2024

# Considerations for a New AI Agency: Risks, Framework, and Inter-Agency Coordination

Robert W. Stewart

The release of the first consumer-focused generative artificial intelligence (AI) tool – ChatGPT 3 – gave the public the ability to interact with chatbots that rely on a machine learning technique using deep learning and Large Language Models (LLMs) to produce human-like responses to their prompts.[1] However, AI tools have been around long before the introduction of these generative tools. We are used to the Google search function auto-filling in our inquiry and platforms like YouTube recommending eerily relevant videos based on our viewing history.[2] Although AI-based consumer algorithms have produced concrete harm long before the release of these new tools,[3] based on the increased conversation and fear about AI in the popular discourse, governments have rushed to be the first mover in passing a comprehensive regulation for this fast-moving technology and its risks.[4]

Most recently, the Biden-Harris administration has issued an Executive Order titled "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (AI EO). It came just a few days before the EU held its AI Safety Summit but months after the EU released a preliminary draft of its AI Act that is awaiting "trilogues" before a final version can be passed.[5] In a seemingly shocking first in the technology industry, top executives of the companies who have developed the leading AI products – OpenAI CEO Sam

[1] Alyssa Stringer and Kyle Wiggers, "ChatGPT: Everything you need to know about the AI chatbot," *TechCrunch*, Nov. 6, 2023, https://techcrunch.com/2023/11/6/chatgpt-everything-to-know-about-the-ai-chatbot/.

[2] Danny Sullivan, "How Google Autocomplete Works in Search," *The Keyword* (Apr. 20, 2018), https://blog.google/products/search/how-google-autocomplete-works-search/; *see also* Cristos Goodrow, "On YouTube's Recommendation System," *Inside YouTube* (Sep. 15, 2021), https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.

[3] Nicole Turner Lee, Paul Resnick, and Genie Barton, Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms, *Brookings* (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[4] Tom Wheeler, "The Three Challenges of AI Regulation" *Brookings* (June 15, 2023), https://www.brookings.edu/articles/the-three-challenges-of-ai-regulation/

[5] The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Altman, Microsoft's Brad Smith, and Google's Sundar Pichai – have called on governments to regulate them.[6] This is in stark contrast to the executives running social media platforms a decade ago. While these platforms are not treated as information publishers as per Section 230 of the 1996 Communications Decency Act, and, thus, not liable for their users' posts, they showed no rush to responsibility as the AI executives seem to now.[7] Developers of this ground-breaking technology realize the grave risks of runaway AI and their inability to understand how their "black box" models work.[8] For example, Sundar Pichai and his co-executives have equated the gravity and widespread impact of this new technology to electricity or fire.[9] Someone in possession of a technology as transformative as electricity might not want to deal with such a responsibility. However, some are suspicious of industry calls for regulation.[10] The ones lobbying governments around the world are the same ones who are leading in the industry.[11] They have the most resources, computing power, talented data scientists, and the most users – and the largest market share. Their cries for regulation could be an attempt to cement their position as market leaders and first movers to keep out small start-ups from gaining market share by requiring them to hire teams of lawyers to help them navigate complex regulatory hurdles.[12]

[6] David McCabe, "Microsoft Calls for A.I. Rules to Minimize the Technology's Risks," *The New York Times* (May 25, 2023), https://www.nytimes.com/2023/05/25/technology/microsoft-ai-rules-regulation.html;

[7] Communications Decency Act of 1996, (CDA), Pub. L. No. 104-104 (Tit. V), 110 Stat. 133 (Feb. 8, 1996), codified at 47 U.S.C. §§223, 230.

[8] Cat Zakrzewski, Cristiano Lima and Will Oremus "CEO behind ChatGPT warns Congress AI could cause 'harm to the world'," *The Washington Post*, (May 16, 2023), https://www.washingtonpost.com/technology/2023/05/16/sam-altman-open-ai-congress-hearing/

[9] Prarthana Prakash, "Alphabet CEO Sundar Pichai says that A.I. could be 'more profound' than both fire and electricity—but he's been saying the same thing for years," *Fortune* (Apr. 17, 2023), https://fortune.com/2023/04/17/sundar-pichai-a-i-more-profound-than-fire-electricity/

[10] Gerrit De Vynck, "Big Tech wants AI regulation. The rest of Silicon Valley is skeptical." *The Washington Post*, (Nov. 9, 2023), https://www.washingtonpost.com/technology/2023/11/09/ai-regulation-silicon-valley-skeptics/

[11] *Id*.

[12] *Id*.

This is akin to pioneers discovering gold on a new land, growing their wealth, building a castle, and then building a miles-long moat around the land to prevent others from digging.

