BRIEF FOR PRODUCT MANAGERS & DESIGNERS

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

We believe four steps are necessary to implement an assessment protocol suitable for measuring structural harms, such as the mental health effects in teenagers. These steps do not necessarily need to follow this sequence, nor are they equally important in all contexts. For small products that do not have the necessary scale to implement these procedures, we believe there is still value in planning for future considerations, and anticipating the architecture that might be needed to enable measurements of harms. Note: these procedures are modeled on social media systems; other products (e.g. generative AI tools) may necessitate alternative methods to implement such a protocol.

If you have feedback or would like to contribute alternative methods to this project, please contact the authors. Ongoing conversations are now in process with technologists and health professionals, and we are seeking additional participants in these discussions.

Step 1: Define potential harms and select metrics

We recognize that product leaders may or may not have discretion over evaluation of their own processes. Nonetheless, the agenda of this workstream is evaluation of metrics often outside of normal workstreams – and so even if health metrics are not reflected in organizational objectives, establishing representative assessments of population health metrics alongside business metrics is a critical first step in evaluating whether there even is an issue which compels attention.

Platforms may choose to orient evaluation around different types of structural harms. While we encourage thoughtful, product-specific considerations of this question, we also believe there is value in default options. Specifically, we offer two classes of potential harm that may be applicable to many different use cases, with a particular emphasis on the first: (1) avoiding reductions in standard measures of adolescent mental health; and (2) avoiding reductions in standard measures of social trust.

We recommend these two choices primarily for two reasons. First, there is widespread consensus that reductions in either of these metrics is *prima facie* harmful for human populations regardless of most other considerations. And second, there are no obviously contentious social valences to the outcomes of these goals – a happier, more socially cohesive society is one that is preferable under most any democratic political ideology.

For any choice of potential harms to evaluate, a regime is only as effective as its metrics. While there are numerous legitimate choices available, we highlight a few specific options that are often used based on experts consulted for this project.

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 Hospital, offers a simple tool that is widely used.[1] An alternative two-question version provides an even simpler set of questions:[2]

- 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.[3] We also suggest considering the use of the Photographic Affect Meter as a supplementary measurement system.[4]

For social trust, surveys similarly are the standard measure of evaluation. Pew Research has asked a short battery of questions for years, which are quite similar to the academic literature.

These include:[5]

- Generally speaking, would say that [most people can be trusted / most people can't be trusted]
- Do you think most people [would try to take advantage of you if they got a chance / would try to be fair no matter what]
- Would you say that most of the time people [try to help others / just look out for themselves]

Step 2: Implement metrics within existing product testing processes, or via product holdouts

Product teams deploy different protocols for evaluation of progress and for reporting and documentation. Our starting position is to say: if there are existing procedures used for business metrics, if at all possible incorporate the metrics from the preceding step into those existing practices. Our goal with this project is not to reinvent the wheel.

[1] https://www.childrenshospital.org/sites/default/files/2022-03/PHQ-8.pdf
[2] https://www.hiv.uw.edu/page/mental-health-screening/phq-2
[3] https://onlinelibrary.wiley.com/doi/abs/10.1002/eat.20679
[4] https://dl.acm.org/doi/10.1145/1978942.1979047
[5] https://www.pewresearch.org/politics/2019/07/22/the-state-of-personal-trust/

To the extent those systems either do not exist or cannot be applied, there are several potential options, likely contingent on the scale of the product and resources of the product team. In all cases, our goal is for a statistically-valid randomized controlled trial operating at a properly-powered scale.

Likely the simplest mechanism in the event there is not a preexisting system is a generalized holdout, with a goal of assessing all changes made to a product (in aggregate) over a defined period of time. By default, we believe a three month window is a reasonable selection.

Under such a regime, a defined audience of users will receive no product changes over the evaluation window; at the end of that period, a sample of that pool of users will be assessed using the metrics in Step 1 to a statistically equivalent pool of users who receive all product changes over the three month period.

Alternatively, if the product is sufficiently scaled and it is technologically feasible, incorporating the metrics from Step 1 can be applied more narrowly to specific tests associated with particular feature rollouts.

If interested in technical assistance, we can put stakeholders in touch with former employees of major technology platforms who were responsible for implementing A/B testing protocols of this format for product optimization.

Step 3: Define documentation and mitigation procedures

Consistent documentation procedures are another necessary step, if they do not exist already. If there is already a mechanism (like a dashboard for experimental reports), harm metrics can be built in directly; if there's not, a simple tracker will be needed. The goal in this reporting should be to help subsequent readers understand not just the outputs of the data collected, but the decision tradeoffs considered in choosing what product features to implement and which to reject. Precisely because causal mechanisms for the outcomes of experiments may not be understood, specific explanations are not a prerequisite for documentation unless experimentation is specifically designed to respond to potential hypotheses.

A few existing methodologies exist to support this work. These include:

- Model Cards[6]
- Datasheets[7]
- HELM[8]
- Reward Reports[9])

[6] Mitchell, Margaret, et al. "Model cards for model reporting." *Proceedings of the conference on fairness, accountability, and transparency*. 2019.
[7] Gebru, Timnit, et al. "Datasheets for datasets." *Communications of the ACM* 64.12 (2021): 86-92.
[8] Liang, Percy, et al. "Holistic evaluation of language models." *arXiv preprint arXiv:2211.09110* (2022).
[9] Gilbert, Thomas Krendl, et al. "Reward reports for reinforcement learning." *arXiv preprint arXiv:2204.10817* (2022).

The forms listed above respectively deal with machine learning models, properties of the data, benchmarks for comparing different large language models (LLMs), and the intended goal of particular AI systems. The precise form of this reporting is less important than that it be consistent and aligned with accountability mechanisms within a given internal organization.

Step 4: Establish a cadence for feedback and system updates

Infrastructure of the type described here is not meant to be static. While updating metrics and systems too frequently may lead to incomparable results, organizations may choose to make changes at a reasonable period (perhaps every six months or yearly). A growing community of practice is interested in these metrics and may be available for additional external support.

Over the longer term, new evaluation areas with new metrics may be monitored, or improved measurement systems may be designed. Product managers may be in a position to compare possible interventions against each other enabling product changes to be more actively justified and defended. When product managers have discretion over what implementation would best enact specific design priorities, the wider interests of both the company and populations using the platform will be better served. We expect these evaluations would be integrated into general announcements and comprise a new dimension of internal, company-wide accountability.