Trade Regulation Rule on Commercial Surveillance and Data Security Rulemaking

We are pleased to submit a public comment in response to the trade regulation rule on commercial surveillance and data security.

This comment is structured in response to the following description of unfair practices covered under Section 5 of the Federal Trade Commission (FTC) Act of 1914:

“Generally, a practice is unfair under Section 5 if (1) it causes or is likely to cause substantial injury, (2) the injury is not reasonably avoidable by consumers, and (3) the injury is not outweighed by benefits to consumers or competition.”[1]

We make two recommendations. Firstly, because AI systems change the behavior of populations exposed to them, harms that stem from systemic behavior changes (e.g., spread of disinformation on social media) must be measured in addition to direct algorithmic harms. Secondly, that any federal agencies that procure vendored AI systems submit annual reports that describe the foreseeable impacts of the system on consumer behaviors. These policies are necessary to ensure the Section 5 mandate remains enforceable as automated systems increasingly make strategic decisions based on human data. The types of harms covered by Section 5 will become both increasingly common and difficult to evaluate without requisite rulemaking.

Today, regulators and policymakers focus on litigating isolated algorithmic harms, such as model bias or privacy violations. But this agenda neglects the persistent effects of AI systems on consumer populations. Meanwhile, leading research labs increasingly concentrate on such persistent effects by attempting to specify the purpose or “objective” of AI systems in ways that limit negative outcomes.[2,3,4,5,6]

Our comment highlights the need to reconcile these agendas through new forms of rulemaking so that the FTC can take an *ex ante* approach to AI-enabled consumer vulnerabilities. In particular, the FTC has a duty to track foreseeable harms that result from how systems make decisions based on consumer data. This tracking should be achieved by documenting how AI systems impact consumer behaviors over time.

The argument of our comment proceeds as follows:

---

[1] Quoted from here: https://www.regulations.gov/document/FTC-2022-0053-0001
[2] Russell, Stuart. "Human-compatible artificial intelligence." *Human-Like Machine Intelligence* (2021): 3-23.
[3] Andrus, McKane, et al. "AI development for the public interest: From abstraction traps to sociotechnical risks." *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2020.
[4] Zhuang, Simon, and Dylan Hadfield-Menell. "Consequences of misaligned AI." *Advances in Neural Information Processing Systems* 33 (2020): 15763-15773.
[5] Stray, Jonathan. "Beyond engagement: Aligning algorithmic recommendations with prosocial goals." *Partnership on AI. https://www.partnershiponai.org/beyond-engagement-aligning-algorithmic-recommendations-with-prosocial-goals* (2021).
[6] Hendrycks, Dan, et al. "Unsolved problems in ML safety." *arXiv preprint arXiv:2109.13916* (2021).

1. **Existing AI regulations focus on removing bias or unfairness from model outputs, rather than tracking the outcomes generated by the system itself over time.**

Existing policies to regulate the performance of machine learning systems tend to focus on the accuracy or bias of their classifications. As an example, the EU General Data Protection Regulation enshrines a “right to explanation” for algorithmic decisions.[7] Within this paradigm, “harm” is defined in terms of system errors that result in injury or damage. For instance, a self-driving car might misrecognize stop signs or lane markers, causing it to take actions that are unsafe.

These regulatory techniques are limited, because they focus on singularly impactful decision points. Modifying such points generates solutions that may work in the short term, but leaves the behavior of the system as a whole untouched. Many consequential effects of AI systems emerge over time as they change how humans make decisions or interact with each other. For example, a social media algorithm may successfully recommend content to users in ways that keep them on site, but at the cost of generating habit-forming effects that make them addicted to its feed.[8-9] Meanwhile, self-driving cars may not just learn to navigate roads safely, but also cluster in ways that limit the flow of traffic.[10] By affecting human drivers’ access to critical intersections, these fleets will impact access to public roads themselves. In both cases, the system’s operation has cumulative effects that companies often fail to track or mitigate.

Structural harms arise whenever these effects tamper with individual rights or protections, like online freedom of speech or keeping public transit accessible.[11] Those protections may not have been properly addressed in advance of the system’s deployment. Or they may be abused by the system itself, as in the above example of social media addiction. Either way, outcomes remain poorly documented.[12] Their effects are difficult for both experts to diagnose and laypeople to perceive, as they require a birds’ eye view of how the system interacts with social contexts over time.