In the spring of 2023, Sam Altman told the Senate Judiciary Committee that there was a need for "a new agency that licenses any effort above a certain scale of capabilities and could take that license away and ensure compliance with safety standards."[13] This paper will discuss, in Part I, the nature of machine learning algorithms as a subset of all algorithms, the harms such models have caused and now pose, and what problems are worthy of some government regulation; in Part II, the current rules, mainly from the Federal Trade Commission, and to regulate the risks of these models; and, in Part III, what a new AI agency might look like based on the current proposals and how to prevent haphazard, delayed, and fragmented government response that comes with agency turf wars. This paper ultimately argues that a new agency, if Congress authorizes one, must respond and work closely with other agencies to ensure efficient and effective regulation.

## Part I: AI RISK AND HARMS

Algorithms, particularly in the form of software and applications, are increasingly influential in our daily lives, powering devices and platforms such as personal computers, smartphones, search engines, and online stores. They play a crucial role in organizing and filtering vast amounts of information. Machine learning algorithms represent a revolutionary category of algorithms. Unlike traditional algorithms that follow predefined instructions to solve

[13] U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, and the Law "Oversight of A.I.: Rules for Artificial Intelligence," (May 16, 2023).

specific problems, machine learning algorithms learn from data and improve over time, enabling them to address problems innovatively and adapt to new information.

The first step in the working of a machine learning algorithm is data input.[14] This data can be in various forms, such as images, text, numbers, or any other measurable attribute of the phenomenon being observed.[15] The algorithm is then "trained" using a subset of this data.[16] Training involves feeding the data to the algorithm and allowing it to learn and identify patterns.[17] The algorithm then iteratively adjusts its operations based on the accuracy of its predictions compared to the known outcomes in the training data.[18] During training, the algorithm makes predictions or decisions based on the data it has, and these outcomes are compared with the expected results.[19] The differences between the predicted and actual outcomes are used to adjust the algorithm.[20] This adjustment helps the algorithm to make more accurate predictions in the future.[21] After training, to determine if the algorithm has genuinely "learned" to identify patterns or if it's just memorizing specific data, the algorithm is tested with a new set of data that it has *not* seen or encountered before.[22] Once trained and validated, the algorithm can be used in real-world applications to make predictions or decisions based on new data.[23]

There are a variety of techniques used. In brief, here are a few important ones to the discussion. Supervised learning algorithms are trained using labeled data, which means that each

[14] Brown, Sara. "Machine Learning, Explained" *Massachusetts Institute of Technology Sloan School of Management*. (April 21, 2021). https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
[15] Id.
[16] Id.
[17] Id.
[18] Id.
[19] Id.
[20] Id.
[21] Id.
[22] Id.
[23] Id.

example in the training dataset is tagged with the correct answer.[24] The algorithm then learns to predict the output from the input data.[25] For instance, a supervised learning algorithm could be used for email filtering, where it learns to classify emails into 'spam' and 'non-spam' categories.[26] In unsupervised learning, the algorithm is trained using data that is not labeled.[27] The goal here is to explore the structure of the data to extract meaningful information without guidance.[28] Reinforcement learning uses a system of rewards and penalties to compel the computer to solve a problem by itself.[29] It's often used in areas where decision-making is sequential, like in game-playing or autonomous vehicles.[30] Ultimately, researchers take these basic principles and model their algorithms as (really large) human brains and layer those in various ways in what is known as a deep neural network.[31] They often become so complex and constantly evolving that they are so-called black-boxes.[32]

Before this discussion moves to unique problems caused by these so-called large foundation models, there are three examples, from before the advent of generative AI and deep neural networks, of traditional machine learning algorithms deployed by Google and IBM failing in unexpected ways. Google's image recognition system, trained to classify images using supervised learning, mistakenly identified images of black individuals as "Gorillas."[33] IBM's

[24] Dulua, Julianna. "Supervised vs. Unsupervised Learning: What's the Difference?" (March 12, 2021). *IBM* *https://www.ibm.com/blog/supervised-vs-unsupervised-learning/#:~:text=Supervised%20learning%20is%20a%20machine,accuracy%20and%20learn%20over%20time*
[25] Id.
[26] Id.
[27] Id.
[28] Id.
[29] Kaelbling, Leslie P.; Littman, Michael L.; Moore, Andrew W. *(1996).* Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*. **4**: 237–285. *arXiv*:*cs/9605103*
[30] Id.
[31] Brown, Sara. "Machine Learning, Explained" *Massachusetts Institute of Technology Sloan School of Management*. (April 21, 2021).
[32] Id.
[33] Zhang, Maggie. "Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software" *Forbes* (July 1, 2015). https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=67b51ce0713d

supercomputer Watson shows another example. Even though it was victorious in its *Jeopardy!* match against two of the best human players, it committed an error that even a novice human player wouldn't make.[34] In final *Jeopardy!*, under the category "U.S. Cities," Watson chose a city that was not in the United States.[35] The hint included that the city has its largest airport named after a WWII hero and its second largest after a WWII battle.[36] As the humans knew, the answer was "Chicago."[37] Yet, Watson responded, with low confidence, "What is Toronto?????"[38] Importantly, researchers and technologists responsible for these algorithmic failures could not precisely figure out the reasons the model failed unexpectedly.[39]

