Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights:

Response to NTIA Request for Public Input (NTIA–2023–0009)

Transformative Futures Institute

March 27, 2024

Stephanie Weiner, Chief Counsel, National Telecommunications and Information Administration | Travis Hall, Acting Associate Administrator at US Department of Commerce, National Telecommunications and Information Administration

Subject: Openness in Artificial Intelligence Models Request for Comment (NTIA–2023–0009)

Dear Mr. Hall and Ms. Weiner,

We greatly appreciate the opportunity to provide comments regarding the benefits and risks of Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights. This dialogue between the public and private sectors is crucial to enable a comprehensive understanding of the potential risks associated with these technologies and how best to manage them.

The Transformative Futures Institute (TFI) is a research group concerned with understanding and mitigating the risks from emerging technology, specifically artificial intelligence (AI). TFI has previously submitted responses to NIST regarding the development of the AISIC and the request for information regarding President Biden's Executive Order on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" on May 2, 2024 (Gruetzemacher et al., 2024).

Below are some of TFI's high-level points and recommendations on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights:

1) We believe that AI systems have the potential to bring immense benefits to society and that foundation models without a general-purpose design, such as those for a wide variety of scientific applications, should continue to be open sourced. As rightly noted by NTIA in the announcement, the primary concern is with the more general dual use systems at the edge of the frontier of research, a narrow set of the most computationally intensive, large-scale AI systems.
2) Open foundation models with widely available model weights can have many important benefits, such as enhanced transparency and equal distribution of capabilities to the less privileged, but they can also have many risks. These include easier and broader access for malicious actors and misuse, system failures (without adequate testing and evaluation), and little opportunity for continual system evaluations to identify harmful properties, provide critical safety updates, or recall misaligned systems.
3) The release of open foundation models with widely available model weights also rapidly diffuses such capabilities to adversary nations, state competitors, and non-state actors in equal measure. Thus, while open weight models can help democratize access, this includes hostile intelligence services, global militaries, organized crime groups, and cyber threat actors. This rapid diffusion has the potential to shift the balance of power.
4) As systems increase in capability and generality, it will be increasingly important to test and evaluate systems for safety and stability prior to public release. Due to emergent, unanticipated properties and capability shifts, understanding this range of capabilities will be crucial to preventing dangerous accidents, the acceleration of misuse (e.g., favoring cyber offense), and for the U.S. to maintain technological dominance.
5) While the risks from current systems remain limited, increased capability and generality will only broaden the scope of misuse (both quality of threat and actor) and the potential for unexpected failures and emergent and autonomous behavior. As autonomous agent systems become more capable and widely available, the extent of impact from such capabilities (e.g., offensive cyber operations with self-replicating, adaptive malware) or our capacity for control is unclear.
6) To maximize the benefit of foundation models, developers can offer structured access to models (e.g., through APIs) to a wide range of users while maintaining the capability to prevent harmful misuse, block or filter dangerous content, and conduct continual safety evaluations. This way, developers retain the capability to provide safety updates and required alignment mechanisms if dangerous capabilities surface, including potential shutdown or rollback if necessary. This puts control in the hands of governments and technology developers, not strategic competitors or non-state actors. Organizations that wish to pursue open release should use a staged approach, allowing sufficient time for technical evaluations and risk management processes to unfold.
7) To effectively test and evaluate these systems for safety and reliability prior to deployment (and continually with staged release), what is needed is a robust and credible third-party testing and evaluation regime with the proper infrastructure, talent, and organizational confidence. This is needed at the international level, with a shared understanding across government, industry, and academia; there must be trusted agreed-upon standards across different schools of thought, as well as a willingness for experimentation as systems adapt and increase in capability. This adaptive ecosystem could greatly accelerate our ability to understand foundation models before deployment to implement more reliable safety practices and guardrails on frontier systems.

In the following sections (starting page 4), we provide comments on question 2 and question 5 (including subquestions). Sections with no response are not indicative of the question's lack of importance but rather our lack of expertise in these areas. Thank you again for the opportunity to comment on the NTIA's Openness in AI Request for Comment.

If you need additional information or would like to discuss further, please contact Kyle A. Kilian at kyle@transformative.org or Ross Gruetzemacher at ross@transformative.org.

