uniquely the responsibility of the government, which owns the e-passport system.

Watermarking

For many creative generative AI applications, adding invisible watermarking in outputs is possible (https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/) and would quite helpful (to assist defender AIs against spam). This can be verified via an API provided by the creator. This is easy and arguably a pressing need, while not incurring a large privacy cost (bit would not be user identifying but would identify it as AI-generated or not).

Weight tainting

For models where strict traceability of weights to a user is required, weight tainting is a theoretical possibility, although untested to our knowledge. In weight tainting, authenticated users download per-user altered weights which have been perturbed with a statistical transformation of unique information associated with the user account (like a pseudorandom number), while the base weights and the specific information tag that was associated are kept offline in secure storage. If those weights are found again in the wild, they can be compared by the deployer with the offline information and attributed to a specific user, enabling legal action. Similarly, identifying which patterns comprise the “serial number” (to remove) would require difficult-to-coordinate collaboration of users, especially for large models. While we do not recommend this as a default or legally required procedure, and there are various technical considerations to investigate, it may be appropriate for some types of highly sensitive models, and can be combined with per-user watermarking to trace outputs as well.

Early Access Tokens

Early Access Tokens are a hypothetical approach that can help broaden the scope of a gift economy in AI development. These work by tracking contributions to open-source software and science, and enabling early access to AI source code and weights for accounts with high karma, using a token tied to a physical security key. Since the overlap of malicious and well-intentioned users is relatively small, this potentially can be useful to allow asynchronous, unplanned sharing of code and weights with responsible collaborators across corporate boundaries, while adding friction to adversarial access (karam is required to be accumulated). A similar, more manual process, is sharing user-tainted weights preferentially with known security researchers, however, this may miss many of the benefits of unplanned, asynchronous development.

Differentiable Credit Licenses

Similarly, differentiable credit licenses can be used in combination with the karma systems to help defend against free-riding financial attacks in model development. In

combination with weight tainting (which adds detection capabilities), this could potentially add deterrence needed to enforce payment, but still allow profitable collaboration well beyond the narrow boundaries of a single company, an capability exceedingly needed if we are to preserve a healthy larger economy.

Towards Public-Private Partnership (re: 4)

In our assessment, the government's support initially should be in administering standardization processes and RFCs (such as this one), legislating well-scoped mandates to add transparency to models and model outputs for high-scale deployments, funding defensive research, supporting responsible disclosure programs which would otherwise be underinvested by the private market, and participating in and administering open standards bodies, as well as negotiating relevant bioweapons treaties with foreign powers to prevent shared risks. Finally supporting testing and scaling approaches outlined in (3) could position the United States to handle the AI transformation well.