Comments of the R Street Institute
Case No. 2021-001-FB-FBR—Facebook Oversight Board

The R Street Institute respectfully submits these comments in response to the request for public comment issued by the Facebook Oversight Board in connection with its consideration of the decision by Facebook pertaining to Donald J. Trump referenced in Case No. 2021-001-FB-FBR. R Street is of the view that: a) context, and therefore case-by-case adjudication is critical to a valid and appropriate adjudication of the issues presented; b) that an operationalized framework is necessary to particularize more finely the considerations applicable to content moderation decisions; and c) that the same framework applicable to private citizens ought to be applied to political actors, with due regard for the different context in which their expression arises.

There is no universally accepted content moderation framework guiding Facebook's practices or the Oversight Board's subsequent analysis. Facebook is a private corporation, not a government, and is free to adopt any legal content moderation practices that suit its community and culture. The Oversight Board is chartered to "review content enforcement decisions and determine whether they were consistent with Facebook's content policies and values," and "will pay particular attention to the impact of removing content in light of human rights norms protecting free expression." In practice, Facebook and the Oversight Board have made, perhaps implicitly, a determination to apply principles of international human rights law (IHRL). R Street concurs that IHRL is a reasonable source for the Board's decision-making and its review of Facebook's actions. We believe, however, that most IHRL is stated at too-high-a-level of generality to be of significant practical usefulness to the Board and inadequately specific to provide guidance to Facebook and notice to users as to how content decisions will be made.

Framework Factors -- We, therefore, offer a framework for suggested analysis that identifies neutral principles and factors on which we believe the Board should rely for its decision-making. In doing so, we are cognizant of the fact that multifactor tests are, themselves, somewhat ambiguous and even sometimes subject to manipulation. Nevertheless, we believe that experience in both domestic and international legal systems over the past 50 years has demonstrated the utility of explicitly identifying relevant factors and issues for consideration in rendering judgment. The exposition of factors and their application to various fact-based scenarios will allow the development of, in effect, a common law of content moderation and, over time, provide greater transparency and clarity in evaluating Facebook's actions and the Board's responses:

- Truth or Falsity – The falsity, vel non, of a social media post is relevant to the balance between harm and freedom of expression; for example, Holocaust denialism and vaccine disinformation are exemplars of exceptionally damaging falsehoods. Of particular concern should be purposeful coordinated disinformation. Discussions of sensitive topics can add positive value, but the purposeful spread of disinformation adds a particularized harm.
- Harmfulness – Not all false speech is harmful (e.g., Grimm's Fairy Tales); nor is all truth harmless. Whether or not particular speech raises significant questions of harm requires assessment on a case-by-case basis. The relevant analysis would begin with the real-world context in which the content arises, and take into account whether real-world physical injury

(e.g., riots or child sex trafficking) is likely to occur in addition to or instead of other kinds of harm. Also relevant is the size, scale, audience and scope of the content's reach: Larger megaphones with bigger audiences are both more influential when they speak, and also more dangerous. There is a spectrum from an average citizen to a blue-check influencer to Elon Musk to Jair Bolsonaro. Political figures have outsized megaphones and thus rank high for potential harm; we do not believe they ought to have any specialized license to cause harm.

- Imminence – Harms that might materialize over a longer time frame are less susceptible to restriction than content directed at imminent events. International law allows greater action when harms are imminent and may permit pre-emptive action. This is especially salient when actual harm is ongoing.
- Incitement – We list this separately because, unlike the harmfulness section, which looks at the context within which the speech occurs, this inquiry looks at what the content provider actually intended. While the Board may wish to look at broader international restrictions on violent speech, at a *minimum*, content which is "directed to inciting or producing imminent lawless action" and is "likely to incite or produce such action" should be more readily subject to restriction. We acknowledge that, in applying this factor, the Board will ultimately face the hard challenge of the incitement of violence against illegitimate, repressive governments.
- Appropriateness of Sanctions – Having determined whether or not particular content is problematic, the Board should next evaluate the appropriateness of the sanction/punishment imposed. The following factors are relevant:
  - Deterrence – Is a sanction necessary to stop others from doing similar acts;
  - Disablement – Is a sanction necessary to stop the actor from repeating similar acts;
  - Contrition – The extent to which the actor acknowledges the nature of their prior acts;
  - Availability of Effective Alternative Sanctions – Can the harms be mitigated through interim measures (e.g. partial deletion/public correction/gating of distribution); and
  - Proportionality – Does the sanction fit the act.

President Trump – While the question before the Board pertains to prohibiting Trump from future posting, his final posts serve as background for our analysis. Based on the foregoing factors, Facebook's decision to remove those posts and prohibit former President Trump from future posting was well justified. His posts about the election were demonstrably false and, in the real-world context of heightened political tension, especially inflammatory and harmful. Not only was violence imminent, but it was ongoing at the time the content was posted and, in context, could reasonably be read as an incitement to further violence. Particularly in the context of violent acts intended to disrupt the lawful transition of government authority, Trump's content was an incitement to lawless action. Finally, Facebook had every reason to determine that alternative sanctions short of prohibiting future posts would not mitigate present or future harm.

R Street believes that it is reasonable to justify indefinite suspension by determining that Trump has a continuing political and public role and will continue to post content worth removing under these same factors. It seems clear to us that he has expressed no contrition and that the deterrence of other like-minded actors would be a positive benefit. However, this is ultimately a predictive judgment; restoring his posting privileges is defensible under different predictions. Even then, we would encourage any future restoration to include strict warnings regarding further harmful activity, with permanent suspension as the final escalatory step.