Best regards,

Kyle A. Kilian | Deputy Director
kyle@transformative.org; kyle.a.kilian@gmail.com
Phone: + 1-720-563-9267

Ross Gruetzemacher | Executive Director
ross@transformative.org; rossgritz@gmail.com

TFI's Comments on Specific Items (Questions 2 & 5) in the *NTIA Openness in AI Models Request for Comment*

Note: This comment attends to only a subset of questions (2 & 5) due to our expertise in these areas and not others. We number the questions (bold) and subquestions (italics) as defined in the Request for Public Input.

Q2. How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?

a. What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?

A key concern with releasing dual use foundation models[1] with widely available model weights (referred to here as open foundation models or open weight models) is the inability to conduct thorough evaluations of unpredictable systems or repair and recall models found dangerous. This is an important public safety concern as the community remains unclear on the degree and character of risk as systems increase in capability and generality. With open weight models, there is minimal opportunity for critical testing and evaluations (T&E) throughout the model lifecycle, with no chance of continual evaluations, which may be crucial with the potential for emergent capabilities. These emergent properties can arise organically from increases in scale, and the only way to identify these capacities is through either careful T&E or trial and error in deployment (which could include malign actors). The science of evaluations remains nascent, with fragmented schools of thought on how to appropriately conduct T&E and no standardized process (Gruetzemacher et al., 2024); however, the process is crucial to identify potential risks, range of capabilities, and their potential impacts on safety and security prior to widespread deployment.

Another critical point to take into consideration is the proliferation of previously inaccessible capabilities to U.S. adversaries or non-state actors. Open foundation models open a range of capabilities to the universe of actors, potentially disadvantaging the U.S. for technological and military dominance.

[1] For this comment, we'll be restricting our comments specifically to frontier AI foundation models—large-scale systems requiring substantial computational resources—that are designed to complete any possible task (generally termed foundation model or frontier model). There are two aspects to the term open foundation model. First, is ***dual use***, which refers to systems with a high degree of capability and generality—e.g., it can be used in a variety of domains—developed at the edge of frontier research, and second, is the wide availability of model weights. This comment will refer specifically to this class of system with widely available model weights, which we'll reference as ***open weight models*** or ***open foundation models***. This distinction could have important legal implications down the line with increasingly capable and general systems.

While AI companies work to integrate safety measures prior to deployment, recent research has shown that fine-tuning foundation models (and RLHF data) can easily reverse safety measures and guardrails, even with benign intent and using trusted data sources (Kapoor et al. 2024). This is true with both open and closed models. However, open weight models make this process far easier for actors intent on using state-of-the-art systems for malicious purposes and without attribution. This is evident from the speed at which criminal actors deployed the first open models onto the dark web (e.g., XXXGPT, WormGPT, FraudGPT, etc.), training, and finetuning variations on dark web content (Kaspersky, 2024). Additionally, open weight models can facilitate the exploitation of weaknesses (damaging if used in safety-critical applications), amplification of bias, and potentially the easier development of cyber exploits with no capacity for attribution (a considerable problem in cyber). Ultimately, widely available open weight models can hinder our capacity to control frontier systems—through regulatory means or unanticipated emergent capabilities.

b. Could open foundation models reduce equity in rights and safety-impacting AI systems (e.g. healthcare, education, criminal justice, housing, online platforms, etc.)

A good portion of AI safety work is applied to decreasing inequality and preventing algorithmic bias in policing, housing, and healthcare, and companies spend a good chunk of their efforts on this end. While these issues will increase in prevalence as generative AI saturates economies, they are not the key concerns with open weight models. A reduction in equity in rights could be possible or exacerbated in certain cases (e.g., authoritarian regimes that otherwise wouldn't have access), but it is more likely that open foundation models would increase equity while democratizing access and altering power asymmetries. The places where such models would likely reinforce inequality, especially where there are economic incentives and race-to-the-bottom dynamics, seem unlikely to be accelerated by open weight models.

However, there is no control over the distribution of open weight models, which would invariably include bad actors and authoritarian governments, bringing new mechanisms of control for illiberal regimes. While open foundation models would likely increase equality in Western states, they could do precisely the opposite in less privacy-conscious regimes. In fact, given the restriction in U.S. law that prohibits U.S. intelligence and law enforcement agencies from accessing U.S. citizen data, there is a clear advantage for adversary regimes that can go all in on using personal data for model development (clearly impacting privacy) to compete with Western states or for domestic surveillance. While this could be true for open or closed models, for states that are less sophisticated technologically, open weight models would hand this capability to them.