In the years since these major incidents, AI and machine learning models have increased in complexity, efficiency, and opacity, which makes it even more impractical or impossible to look into the model's decisions to identify the source of its errors. The ability of these models also has significantly increased, meaning that they can produce benefits that the human mind cannot produce alone or in groups but also can be misused by malicious actors to carry out harm on a greater scale than a TV gameshow or a job application. The AI EO categorizes some of these risks with cutting-edge foundation models.

The AI EO defines these powerful, new models as dual-use foundation models:

> an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be *easily modified* to exhibit, high levels of performance

[34] Kanalley, Craig. "Watson's Final Jeopardy Blunder In Day 2 of IBM Challenge" *HuffPost* (Dec. 6, 2017) https://www.huffpost.com/entry/watson-final-jeopardy_n_823795
[35] Id.
[36] Id.
[37] Id.
[38] Id. See generally Markoff, John. "Computer Wins on 'Jeopardy!'" Trivial, It's Not" *The New York Times* (Feb. 16, 2011) https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html
[39] Baker, Stephen. "Final Jeopardy: Man vs. Machine and the Quest to Know Everything" *Houghton Mifflin Harcourt* (2011).

> at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters

(Emphasis added).[40] These models allegedly possess the dual capability to revolutionize numerous beneficial applications while simultaneously posing substantial risks to security, public health, and safety.[41] These large, "open-source" foundation models reduce the cost and technical expertise needed to build both harmful and beneficial machine learning systems on top of those foundation models.[42] It is difficult to distinguish between benign and malevolent applications of these models because they lower the barriers to applying machine learning in various contexts.[43] They can be adapted to many narrow tasks with little new training data and be fine-tuned from the models' general-purpose nature to bypass safeguards to meet malicious ends such as the identification of dissidents by oppressive governments or the creation of targeted weapons.[44] The AI EO lists an example of a dual-use foundation model as one that, "substantially lower[s] the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons."[45] If a non-expert were able to access a model like this that was trained on primarily biological inputs that enable researchers to discover novel biologics and pharmaceutical compounds, he or she could "easily" fine-tune the parameters of such a model to predict and create not health-promoting or disease-fighting drugs but the cheapest, most toxic, most contagious, and easily-designed molecules.[46]

[40] The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, (Oct. 30, 2023).
[41] Id.
[42] Id.
[43] Id.
[44] Id.
[45] Id.
[46] Id.

Moreover, as far as regulation is concerned, it is difficult to identify which foundation models are dual use. These models, by their very nature, are expansive and versatile. They are trained on diverse datasets, enabling them to perform exceptionally across various contexts, making it difficult to pinpoint when a model transitions from benign to dual-use, especially when the model weights are publicly available. It seems that the government can classify a model as dual-use if it reaches some threshold capacity. This discussion will reference sections of the AI EO later to demonstrate features of a potential new AI agency that would, in part, set such capacity thresholds and certain requirements for models that meet those thresholds.

The AI EO lists one example of how a dual-use foundation model could pose a risk to security and public safety: "by… permitting the evasion of human control or oversight through means of deception or obfuscation."[47] This scenario comes to fruition when the underlying mechanisms of dual-use foundation models are opaque or rendered obscure by the nature of their complexity.[48] As previously mentioned, these self-improving black box models are deep learning statistical machines that update the weights on the billions of parameters of its model in response to several factors including human feedback, The AI EO emphasizes here that the complexity of the entire system evades not just human control but human understanding of how the model outputted certain predictions based on the inputs.[49] Aside from outright deception, the ability of these models to be unintelligible – even to their developers – underscores why they pose a threat (why they are dual-use).[50] "What makes [these models] valuable is what makes them uniquely

[47] Id.
[48] Id.
[49] Id.
[50] Id.

hazardous."[51] The next section addresses how current rules and the rule-making authority of existing agencies, particularly under the Federal Trade Commission, could and have been addressing some of these harms before discussing what a new, centralized agency could do to address these unique, inherent problems with advanced foundation machine learning models.