2. **Documentation is neither standardized nor legally required for the operation of highly capable AI systems. Moreover, existing approaches fail to capture or account for the emergent effects generated by deployed AI systems.**

---

[7] Selbst, Andrew, and Julia Powles. ""Meaningful Information” and the Right to Explanation." *Conference on Fairness, Accountability and Transparency*. PMLR, 2018.
[8] See for example Allcott, Hunt, Matthew Gentzkow, and Lena Song. "Digital addiction." *American Economic Review* 112.7 (2022): 2424-63.
[9] Allcott, Hunt, et al. "The welfare effects of social media." *American Economic Review* 110.3 (2020): 629-76.
[10] See for example recent behavior exhibited by Cruise vehicles in San Francisco: https://www.cnbc.com/2022/07/01/self-driving-cars-from-gms-cruise-block-san-francisco-streets.html
[11] For more discussion of structural harms see: https://www.technologyreview.com/2022/08/09/1057171/social-media-polluting-society-moderation-alone-wont-fix-the-problem/
[12] Roberts, Sarah T. *Behind the screen*. Yale University Press, 2019.

Machine learning practitioners are now developing documentation protocols to address the structural harms that arise from system behaviors. One of these tools is the datasheet, a framework for documenting the specific dataset available to a system. These datasheets may include the features on which the dataset was labeled, the motivation for that labeling strategy, the intended use of the dataset in question, and known limitations or missing data.[13] Another important example is model cards, a framework for documenting the model that has been learned to complete some machine learning task. Model cards may include the known accuracy thresholds of the model on specific tasks such as facial recognition (e.g., error rates for men vs. women), the organization responsible for its training, metrics of interest, and known ethical considerations.[14]

Both these approaches are invaluable for understanding the capabilities of particular systems and are necessary for the robust documentation of potential harms. However they are not sufficient for tackling the emergent effects of AI systems, for two reasons.

First, they do not track how distinct components of the entire system architecture interact with each other over time. For example, a self-driving car may learn unusual swerving behaviors based on how its vision and navigation subsystems interact, impacting traffic even if it doesn't hit anything.[15] Datasheets and model cards fail to capture such effects, which may be opaque to designers as well as consumers.

Second is the lack of enforceable standards for good performance in safety-critical settings. This is pivotal in any rulemaking context, but is particularly urgent for AI systems today. Beyond the discretion of in-house machine learning teams, there are no technical barriers to deploying model classifiers that are known to be deeply flawed. In that absence, companies like Hugging Face have stepped up to serve as a one-stop shop for documentation of popular machine learning resources.[16] However, there is no equivalent for social media applications whose algorithms are proprietary and secretive. While voluntary commitments to transparency are valuable, consumer rights require more than acts of good faith from designers. What is needed is a reporting protocol for structural harms so that design priorities remain aligned with critical consumer protections.

3. **While research on the full sociotechnical effects of system harms is ongoing, their potential for more significant risks has been established.**

Self-driving cars that can alter the flow of traffic will change how roads work. Over time, social media algorithms that affect users' tastes will transform how people communicate. Improving the accuracy of a predictive model does nothing to reveal the types of structural harm

---

[13] Gebru, Timnit, et al. "Datasheets for datasets." *Communications of the ACM* 64.12 (2021): 86-92.
[14] Mitchell, Margaret, et al. "Model cards for model reporting." *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Atlanta 2019.
[15] For a theoretical explanation of this problem, see Gilbert, Thomas Krendl. "Mapping the Political Economy of Reinforcement Learning Systems: The Case of Autonomous Vehicles." *Simons Institute Newsletter* (2021).
[16] https://huggingface.co/

implicit in these disruptions. Instead, regulators need the means to anticipate and evaluate AI system outcomes before they happen.

Fortunately, technical methods can generate insights into possible outcomes that make potential harms more foreseeable. For example, some AI systems are able to learn continuous behaviors by taking multiple actions or “decisions” in sequence. Consider a YouTube algorithm that learns how users watch content differently based on its own past recommendations, and then factors that into its future ones.