On the other hand, wider access could empower citizens to resist or evade such efforts, but the balance is not in their favor. Beyond governments, this widespread access could create new variations of inequality we have yet to consider, providing disparate actors across industries new means to exploit workers and disadvantaged persons without the technical bandwidth. Cyber threat groups could create tailored platforms for exploiting underage persons on the dark web, spreading disinformation, or right-wing groups targeting migrant communities, to name three possible examples. At the same time, open weight models level the playing field wide enough that the proliferation of deep fake media and disinformation could saturate the information ecosystem, accelerating truth decay and decline in trust in information (Kavanagh, 2018).

c. *What, if any, risks related to privacy could result from the wide availability of model weights?*

It is plausible that models trained on non-public data could eventually be used intentionally or unintentionally for access to this information or for illicit purposes, impacting privacy and public trust. If an AI company (or firm using open models) is willing to release model weights based on private data they are not willing to release, the long-term implications on privacy and access to this information are unclear. Thus, releasing open weight models trained on non-public data is taking a gamble on whether or not future work will allow others to extract the data or use it in novel ways for criminal enterprises. The presumption should be that such releases could violate the privacy of any persons whose data was used. The opportunities for misuse are likely far wider than what we currently understand. This is why we must carefully weigh the costs and benefits of open weight models, considering the potential trajectory of the technology.

d. *Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy?*

The release of open foundation models with widely available model weights can level the playing field substantively for state and non-state actors intent on causing harm with few mechanisms for stemming misuse or levying attribution (including functions such as OFAC sanctions or hardening infrastructure from known actors). This has the potential to exacerbate security risks to public safety, democracy, defense, and the economy. It is already known the current threats we face from state and non-state actors have the potential to be enabled by AI, and this could be accelerated by open weight models; examples of these changes are already visible on a limited scale—such as automating cyber exploits (Arntz, 2022), phishing attacks (Desai et al., 2024), and disinformation—radical changes in tactics remain limited. However, the key concerns for open foundation models are the scale and variety of actors that can participate and the rapid pace of change. As models become more sophisticated, new attack vectors will proliferate as well (such as automated and adaptive agent systems).

While the vision of AI democratization is understandable and open foundation models do level the playing field for individuals and small businesses, it does the same for low and high-resourced bad actors alike, including state defense and intelligence services; indeed, the same open weight model that is released on GitHub to the AI community, with the same level of capabilities, is also released to adversarial state and non-state actors. While it is important to remain optimistic, the U.S. Government must take into account worst care contingencies and plan on how to minimize them.

i. How do these risks compare to those associated with closed models?

Crucially, closed models can be subjected to ongoing testing and evaluation (T&E) processes to ensure continued safe operation, and harmful capabilities can be patched or updated. Once the open foundation model is *in the wild*, any potential emergent capability or misalignment is accessible and at the whim of the user. This could include relatively benign bugs or more significant misalignments that could lead to cascading failures. Preventing such failures is inherent in NIST's AI Risk Management Framework (Zilong et al., 2024) and President Biden's "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" with voluntary agreements by leading AI companies to conduct evaluations and red teaming before widespread deployment (Executive Order 14110). As noted in sub-question *a*, without T&E, emergent capabilities, and risky attributes could be unintentionally released, impacting good and bad actors alike and all connected systems. Closed or restricted weight models can allow iterative evaluations and updates to ensure such emergent side effects are not released unintentionally.

Comparing the risks of open weight foundation models to more restricted *open models* centers on the level of accessibility, ease of implementing potentially harmful applications, unintentional release of misaligned or unsafe systems, and inability to reverse these decisions. With respect to misuse, wider accessibility to open foundation models can aid bad actors in the discovery of vulnerabilities in infrastructure (e.g., vulnerabilities in power grids, transportation networks, or communication networks) to a wider range of actors. Indeed, a key concern is the expanded range of attackers capable of exploiting vulnerabilities in connected systems.