## Part II REVIEW OF CURRENT RULES AND INTER-AGENCY EFFORTS

Recently, the proposals of the AI EO call upon inter-governmental coordination for the regulation of certain machine learning models and their risks.[52] However, agencies, like the Federal Trade Commission (FTC) have made efforts to make rules that curb the risks of AI. The FTC's rule-making authority arguably covers many of the harms that inherently result from the design and deployment of AI, including the black box problem, transparency, and algorithmic bias and discrimination.[53] Additionally, its requirements for disclaimers, accountability in the form of a duty to monitor for misuse, impact assessments, audits, and enforcement actions could serve as useful tools in effectively regulating the harms of AI models.[54]

Importantly, the FTC does not have a purview to regulate AI per se but has the authority to regulate the consumer harms that result from the use of AI.[55] In many instances, because harms are built into the design of the models, the FTC could have the authority to effectively regulate AI models directly.[56] By this general approach, the FTC has not adopted a definition of

[51] Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN L. REV. 83 (2017) (addressing the risks of machine learning models)
[52] The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, (Oct. 30, 2023).
[53] Federal Trade Commission, FTC Report to Congress: Combatting Online Harms Through Innovation (2022) https://www.ftc.gov/system/files/ftc_gov/pdf/Combatting%20Online%20Harms%20Through%20Innovation%3B%20Federal%20Trade%20Commission%20Report%20to%20Congress.pdf
[54] Id.
[55] Federal Trade Commission, A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority (May 2021) https://www.ftc.gov/about-ftc/mission/enforcement-authority
[56] Federal Trade Commission, FTC Report to Congress: Combatting Online Harms Through Innovation (2022)

AI because, for its purposes, it is only concerned with regulating consumer harms, unfair business practices, and deception that result from automated decision-making broadly, especially considering that there are many competing definitions of AI, which, if it adopted one, the FTC might unnecessarily narrow its rule-making authority.[57] In interpreting its mandate from Congress, the FTC, "assume[s] that Congress is less concerned with whether a given tool fits within a definition of AI than whether it uses computational technology to address a listed harm."[58]

In its correspondence with AI companies, the FTC has heard arguments from executives and engineers that essentially amount to them playing the "black box" card when asked to substantiate claims that it does not deceive its users.[59] By claiming that the complexity and evolving nature of the underlying models evade their understanding, these developers have tried to avoid the FTC holding them liable for deceptive practices.[60] As mentioned in the AI EO, the ability of dual-use models to deceive implies that the FTC would have proper authority over such matters.[61] No actual deception is required; the FTC requires only that deception associated with a machine learning model be reasonably foreseeable by the company in connection with its products.[62] Equally relevant for dual-use models and any machine learning models, the FTC does not require that the company developing such models have the intent to deceive the users of its products.[63] Additionally, the FTC has required companies to institute reasonable consumer injury

[57] Id.
[58] Id.
[59] Michael Atleson, Keep your AI claims in check, Federal Trade Commission (Feb. 27, 2023). https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check
[60] Id.
[61] Id. *See generally* The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, (Oct. 30, 2023).
[62] Id.
[63] Michael Atleson, Chatbots, deepfakes, and voice clones: AI deception for sale, Federal Trade Commission (March 20, 2023). https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale

deterrence measures before it releases its products and has brought enforcement actions against those companies that have not taken such measures.[64] This shows that the FTC has enforcement authority ex-ante; it could prevent a potentially harmful product from being deployed.[65]

Congress has called on the FTC to investigate AI companies for the potential ability of models to be used in ways that would "cause substantial injury to consumers."[66] In response, the FTC has acknowledged that a model could cause such injury in many ways that are a result of gaps in training data, misclassifications, or other algorithmic flaws, which allows its enforcement power over unfair trade practices to extend to the underlying algorithms.[67] The following are some examples of recent agency efforts to mitigate certain harms that could result from anything under the umbrella of unfair practices.

The FTC has worked with other agencies like the Consumer Financial Protection Bureau (CFPB), Department of Justice, and Equal Employment Opportunity Commission (EEOC), to express increasing concern about the potential for bias and discrimination in AI tools.[68] This concern is particularly focused on the issues of inaccuracy, biased datasets, and the opaque nature of many AI models, which often result in unfair business practices and discriminatory outcomes.[69] As mentioned, for the FTC to initiate an enforcement action, the harm caused must be substantial and a consumer cannot reasonably avoid it on their own. This includes situations where AI is involved wholly or partially in making decisions that affect consumers in critical

[64] Id.
[65] Id.
[66] 15 United States Code § 45(n). See also FTC, FTC Report to Congress: Combatting Online Harms Through Innovation (2022)
[67] FTC, FTC Report to Congress: Combatting Online Harms Through Innovation (2022)
[68] Chopra, Rohit; Clarke, Kristen; Burrows, Charlotte A.; and Khan, Lisa M. "Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems" (Apr. 25, 2023) https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf
[69] FTC, FTC Report to Congress: Combatting Online Harms Through Innovation (2022)

areas like credit, employment, insurance, or housing.[70] In these cases, the FTC requires companies to provide clear explanations to consumers when AI influences decisions that might adversely affect them.[71]