To automatically choose actions in these settings, practitioners often use techniques such as *reinforcement learning* (RL) or planning. These techniques enable AI systems to learn how to take actions that change their environment according to their objectives (in contrast to simply e.g. classifying a datapoint). By optimizing for long-term outcomes, these techniques can in theory learn more complex and more desirable behaviors than single-step classifiers. For example, using RL in social media applications seems to increase long-term engagement metrics and reduce how much clickbait content is shown to users (because clickbait is bad for user retention).[17]

However, the increased power of RL also raises the stakes for potential harm. Firstly, it is often hard to understand the systems’ actions even under close scrutiny:[18] the solutions that these systems converge to might be more complex than what humans can easily reason about. For example, these techniques are what enabled AI bots to beat the world champions at Chess and Go. Still today, research continues to unpack and understand the strategies of a famous AI system for these games that was released in 2017.[19] If systems based on these approaches were to take systematically harmful actions, it could be hard to determine the nature of the failure even after the fact, let alone anticipate it.

Moreover, when functioning in human contexts, such systems will have incentives to manipulate people they interact with.[20] This is especially concerning in the context of social media,[21,22] where an increasing number of companies are using RL.[23,24] RL’s potential to shape long-term behavioral outcomes beyond short-term effects mirrors the monopolistic strategies

---

[17] Using RL on YouTube saw the largest single launch improvement in 2 years, https://www.youtube.com/watch?v=HEqQ2_1XRTs&t=83s

[18] Puiutta, Erika, and Eric Veith. "Explainable reinforcement learning: A survey." International cross-domain conference for machine learning and knowledge extraction. Springer, Cham, 2020.

[19] McGrath, Thomas, et al. "Acquisition of chess knowledge in alphazero." Proceedings of the National Academy of Sciences 119.47 (2022): e2206625119.

[20] Krueger, David et al. “Hidden Incentives for Auto-Induced Distributional Shift.” ArXiv abs/2009.09153 (2020): n. pag.

[21] Carroll, Micah et al. “Estimating and Penalizing Preference Shift in Recommender Systems.” Proceedings of the 15th ACM Conference on Recommender Systems (2021).

[22] Evans, Charles and Atoosa Kasirzadeh. “User Tampering in Reinforcement Learning Recommender Systems.” ArXiv abs/2109.04083 (2021): n. pag.

[23] Chen, Minmin, et al. "Top-k off-policy correction for a REINFORCE recommender system." Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019.

[24] Gauci, Jason, et al. "Horizon: Facebook's open source applied reinforcement learning platform." arXiv preprint arXiv:1811.00260 (2018).

used by Amazon and other Big Tech companies.[25] Given their lack of interpretability, it might be hard to systematically prove that such systems are being (or not being) manipulative.

These concerns extend also to other types of systems beyond RL.[26] For example, the ways that RL explicitly incorporates sequential decisions mirror system failures observed in electrical grids and transportation networks.[27] Meanwhile, rent setting algorithms may be artificially and systematically driving up rent prices and stifling competition.[28] Whether the system uses RL or not, sequential decision criteria may introduce preventable types of consumer harm within human domains.

At present, regulators rely on the benevolence of the companies involved to assess, monitor, and limit consumer harms. This kind of monitoring is against companies' economic interests because automated decision making techniques such as RL can lead to significant increases in revenue.[29] Companies have little incentive to leave a paper trail which could expose them to legal liability. One remaining option might be to rely on external audits, but structural harms are difficult to assess from outside the companies themselves.

4. **The FTC must codify rules that mitigate the potential safety risks to consumers from automated systems.**

Algorithmic harms remain poorly documented. Whether or not a given system generates concrete injuries, its impacts remain hard to measure and difficult to weigh against its benefit to consumers. This means that automated decision-making can generate harms that are not reasonably avoidable by consumers. Absent appropriate oversight, they must be seen as unfair practices according to criterion (2) of Section 5 of the FTC Act, and possibly also criteria (1) and (3) depending on the scale and intensity of their effects.