At the same time, closed models that allow API access seem likely to have similar vulnerabilities as open models where safety features can be reversed through finetuning. Additionally, theft of model weights and sensitive IP by cyber criminals or through espionage, especially from well-resourced actors, is also likely. However, the widespread proliferation of open weight models will shift the number of potential actors and thus the potential for AI-enabled attacks and disinformation at scale. Leaders and regulators must recognize that open release can enable misuse and state integration of powerful capabilities far more easily, which could disrupt the offense-defense balance between states (less capable and powerful) and potentially shift the balance toward cyber offense.

ii. How do these risks compare to those associated with other types of software systems and information resources?

The primary risks from open foundation models are a novel product of the specific paradigm, model generalizability, capacity for transfer learning (and fine-tuning), and unique characteristics that increase with scale (data, compute, or model parameters). Such risks are not present, at least in the same way, with any other type of software—open source or otherwise—in other domains. There are comparable examples that can be drawn in areas of misuse, such as illicit weapons development or biological weapons design, but risks from LLMs are ultimately unique, requiring specialized attention and consideration by developers and lawmakers. In fact, this generalizability acts as a force multiplier for other categories of risks seen in other technological systems (e.g., synthetic biology design) and will increasingly evolve over time.

Another key difference, as discussed, is the capacity for reliable T&E prior to deployment. Other high-security software systems have been tested in ways that are not currently reliable or feasible for LLMs (e.g., extensive specifications prior to development, architecture maps, unit tests, and, finally, verification and validation of the systems). In these domains, final secured systems have been air-gapped or remain entirely inside secured networks. The problem with T&E and consistency in measurement is well understood by NIST's AISC, industry partners, and U.S. Defense officials who insist on safe and reliable systems (Flournoy et al., 2020). Requiring a degree of restrictions on open foundation models to allow continued T&E and the potential for recall is a smart policy to protect public safety and national defense.

e. *What, if any, risks could result from differences in access to widely available models across different jurisdictions?*

The generalizability of foundation models makes them inherently unpredictable, independent of jurisdiction. Therefore, a jurisdictional approach is no safer than completely open-sourcing model weights (i.e., because of how safety inputs can be removed with fine-tuning). Notwithstanding, it is probable that if an open weight model is available in one jurisdiction, malicious actors could easily transfer it to another.

However, foundation models without a general-purpose design, such as those for a wide variety of scientific applications, should continue to be open-sourced. As models grow more general purpose in nature, challenges can arise in legally differentiating which models should be open-sourced with widely available weights and which should not. At this point, it may be safest to say that models trained on a variety of data exceeding a certain threshold of computation be deemed general purpose and, therefore, should be governed differently—we would say the weights should not be shared once such classification is attained.

f. *Which are the most severe, and which the most likely risks described in answering the questions above? How do these set of risks relate to each other, if at all?*

Without guardrails, catastrophic risks are possible. The most severe risks include the development of dangerous and novel chemical and biological agents, with the possibility of unintentional release or attack, the automation of strategic weapons systems by state actors that lead to catastrophic failures (possibly triggering interstate conflict through nuclear means), or unintended failures of critical infrastructure through cascading system failures or cyber attack (perhaps using autonomous and adaptive agent systems). As agent systems increase in capability and become more widely available, it is uncertain how such systems will be used, the extent of penetration throughout networks (public and private), their capacity for independent adaptation, and our capacity for control.

Consider the potential for malicious actors to exploit vulnerabilities in critical systems (e.g., power grids)—that otherwise could have been protected, repaired, or identified through evaluations—leading to cascading failures across connected systems. Or the AI integration into a nuclear-armed state's command and control (C2) that experiences a failure, triggering a contagious interstate conflict. The premature integration of powerful, untested systems without adequate T&E (or potential for recall) could lead to such catastrophic events. Considering race dynamics, the premature integration of AI into defense systems could become far more likely as AI increases in strategic importance.

The most likely or near-term risks involve the spread of disinformation at scale, truth decay, and AI-enabled cyber. However, the more medium-term risks of unwise integration of AI into safety critical systems and cascading failures are equally likely, barring a significant accident that drives an increase in regulation (e.g., the Flash Crash of 2010) (Treanor, 2015).
The novelty of the risks and the demonstrated ability to remove safety guardrails—intentionally or unintentionally—are both incredibly concerning and likely, particularly as models grow in capability. With emergent properties, increased capability, and generality, continual evaluation and reasonable guardrails will become crucial (potentially impossible with unrestricted release). At the same time, likelihood and severity could also be a question of scale: with more opportunities for failures or attacks due to the ready availability of open foundation models, misaligned systems are more likely to slip through the cracks and through a greater number of connected systems.