Another key area of the FTC's focus is the transparency of AI systems, both within companies and to the public. Particularly, the FTC has released a statement of best practices for companies' transparency, such as testing their algorithms, publishing independent audit results, and disclosing how the company uses consumer data.[72] Additionally, the FTC emphasizes that data safety measures should be robust and designed into companies' models to prevent any foreseeable substantial harm that could result from data breaches.[73] Moreover, the FTC has increased its emphasis on the duty of these companies to monitor their AI products for misuse. This includes regular audits and providing redress for erroneous or unfair algorithmic decisions.[74]

Relevant to the justifiably suspicious calls from leading AI companies for new regulations requiring pre-market approval and licenses, the Chair of the FTC has argued that the FTC will work to prevent collusion and concentration in the AI marketplace.[75] The FTC learned from the rise of revolutionary social media platforms wherein a few private companies with all of our data have managed to wield outsized power.[76] Still reckoning with the monopolistic

[70] Id.
[71] Id.
[72] Elisa Jillson, Aiming for truth, fairness, and equity in your company's use of AI, Federal Trade Commission (April 19, 2021).
[73] Michael Atleson, Chatbots, deepfakes, and voice clones: AI deception for sale, Federal Trade Commission (March 20, 2023).
[74] Slaughter, Rebecca K.; Kopec, Janice; and Batal, Mohamad, *Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission*, Yale J. L. & Tech. (Aug. 2021).
[75] Khan, Lisa M. "We Must Regulate A.I. Here's How." *The New York Times*. (May 3, 2023) https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html
[76] Id.

tendencies of large social media platforms, the FTC Chair is wary that efforts to broaden the range of risks from AI to include hypothetical, existential threats of AI, much like the ones cited in the AI EO, to pressure the government to regulate even more will cement, "the market dominance of large incumbent technology firms," permitting them to easily outcompete, "against downstream rivals."[77]

All of these proposals, rules, and enforcement actions are still based on the outputs of AI and are not concerned with regulating algorithmic processes. However, the next section will discuss how a new agency might take a more hard-edged approach involving capacity thresholds, reporting requirements, and approvals, and how such an agency could coordinate with agencies like the FTC that are heavily involved and interested in mitigating AI risks and harms.

## Part III: REGULATORY DESIGN AND INTER-AGENCY COORDINATION

### A: Regulatory Design

The proposed new agency will likely adopt approaches from similar agencies in regulating dangerous technologies like nuclear reactors and biological agents. Before building a reactor, one needs a license from the NERC. Before selling a new pharmaceutical, one needs to make a safety case to the Food and Drug Administration (FDA).

One proposal for a new agency modeled after the FDA involves a multi-faceted approach.[78] The reason to model the new agency after the FDA is that processes underlying complex pharmaceutical drugs, much like those that make up advanced machine learning

[77] Id.

[78] *See generally* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017) (proposing a federal agency to oversee AI and machine learning)
algorithms)

models, are difficult to understand.[79] So, unlike the FTC's focus on output and harm, this agency would be in response to calls for regulation of the underlying models. The agency could act as a standards-setting body, developing categories for classifying algorithms based on their complexity and setting guidelines for their design, testing, and performance.[80] Establishing categories underlies Altman's suggestion and echoes the EU Artificial Intelligence Act's risk-based tiered approach to regulation.[81] OpenAI proposes to set capability thresholds, wherein the strictest regulations (or even an outright ban, in the case of the EU AI Act's category of prohibited risk) would apply to models that could persuade, manipulate, or influence, or models used to create novel biological agents (drugs and bioweapons).[82]

The proposal suggests a soft approach that would encourage explainability and transparency standards and a hard approach requiring pre-market approval for complex algorithms used in critical infrastructure to ensure they meet safety and performance standards before deployment e.g., a self-driving car algorithm might need to match the safety-per-mile of typical vehicles driven in a particular year to gain approval.[83]

The new AI EO proposals overlap some of these suggestions but provide a more specific framework that an AI agency might adopt in the future and advocate for an interagency approach as opposed to a new AI agency (as the President does not have the authority to create a new agency through an Executive Order). This discussion will briefly touch on some of these

[79] *Id*.
[80] *Id*.
[81] Sam Altman, Greg Brockman, and Ilya Sutskever, "Governance of Superintelligence" *OpenAI*, (May 22, 2023), https://openai.com/blog/governance-of-superintelligence; *see generally* European Commission, "REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS."
[82] Altman et. al.
[83] Tutt at …

proposals but will focus on some of the rules concerning dual-use foundation models' capacity thresholds and concomitant requirements.

