BRIEF FOR REGULATORS

This document is intended to accompany *Accountability Infrastructure: How to implement limits on platform optimization to protect population health*. It is intended to offer a brief cheat sheet for implementing the protocol within a regulatory regime. For the motivations, reasoning, consideration of drawbacks, and other complications, please consult the full paper at https://www.platformaccountability.com.

The goal of this kind of plan is to ensure population health for at-risk groups exposed to a product. It is not to direct any particular technology decision or product choice. To achieve this goal, policy must enable a statistically valid series of assessments assessing the effects of exposure to a product, inspired by consumer protection rules that apply to physical goods and drug protocols. While many of those requirements are cumbersome and highly regulated, the mission of this system is to leverage existing systems and methods that most technology platforms *already* use to assess potential harms on at-risk populations. Every effort should be made to make systems flexible to use existing tools rather than creating new measurement practices – so long as those systems adhere to necessary standards reviewed by third parties.

This document, clearly, is not intended to provide legislative language. Instead, it offers a few simple frameworks for initial implementation of the protocol's mechanism design. Note: these procedures are modeled on social media systems; other products may necessitate alternative methods.

If you have feedback or would like to request support for implementing the methods described here, please contact the authors. Ongoing conversations are now in process with technologists, health professionals, and regulators, and we are seeking additional participants in these discussions.

Transparency into existing decision making processes

The first regulatory opportunity does not require any new data collection, but instead focuses on existing experimental practices. Large technology products – and specifically the ones which should be subject to this type of regulation (see below for additional description of size parameters) – already implement A/B testing and holdouts as part of their normal team-level decision making procedures. For instance, Meta has a dashboard, internally called Deltoid, which houses test results, including the various metrics assessed and how they shift based on experimental conditions.

Our first, and easiest-to-implement recommendation is to set requirements that these dashboards and test results be subject to transparency requirements. Tests that incorporate changes designed to be applicable to scales/procedures capable of societal-scale harms should be subject to this type of review. The experiments associated with system architecture of large

social media platforms, specifically including dashboards like Deltoid, would be subject to oversight.

One particularly appealing aspect of this approach is that these dashboard metrics *do not require disclosure of personally identifiable information (PII) of users*. Access to PII is a legitimate sticking point for many regulatory transparency strategies, especially those which empower third-party reviewers. But because this class of data deals with the summary results of those tests and how results intersect with internal business decisions, no PII need be transferred to reviewers, except potentially for (much more) limited auditing/validation purposes.

At the same time, companies do have legitimate business interests in protecting information in these dashboards from competitors. We would recommend some careful methods that balance these very real private interests with the public. We believe there are three ways that this type of regulatory structure could be implemented to enable access to the relevant data without putting undue burden on companies:

1. The scope of publicly-reported metrics can be limited to those which potentially implicate broader societal interests. For instance, this might relate to metrics like hate speech, quantifiable misinformation on subjects like non-partisan voting information, and engagement with direct implications for health like time spent at night among teens.

2. Direct review of internal dashboards and metrics can be limited to those who have a legitimate need, such as regulators or accredited academics. While we believe the public should have access to information, this can be limited to a further-summarized distillation of overall effects.

3. The publication of results can be released over time such that any specific test or result is outdated by the time of its release. Potentially, special considerations for non-public review by regulators might be triggered for particularly high-risk product changes, but these updates need not be public.

Ultimately, a transparency requirement requires public access at least to the metadata of these results. Under data access rules such as those proposed under the Platform Accountability and Transparency Act or via a designated agency such as the Federal Trade Commission, experimental results among the largest companies could be documented.

Size requirements

Similar to proposals such as the DSA,[1] we believe that only larger products should be subject to requirements of the profile described here. While a single binary (big enough to be subject or not subject) is actionable, we prefer a scaled set of increasing requirements designed to induce participation from companies earlier without adding undue regulatory burdens. Here is one method for considering scale changes (and recognizing that such a proposal will require a suitable definition of a "monthly user").

Number of Monthly Users	Requirement
1 million	Submitted plan for metrics and methods for evaluation of potential structural harms
10 million	Consistent data collection on potential structural harms
50 million	Quarterly, enforceable assessments on product aggregate effects on structural harms, with breakouts for key subgroups
100 million	Monthly, enforceable assessments on product aggregate effects as well as targeted assessments of specific product rollouts for any subproduct used by at least 50 million users, with breakouts for key subgroups

Requirements for product-wide holdouts for health metrics

While a range of potential metrics might be considered for regulation, we believe the initial focus should remain squarely on mental health effects among at-risk populations, especially teenagers. (As a secondary focus, we suggest the establishment of metrics associated with societal trust – but because these are less timely or consensus, for regulatory purposes we believe the immediate focus should remain on mental health.)