As leading experts on the likely impacts of advanced AI capabilities, we believe such harms are likely to become both more common and more impactful in the coming years. Recently the Digital Services Act has drawn attention to systemic risks created by large online platforms[30], and the EU AI Act expands this lens to include other kinds of AI systems.[31] While these developments are encouraging, existing standards and protocols remain stuck in a static, *ex post* accountability regime that cannot keep pace with the emerging risks.

---

[25] Khan, Lina M. "Amazon's antitrust paradox." *Yale IJ* 126 (2016): 710.
[26] Jiang, Ray, et al. "Degenerate feedback loops in recommender systems." Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019.
[27] Gilbert, Thomas Krendl, et al. "Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems." *arXiv preprint arXiv:2202.05716* (2022).
[28] https://www.propublica.org/article/yieldstar-rent-increase-realpage-rent
[29] Chen, Minmin, et al. "Top-k off-policy correction for a REINFORCE recommender system." Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019.
[30] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065&qid=1666857835014
[31] https://www.brookings.edu/research/the-eu-ai-act-will-have-global-impact-but-a-limited-brussels-effect/#footnote-2

It follows that the FTC must implement trade rules and protocols to better document automated decisions over time. These rules should reflect how specific types of sequential decisions tie back to social impacts. For example, social media harms often derive from the use of metrics like engagement that are at best unreliable proxies for real-world outcomes that are difficult to encode or directly optimize.[32,33] In transportation, improving traffic throughput may depend on how automated vehicles route along critical points of a road network.[34]

What matters is that both risks and benefits will arise from how automated systems shift prevailing social norms. These norms include accepted patterns of behavior, rules of conduct, and information flow between stakeholders.[35] It follows that new rules should also monitor how consumer data is collected and retained for use in sequential decisions. These rules should strive to reconcile the inherent capabilities of the system with established consumer protections. Such documentation would be a small burden for companies to maintain, while also providing a paper trail to hold them accountable once harms do occur.

5. **To better investigate the consumer impacts of automated decisions, the FTC should support AI documentation that tracks systems' effects over time.**

Present AI documentation techniques like Model Cards and Datasheets provide strictly *ex post* evaluation of AI components that have already been built. But to apply the FTC Act, there is a need for *regularly updated* documentation that tracks whether 1) potential consumer injuries qualify as substantial, and 2) whether or not they are outweighed by other benefits.

Combining *ex ante* and *ex post* documentation of AI systems would help diagnose distinct risks in specific use cases. If regularly updated and publicly available, it would also support cohesive evaluation of system components and comparison of designers' expectations with post-deployment performance. One such framework is Reward Reports, which track how system behaviors stem from the sequential decisions it was designed to make.[36] Reward Reports and related efforts could inform how the FTC adopts new standards for consumer harms based on Section 5.

Since the passage of the National Environmental Policy Act, federal agencies have been required to submit reports called Environmental Impact Statements (EISs) that explain the likely

---

[32] Milli, Smitha, Luca Belli, and Moritz Hardt. "From optimizing engagement to measuring value." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021.
[33] Bengani, Pri, Stray, Jonathan, and Thorburn, Luke. "What's right and what's wrong with optimizing for engagement?" https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851
[34] Wu, Cathy, Alexandre M. Bayen, and Ankur Mehta. "Stabilizing traffic with autonomous vehicles." *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.
[35] Nissenbaum, Helen. "Privacy in context." *Privacy in Context*. Stanford University Press, 2009.
[36] Gilbert, Thomas Krendl, et al. "Reward Reports for Reinforcement Learning." *arXiv preprint arXiv:2204.10817* (2022).

consequences of their proposed projects on the environment.[37] EISs are a clear model for documenting AI systems. In particular, we recommend that any federal agencies that rely on vendored AI systems submit annual Reward Reports that explain the likely consequences of the system for consumers. If Reward Reports were regularly issued in like manner to EISs, the FTC's Section 5 mandate would be much easier to enact and enforce.

Sincerely,

Thomas Krendl Gilbert
Postdoctoral Fellow, Cornell Tech

Micah Carroll
Ph.D. Candidate, UC Berkeley

[37] Caldwell, Lynton Keith. *The National Environmental Policy Act: an agenda for the future*. Indiana University Press, 1998.