Section 4 of the AI EO called "Ensuring the Safety and Security of AI Technology" shares some features of these regulatory approaches.[84] This is the most relevant section of the AI EO for the present discussion, but parts from the rest of the EO will be addressed in the remainder of the discussion. Section 4 establishes a comprehensive framework for regulating and managing AI that seems to take a similar risk-based approach, specifying high-risk examples of AI use and how to manage that risk.[85] It generally focuses on safety, security, and reliability.[86]

Section 4 mandates the development of guidelines, standards, and best practices for creating safe, secure, and trustworthy AI systems, including resources for generative AI and dual-use foundation models, as well as establishing guidelines for AI red-teaming.[87] The EO requires companies that are developing or have developed dual-use foundation models to report on various aspects of their model such as cybersecurity measures and model performance.[88] The reporting requirements will be discussed in more detail in the next paragraph. This section also addresses the integration of AI in critical infrastructure and cybersecurity, assessing risks in critical sectors, and incorporating the AI Risk Management Framework NIST AI 100-1 from the National Institute of Standards and Technology into safety guidelines.[89] It emphasizes evaluating

[84] The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
[85] *Id*.
[86] *Id*.
[87] Red-teaming is defined as "means a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red-teaming is most often performed by dedicated "red teams" that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system."
[88] *Id*.
[89] *Id*. (referencing AI Risk Management Framework NIST AI 100-1)

AI's potential misuse in creating chemical, biological, radiological, and nuclear (CBRN) threats, particularly focusing on biological weapons, and recommends steps to mitigate such risks.[90] Additionally, it aims to manage synthetic content produced by AI, setting standards for authenticating and detecting such content.[91] The order solicits input on the risks and benefits of dual-use foundation model weights[92] that are widely available to the public.[93] It also establishes guidelines for performing security reviews of federal data to prevent its misuse in developing CBRN weapons or offensive cyber capabilities.[94] Finally, section 4 directs the development of a National Security Memorandum to govern AI used in national security systems, focusing on AI assurance and risk management.[95]

Aside from enabling the NIST director, Secretary of Commerce, and the heads of other relevant agencies to create "guidelines and best practices…for developing and deploying safe, secure, and trustworthy AI systems," the AI EO acts to "ensure and verify the continuous availability of safe, reliable, and effective AI."[96] This section calls on the Secretary of Commerce to establish reporting requirements for dual-use foundation models and, by these requirements, define and appropriately update the technical conditions that qualify "models and computing clusters" to abide.[97]

[90] *Id*.
[91] *Id*.
[92] The EO defines model weights as "a numerical parameter within an AI model that helps determine the model's outputs in response to inputs."
[93] *Id*.
[94] *Id*.
[95] *Id*.
[96] Id.
[97] Id.

First, from whom does this provision require reports? Companies that are creating or have an intent to create potential dual-use foundation models.[98] It also requires any entities that "acquire, develop, or possess a potential large-scale computing cluster to report any such acquisition, development, or possession," but does not require these entities to provide ongoing reports as it does for the developer companies.[99] Section 4.2(a)(i)(A)-(C) requires these developer companies to report any activities related to training or developing such a model, including measures taken to guard the integrity of the development process; who owns or possesses the weights of the hidden layers of the model, including measures to secure the weights; and AI red-teaming test results, including (before NIST develops red-teaming standards) results concerning mitigating potential creation of biological weapons, discovering software weaknesses, using software that influences events, and assessing the chances that a model could replicate itself.[100]

Section 4.2(b) includes thresholds for when it deems a model or computing cluster to rise to the level of a dual-use foundation model.[101] For a model, "trained using a quantity of computing power greater than $10^{26}$ integer or floating-point operations," or for models that are trained "primarily" on biological sequence information (e.g., DNA) and only have a quantity of computing power greater than $10^{23}$ integer or floating-point operations."[102] For a computing cluster, that cluster must include, "a set of machines physically co-located in a single datacenter, transitively connected by data center networking of over 100 Gbit/s, and having a theoretical

[98] Id.
[99] Id.
[100] Id.
[101] Id.
[102] Id.

maximum computing capacity of $10^{20}$ integer or floating-point operations per second," for it to be subject to compliance to ongoing reporting requirements.[103]

Many of these requirements are concerned with capacity, as Altman suggested, instead of the likelihood of harm as the FTC examines. However, in effect, these requirements for foundation models were created because they are dual-use. By definition, their size and features inherently hold the potential for harm. Additionally, this AI EO accounts for CBRN threats, which exceed the degree and nature of the discriminatory harms to consumers that the FTC has and plans to regulate. As such, there is not a one-to-one overlap, but differences that suggest these agencies will exist in a shared regulatory space and should thus coordinate to avoid redundancies and fragmented approaches. This is particularly important to avoid situations like the government's lack of regulatory coordination contributing to its failure to prevent the terrorist attacks on September 11, 2001.[104]

## B: Avoiding Regulatory Capture

One concern mentioned with creating a new agency to institute hard-edged capacity thresholds, pre-market testing and approval, and model disclosures is that there is significant overlap with other agencies such as the FTC. Creating an entirely new agency, on top of the regulations proposed and enforced by existing agencies, places more regulatory hurdles in front of AI companies and potentially gives incumbent AI companies a first-mover advantage and a

[103] Id.

