Comments of the R Street Institute
Policy Advisory Opinion 2021-02–Oversight Board

The R Street Institute respectfully submits these comments in response to the request for public comment issued by the Oversight Board in connection with its consideration of a request by Meta for a policy advisory regarding the "cross-check" system used to help detect some "false positive" content takedowns in certain circumstances. R Street recognizes the value, and arguably the necessity, of a user-centric system such as Meta's cross-check to supplement the normal operation of automated content review mechanisms. However, as detailed below, such a system must provide far more transparency, and on the basis of such transparency should be reviewed and improved over time with the benefit of input from external perspectives to ensure that its output is balanced and on net positive for the entire ecosystem.

The *sine qua non* of the cross-check system is Facebook's extensive use of automation for content moderation at scale. One of the richest propositions discussed in the development of R Street's multi-stakeholder content policy project was the use of automation to detect potential policy violations in real-time. At the scale of Facebook, automation is essential for effective content moderation and can serve to greatly mitigate harm through techniques such as "virality circuit breakers" driven by metadata associated with online activity. However, as our project participants were quick to note, automation struggles with context, which often requires a very human understanding of evolving offline social cultures and structures. Furthermore, automation can have disparate impacts, with racial bias in hate speech detection as a known example cited in our final report.

With these inconsistencies in mind, a properly designed system such as cross-check, can layer in a degree of pre-programmed review that is tailored to known challenges for automated systems. Substantial external input into the design and operation of such a system, and adequate disclosure of its impact in practice, could greatly improve the issue of online harm and increase trust and confidence in the entire social network.

However, that has not been the case thus far with Meta's cross-check system. R Street appreciates the Oversight Board's intent to offer constructive guidance regarding improvements to the system. Many such improvements will focus on the mechanics of the system, such as the factors used in its review. Creativity is possible in these considerations above and beyond the current state; for example, a geographic content-based scope could be used to trigger certain content in "hot" geopolitical regions regardless of the identity of the author. Because of the wide-ranging scope of creative possibilities, which will likely be the topic of other submissions to the Board, R Street will focus the remainder of this submission on structural considerations around the system, such as how it is evaluated and used.

With regard to the third issue identified for comment by the Oversight Board, R Street submits for consideration that perfect neutrality and a full removal of bias are, in practice, patently impossible. Thus, an explicit focus on *balance* rather than *bias* may be more productive in the Board's consideration of the design of the cross-check process and its implementation and outcomes. The cross-check system today is essentially a list of potentially sensitive users, together with a few articulated criteria concerning why some users are placed on the list and others are not. A focus on *balance* (in contrast to bias) would place more importance on the list, rather than the criteria used to form it. The list of users who receive the benefits of the cross-check system must be balanced and reflect Facebook's community; if it is not, no strength of logic justifying the criteria for selection will save the system in the public eye.

Of course, to be able to gauge whether the cross-check system is balanced in its protections, transparency regarding beneficiaries is a necessity. As noted in the eighth issue identified for comment, transparency into the operations of the cross-check system is the richest opportunity for Meta to invest in to increase system efficacy and public trust. To provide greatest benefit, the cross-check system should make clear to a user whether they benefit from the review and should provide a similar signal to other users of the system. Furthermore, transparency into the rationale for such classification would encourage a public dialogue around, and diverse input into, the cross-check system to improve it over time. External input is the most valuable information available to resist tendencies towards institutional myopia that can otherwise develop.

Returning to the reality of today, the Oversight Board has indicated clearly that sufficient transparency has not been provided to the public, nor to the Board itself. This leaves the Board ill equipped to tackle its significant responsibilities, while increasing public doubt over its legitimacy as an institution. While it is essential for the future of the Oversight Board that such gaps do not appear ever again in the future between Meta and the Board, it is equally essential for the public's currently-fragile trust in Meta—and therefore essential for the success of the company's current and future services—that more transparency be provided broadly, not just to the outside Oversight Board, which continues to be seen by some as closely associated with the company itself.

Content moderation is among the most difficult matters faced by internet companies in the modern era. One central element of that difficulty is very much on display here. As the Oversight Board's call for comment notes, Meta "still has difficulties striking a balance" in practice between the competing values of freedom of expression and its own content policies designed to limit online harm and abuse of the system. That observation reveals that the cross-check system is about more than simply addressing “false positive” content policy decisions. It allows Meta to use automated moderation to maximize the implementation of its content policies, while creating opportunity for manual intervention that enables them to walk back content policy review decisions where the net outcome of a takedown decision would be harmful to the company. In other words, it helps Meta continue to advocate zealously for both freedom of expression *and* the full implementation of its content policies by masking the inherent ambiguity of content policies for which there will always be instances that defy objective policy application as well as tension whenever one person’s “free expression” is another person’s real harm. While Meta is well within its rights to adopt such a “have your cake and eat it too” posture—just as it should be expected to take the steps necessary to protect itself as a company—the societal richness of Facebook and Meta’s other services and the massive socioeconomic consequences of failure call for a different approach.

Certainly, this dynamic extends beyond the scope of the cross-check system itself. However, it is a factor worthy of consideration as part of why such a system exists in the first place, as well as an illustration of an underlying normative balance and tension whose public articulation and open discussion will help build trust generally. Brutal honesty that content policies are human and subjective and therefore impossible to apply with mathematical or algorithmic precision, and a posture (and reality) of listening to public input into the design and use of fundamentally imperfect but necessary technical systems like cross-check, will go far to help address the underlying challenges on display here.

Finally, the cross-check system, like other internal systems at dynamic companies like Meta, will almost certainly evolve over time, which is a good thing. As it evolves, the company will undoubtedly do A/B testing of various options for its evolution. Transparency into such experiments and their outcomes, including the metrics used to evaluate the tests, would be highly beneficial. Not all situations will require public transparency, but many would benefit from high level transparency to the public with more details available to independent researchers and relevant regulators. Such transparency would not only improve trust, but would also facilitate the collection of perspective and information from outside the company, which could substantively improve the system itself.

Sincerely,

Chris Riley
Senior Fellow, Internet Governance
R Street Institute