Q5. What are the safety-related or broader technical issues involved in managing risks and amplifying benefits of dual use foundation models with widely available model weights?

As noted, foundation models are inherently "dual use," which has been the case with all general-purpose technologies (GPT). The more general a system, the wider the array of potential use cases, beneficial or dangerous. The introduction of autonomy opens up entirely new potential risks, many of which we have yet to contend with or imagine. While it is true that all technological revolutions have introduced radical change, none have demonstrated the ability to learn, adapt, and directly influence the environment.

As foundation models increase in generality, the scale and scope of use cases will only increase, posing novel risks, and the difficulties required of T&E will increase in kind. As an adaptive technology that can change as it learns and demonstrates new capabilities, standards for evaluation will remain a moving target, making safety and security especially difficult. The science of evaluation is in a preparadigmatic state, and as a result, the T&E ecosystem is and will continue to be fragmented, making the establishment of consistent and reliable models increasingly difficult.

a. *What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?*

Evaluations are critical in determining the benefits and risks of foundation models (open or closed). Once the model and its weights are released, evaluating novel capabilities or misuse risks becomes challenging. As dual use systems, the range of applications—e.g., novel misuse cases—are usually not known until the GPT has been in use for some time (e.g., the phonograph was invented to record and transcribe telegraph communications not music).

An important component of model evaluations is the overall capability and generality of the model and whether these properties will change or adapt through time with new capabilities. This is exceedingly difficult to measure with one shot and is a strong reason why continual evaluations are crucial. Once the model is released, evaluations or updates to fix dangerous components or prevent potentially dangerous behavior become extremely limited (e.g., security update or fix). For example, if the system is designed with autonomy in mind and takes very little effort from independent developers to enhance these capabilities (or instantiate goal-directed agents), it would be important to understand these boundaries prior to deployment.

Increasing the availability of T&E infrastructure could help identify potential threats early so as to address risks prior to deployment (e.g., the European Commission's AI Testing and Experimentation Facilities) (EDIH, 2023). Establishing an ecosystem of trusted third-party evaluators–from academia, industry, and government–could be invaluable for effective evaluations (Clarke, 2024).

b. *Are there effective ways to create safeguards around foundation models, either to ensure that model weights do not become available, or to protect system integrity or human well-being (including privacy) and reduce security risks in those cases where weights are widely available?*

Safety measures can help reduce the capacity for misuse, but those with access to model weights can easily reverse these measures. At the same time, fine-tuning can also reverse safety measures (described in Q1), and increasingly creative jailbreaking techniques can obviate many of the safeguards put in place by companies. Since open is a spectrum (e.g., open source, open weights, restricted weights, closed, etc.), the optimal path is to keep powerful frontier systems at least somewhere on that spectrum to provide technology developers the chance to identify potentially dangerous capabilities, conduct continued evaluations, safety fixes, and institute updates if needed.

c. *What are the prospects for developing effective safeguards in the future?*

No response.

d. *Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?*

As far as we know, there are currently no means to regain control of or restrict access to an open foundation model once it has been released with widely available weights. This is the primary problem with open weight models; once widely released, the world has access to the capability, for good or ill. There are scenarios where companies, governments, or authorities may wish to kill or decommission such systems—e.g., akin to a replicating virus such as the WannaCry ransomware (Hern, 2017)—and widely released open foundation models would restrict this ability.

e. *What if any secure storage techniques or practices could be considered necessary to prevent unintentional distribution of model weights?*

Standard information security procedures (e.g., ISO 27001 and ISO 27002) and cyber defense should be encouraged for all AI developers to prevent the release or theft of model weights. As systems increase in capability and strategic importance for U.S. interests, this could require additional oversight and mandates (although, as mentioned, this would be confined to only the most capable models at the frontier). Most large AI companies recognize the immense economic incentives, what they have to lose, and the threat from state and non-state actors (Anthropic, 2023). However, preventing intrusions from well-resourced nation state actors (by way of espionage or IP theft) will be difficult to impossible, requiring enhanced efforts by companies and regulators. Sophisticated state actors are highly likely to attempt to infiltrate premier AI organizations and also the external servers on which they train. This risk has been noted by leading think tanks such as RAND (Nevo et al., 2017) and some AI companies cognizant of these security concerns (Anthropic, 2023).