[104] NAT'L COMM'N ON TERRORIST ATTACKS UPON THE UNITED STATES, *THE 9/11 COMMISSION REPORT* (2004), https://govinfo.library.unt.edu/911/report/911Report.pdf

chance to mold the regulations of this new agency. However, there are ways to design an agency that could mitigate the risk of this so-called regulatory capture by private interests.[105]

First, insulation from political and interest group pressures is crucial, especially appropriate given that one Senator, in the aforementioned hearing, asked Sam Altman whether he would like to chair the proposed agency.[106] One effective method is the appointment of agency heads for fixed, staggered terms.[107] This would reduce the influence of any single administration or political wave.[108] Additionally, these heads should possess specific qualifications in machine learning and neural network technologies to ensure decisions are grounded in expertise rather than political or industry bias.[109]

Traditional pillars of independence, such as for-cause versus at-will removal provisions, are essential but not sufficient on their own.[110] While for-cause provisions offer some protection against direct political interference, they can still be subject to legal battles and political maneuvering.[111]

Oversight by external bodies, like the Office of Information and Regulatory Affairs (OIRA), should be considered.[112] While oversight is necessary for coordination and consistency with broader administrative policies, excessive control can undermine the agency's

[105] Rachel E. Barkow, *Insulating Agencies. Avoiding Capture Through Institutional Design*, 89 TEX. L. REV. 15 (2010).
[106] U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, and the Law "Oversight of A.I.: Rules for Artificial Intelligence," (May 16, 2023).
[107] Barkow at 29.
[108] Id.
[109] Id.
[110] Id.
[111] Id.
[112] Id.

independence.[113] A balanced approach would involve limited OIRA oversight, focused more on the process and less on substantive decisions.[114] This will allow the agency to make technical decisions based on AI expertise while maintaining overall alignment with governmental policy.[115]

Moreover, transparency in decision-making processes can mitigate the risk of capture. Open hearings, public comment periods, and clear, data-driven rationale for decisions can help prevent undue influence from powerful tech companies, like Microsoft or OpenAI.[116] This transparency not only builds public trust but also allows for academic and civil society scrutiny, which can provide a counterbalance to industry pressure.[117]

### C: Inter-Agency Conflict and Coordination

Another concern with the creation of a new agency tasked with overseeing AI is that it will conflict with other agencies, like the FTC, whose purviews concern regulating AI's harms. The establishment of a new agency to regulate machine learning models will likely lead to significant conflicts with existing regulatory bodies, resulting in ineffective, overly complicated, or duplicative regulation. However, there are opportunities for cooperation and close coordination.

First, AI has already caused harm, which the FTC addresses and plans to address. With the rise of new dual-use foundation models, the risks are ever-present. Any promise of regulation

[113] Id.
[114] Id.
[115] Id.
[116] Id.
[117] Id.

from a new agency would be delayed because proposals would need to go through the drawn-out legislative process before it could even become law and before an agency can be staffed to start writing and enforcing administrative orders. By the time that happens, AI will do more harm. Fortunately, the AI EO calls for inter-governmental regulation to go into effect in a much shorter timescale.[118] Still, this article is more concerned with how an agency that incorporates some of the hard-edged requirements from the AI EO would coordinate with existing agencies like the FTC.

Additionally, creating and implementing a comprehensive, unified regulatory framework for AI in the United States, is challenging, especially accounting for existing rules and orders and the current efforts of agencies (not just the FTC) like the Federal Communications Commission (FCC), the FDA, the National Highway Safety Administration (NHSA), and others to regulate the effects of AI in their sectors.[119] The FCC, for example, is exploring AI's role in spectrum management and consumer protection from robocalls[120], while the Federal Election Commission is dealing with AI's impact on political ads and the potential for deepfake misinformation.[121] These examples show that existing agencies are already stretching to incorporate AI regulation within their purviews, suggesting that the addition of a new AI-specific agency could lead to duplication of efforts, jurisdictional conflicts, and increased regulatory complexity. Would the

[118] The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, (Oct. 30, 2023).
[119] Tutt at 114-115.
[120] Federal Communications Commission *FCC Fact Sheet: Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts* (October 25, 2023) https://docs.fcc.gov/public/attachments/DOC-397988A1.pdf
[121] Federal Election Commission, *Comments sought on amending regulation to include deliberately deceptive Artificial Intelligence in campaign ads* (August 16, 2023) https://www.fec.gov/updates/comments-sought-on-amending-regulation-to-include-deliberately-deceptive-artificial-intelligence-in-campaign-ads/

creation of a new agency and its concomitant adoption of a unified regulatory framework supplant or supplement existing regulatory schemes?