The first regulatory step is to identify a body of experts suitable for establishing a set of benchmark mental health metrics based in existing procedures outside of technology, or to establish those benchmarks directly. We include a few existing benchmarks here as starting points based on a review of the literature and conversations with experts, but ideally further ideation and iteration will occur under a regulatory process to refine these assessment standards in consultation with a range of experts (including technologists in companies to help assess feasibility).

For mental health, we believe surveys are the most applicable method. For general wellbeing, the standard eight-question patient health questionnaire, such as the one from Children's

[1] https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Hospital, offers a simple tool that is widely used.[2] An alternative two-question version provides an even simpler set of questions:[3]

- Over the last 2 weeks, how often have you been bothered by having little interest or pleasure in doing things?
- Over the last 2 weeks, how often have you been bothered by Feeling down, depressed or hopeless?

For teenagers, the SCOFF benchmarks offer options specific to eating disorders, which may be correlated with a larger set of considerations.[4] We also suggest considering the use of the Photographic Affect Meter as a supplementary measurement system.[5]

Once reviewed and established, regulation will require the consistent measurement of these metrics among representative populations both of the broader userbase, and among teenagers specifically. What do we mean by “consistent tracking”? We mean, specifically, the establishment (at minimum) of a randomized trial using a holdout group of users whose product experience is held static over time, and a treatment group of statistically equivalent users who see the most up-to-date “standard” version of the product, enabling direct comparisons of the users based on the metrics just described.

By establishing a quarterly requirement that both the treatment group and control group be assessed under the established metrics, regulators can assess whether measurable harms are being caused. We leave to regulators to decide whether transparency is sufficient in the case harms are established or whether other incentives are necessary.

If you would like to request support for implementing the methods described here, please contact the authors. Ongoing conversations are now in process with technologists, health professionals, and regulators, and we are seeking additional participants in these discussions or can connect prior participants with interested parties.

Feature-specific evaluations

For larger companies, especially those which already run randomized assessments of the types described here, additional requirements can be established based on product components that have scaled consumer usage. For instance, if a component of a recommendation system reaches 100 million users monthly and is tested independently, regulators can insist that existing testing protocols also include metrics associated with health.

By setting these standards based on assessment methods and outputs, rather than specific procedures for how randomized trials must be implemented, regulators can leverage the

[2] https://www.childrenshospital.org/sites/default/files/2022-03/PHQ-8.pdf
[3] https://www.hiv.uw.edu/page/mental-health-screening/phq-2
[4] https://onlinelibrary.wiley.com/doi/abs/10.1002/eat.20679
[5] https://dl.acm.org/doi/10.1145/1978942.1979047

existing tools and methods that companies build rather than insist on independent forms of measurement that would lead to an undue burden.

Evaluation and transparency requirements

For any of the three experimental processes described here – review of existing A/B tests, evaluation of product-wide holdouts, or evaluation of specific major features – review of the results will be necessary by an accredited set of third party reviewers. We believe there are a number of ways such a system could be legitimately implemented, and the most comprehensive current one is under the Platform Accountability and Transparency Act, which enables accredited researchers to be certified under the National Science Foundation.[6] Working with an adequate government regulator, such as the Federal Trade Commission or the National Institute for Health.

Regulators can limit, at least initially, requirements to validated collection of health data, on the assumption that should products demonstrably reduce metrics to statistically-validated levels, platforms will make changes prior to distribution of results publicly. If such an inducement is insufficient, regulators can set additional punishments for particular levels of harm beginning a set period, say one year, after the launch of the regime.

While it will be necessary for the experts identified earlier to have access to the results of experiments identified under this system, protections must be afforded for companies to protect their legitimate business interests, including trademarks and trade secrets. Steps can be codified to ensure this occurs. We offer two specific methods (besides rules around access tied to accreditation) that can be written into any rules:

1. Required reports can be limited to metrics associated with societal interest (e.g. health metrics identified earlier) and parameters predictive of those metrics, but not underlying parameters themselves.
2. Publication of results can occur only after a delay (e.g. two years) to ensure that specific tests and protocols are outdated for competitors.

[6] https://www.congress.gov/bill/117th-congress/senate-bill/5339/text