Elsewhere, we've proposed establishing a National Center(s) of Excellence (CoE) with a secure compute cluster for AI Safety, Testing, and Evaluation (as part of the NAIRR) and to protect sensitive IP such as model weights (Gruetzemacher et al., 2024). This secure environment would be designed specifically for researching AI safety, testing, and evaluation and for companies to feel secure in storing their data and model weights. This would require that the facility be air-gapped and adhere to protocols akin to those used in sensitive compartmented information facilities (SCIFs) by government officials when handling highly classified information. As highly capable models straddle the frontier of capabilities, such a secure facility could be key to preventing targeted intrusions by the most capable adversaries.

f. *Which components of a foundation model need to be available, and to whom, in order to analyze, evaluate, certify, or red-team the model? To the extent possible, please identify specific evaluations or types of evaluations and the component(s) that need to be available for each.*

No response

g. *Are there means by which to test or verify model weights? What methodology or methodologies exist to audit model weights and/or foundation models?*

No response

References

Anthropic. (2023, July 25). Frontier Model Security. https://www.anthropic.com/news/frontier-model-security#entry:146893@1:url

Anthropic. (2024, March 25). Third-party testing as a key ingredient of AI policy. https://www.anthropic.com/news/third-party-testing

Arntz, P. (2022, September 15). Explained: Fuzzing for security. MalwareBytes Labs. https://www.malwarebytes.com/blog/news/2022/09/explained-fuzzing-for-security

Desai, D. et al. (2024, April 18). 2023 Phishing Report Reveals 47.2% Surge in Phishing Attacks Last Year. Zscaler Blog. https://www.zscaler.com/blogs/security-research/2023-phishing-report-reveals-47-2-surge-phishing-attacks-last-year

(EDIH) European Digital Innovation Hubs. (2023). European Commission. Accessed May 26, 2024. https://digital-strategy.ec.europa.eu/en/activities/edihs

Executive Order 14110. (2023, October 30). Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The Executive Office of the President, United States of America. 88 FR 75191.

Flournoy, M., et al. (2020). Building Trust via Testing Adapting DOD's TEVV Enterprise for ML Systems. WestExec Advisors.

Hern, A. (2017, December 30).WannaCry, Petya, NotPetya: how ransomware hit the big time in 2017. The Guardian. https://www.theguardian.com/technology/2017/dec/30/wannacry-petya-notpetya-ransomware

Gruetzemacher, R., et al. (2024, February 2). Envisioning a Thriving Ecosystem for Testing & Evaluating Advanced AI. NIST AI EO RFI Comments. Docket Number: 231218-0309. https://www.nist.gov/system/files/documents/2024/02/15/ID019%20-%202024-02-02%20Wichita%20State%20University%20et.%20al%2C%20Comments%20on%20AI%20EO%20RFI.pdf

Henderson, P. (2024, January 11). Safety Risks from Customizing Foundation Models via Fine-Tuning. Human-Centered Artificial Intelligence (HAI), Stanford University. https://hai.stanford.edu/policy-brief-safety-risks-customizing-foundation-models-fine-tuning

Kaspersky. Shadowy innovation: how cybercriminals experiment with AI on the dark web. https://dfi.kaspersky.com/blog/ai-in-darknet (Accessed 2024, March 21)

Kapoor, S. et. Al. (2024). On the Societal Impact of Open Foundation Models. arXiv:2403.07918v1. https://arxiv.org/pdf/2403.07918.pdf

Kavanagh, J. & Rich, M.D. (2018). Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR2314.html. Also available in print form.

(NIST) National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

Nevo, S. et al. (2024). Securing Artificial Intelligence Model Weights: Interim Report. Santa Monica, CA: RAND Corporation, 2023. https://www.rand.org/pubs/working_papers/WRA2849-1.html.

Treanor, J. (2015. April 22). The 2010 'flash crash': how it unfolded. The Guardian. https://www.theguardian.com/business/2015/apr/22/2010-flash-crash-new-york-stock-exchange-unfolded

Zilong, L. et al. (2024). Malla: Demystifying Real-world Large Language Model Integrated Malicious Services. arXiv:2401.03315.