Ultimately, this paper argues that, for the most part, each agency and governmental department, with possible guidelines from Congress or the Executive to coordinate any inter-agency or -governmental efforts, should govern AI to the extent that it affects the sectors over which it has a mandate to issue administrative rulings. This approach would likely be more successful in circumventing any potential agency turf wars and would likely be more effective than creating a centralized agency tasked with regulating AI. Because one foundation model can be applied to disparate use cases, from creating video games and art to understanding how a cell works and developing novel biological agents, it would be short-sighted to concentrate the regulation of AI in a central authority. Creating a new, bespoke, centralized agency tasked with merely regulating AI would have such an outsized mandate and unrealistic goal compared to bolstering the current governmental and administrative state tasked with regulating different sectors, including high-risk sectors like defense and cybersecurity, as an extension of their current efforts.

However, if Congress grants the authority to create such a new agency, the overarching goal would aim to avoid conflict, redundancy, and fragmented approaches to ensuring a coherent and unified regulatory framework for AI. To achieve this, a multifaceted approach to coordination, employing various tools and strategies, would be essential. First, establishing clear consultation provisions would be paramount.[122] Discretionary consultation, wherein the new AI agency would be authorized but not required to consult with existing agencies, would foster a

[122] Jody Freeman & Jim Rossi, *Agency Coordination in Shared Regulatory Space*, 125 HARV. L. REV. 1131 (2012)

collaborative environment.[123] Given, however, the complexity and far-reaching impact of AI as well as the potential for a new agency to overpower existing agencies that have AI regulation properly in its sights, mandatory consultation might be more appropriate. This would ensure that before the new agency takes significant actions, it consults with agencies like the FTC or FCC, which have domain expertise in areas affected by AI, such as consumer protection and communications, respectively.[124] An example of this can be drawn from the Endangered Species Act, where agencies are required to consult with wildlife agencies to ensure that their actions do not jeopardize protected species.[125]

Further, default position requirements could be instituted.[126] For example, in matters overlapping with the NHTSA's domain, such as AI in autonomous vehicles, the new agency's default position could be in line with NHTSA's recommendations, deviating only when necessary. This is akin to the Federal Power Act's approach, where the Federal Energy Regulatory Commission must consider other agencies' recommendations as their default, and only deviate from this default in exceptional circumstances.[127]

Concurrence requirements – mechanisms whereby one or more agencies must agree or give their approval before a particular action or policy decision can be finalized – could also be effective, particularly in areas where the stakes of AI deployment are high, and the expertise of multiple agencies is crucial.[128] For instance, the use of AI in hiring and employment, where the EEOC has jurisdiction, might require that any new regulations by the AI agency receive

[123] Id.
[124] Id.
[125] U.S. Gov't Accountability Office, GA 04-590, *Border Security: Agencies Need to Better Coodinate Their Strategies and Operations on Federal Lands*, (June 2004) https://www.gao.gov/assets/gao-04-590.pdf
[126] Freeman at 1159
[127] Id.
[128] Id at 1160.

concurrence (agreement) from the EEOC to ensure non-discrimination and fairness in AI-assisted or -enabled hiring practices.

Additionally, interagency agreements, particularly Memoranda of Understanding (MOUs), would be useful.[129] These MOUs could outline specific areas of responsibility, establish procedures for collaboration, and detail information-sharing protocols.[130] This approach, while flexible, would create a framework for ongoing cooperation and avoid jurisdictional conflicts.[131] For example, an MOU between the new AI agency and the FTC could delineate responsibilities in regulating AI in consumer goods and services, ensuring that both agencies' efforts are complementary rather than duplicative.

Joint policymaking is another critical strategy.[132] By collaboratively developing guidelines and regulations much like the joint statements by other agencies cited earlier in the discussion, the new AI agency and existing agencies can ensure that the regulatory landscape is seamless and coherent. This approach would be particularly useful in areas where AI intersects with multiple regulatory domains, such as AI in telecommunications, where both the FCC's and the new agency's expertise would be vital.[133]

## PART IV: CONCLUSION

Writing a paper in the Fall of 2023 about AI regulation is like shooting at a moving target. However, this discussion argues against centralizing AI regulation in a single new agency

[129] Id at 1161
[130] Id.
[131] Id.
[132] Id. at 1165.
[133] Id.

and advocates instead for leveraging the expertise of existing agencies in specific sectors to work with a new AI agency. It underscores the importance of clear inter-agency coordination and consultation to avoid conflicts and ensure a unified regulatory framework. This approach acknowledges the diverse, dangerous applications of deep neural networks used in foundational models and how the purview and expertise of existing agencies like the FTC can complement a new AI agency tasked with regulating underlying AI processes to effectively navigate the challenges posed by this transformative technology.