Robert Gorwa
WZB Berlin Social Science Center
robert.gorwa@wzb.eu
Michael Veale
University College London
m.veale@ucl.ac.uk

Dear NTIA staff,

We appreciate the invitation from your colleague Christopher Quarles to submit a short comment to this interesting and important RFC. We are two interdisciplinary technology regulation academics (one based in the UK, and one in Germany) that have both been working in some capacity on issues relating to the topic at hand for many years. (MV has in particular been publishing on a wide range of issues pertaining to the regulation of automated decisionmaking and AI systems in Europe and beyond; RG's work focuses on the formal and informal regulation of/by platform services in a comparative context.)

We would like to provide a focused set of suggestions based upon our peer-reviewed academic article, 'Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries.' The paper has been openly available online since November 2023 and will be published in summer 2024 in the legal journal *Law, Innovation and Technology*.[1]

That article is relevant to a few of the questions in the RFC, and is in particularly applicable as it relates to regulatory questions relating to the forms of public or semi-public model access that are mediated via online model hosting intermediaries. A few brief reflections based upon our research follow.

1. How should one consider 'openness' and access when relating to foundation models and model weights? (RFC 1c-i, 7a)

AI systems are increasingly distributed and accessed through hosting intermediaries that we call model marketplaces. These platforms are not the only channel through which distribution occurs, but just as GitHub is deeply embedded in software deployment and development supply chains, platforms like Hugging Face are quickly taking a similar role in machine learning supply chains, where public models are fine-tuned and then deployed on generic cloud hardware in numerous potential applications and contexts.

Given the increasing use of these model marketplaces (and the specific services and APIs that these platforms have developed) in research and organizational practice, one approach to consider 'openness' would be whether the model (and/or weights or other associated information) are easy to download or use via such a marketplace. This kind of definition has the advantage of considering the actual practices of access, rather than more theoretical, potentially arbitrary thresholds based on the *probability of access* in the future by an uncertain number of users or organizational entities.

Model marketplaces are additionally helpful to look at because our research has shown that these platforms already provide some kind of restrictions on access, introducing friction into their interfaces that can and do mediate the 'openness' and availability of different models. In Gorwa and Veale (2024, p. 36), we illustrate that, in response to issues of safety, content moderation, and other concerns, model marketplaces have put in place a spectrum of increasing friction and thus decreasing availability of models depending on (a) developer preferences and (b) some slowly developing content policies. These access options generally involve models that are:

- fully public (download available to anyone via model hosting intermediary)
- behind a low verification wall (download available to users with platform login credentials, or must only provide name and basic contact info, which is not verified)
- require full verification (downloaders must provide phone number or identity in some way)
- whitelisted (download only permitted for vetted users/organizations)

These real-world categories could provide some proxy for model openness, distinguishing between more established 'open for researchers in the community' models of access (akin to model weights/config files/other community knowl-

[1] Robert Gorwa and Michael Veale, 2024. 'Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries.' *Law, Innovation and Technology* 16 (2). Available at https://doi.org/10.31235/osf.io/6dfk3. Hereto referred to Gorwa and Veale (2024).

edge circulating via mailing lists or other finite specialist groups) and emerging 'totally public for experimentation and deployment by any individual anywhere' models of access.

Naturally, these types of access conditions are not perfect in terms of actually limiting access to models. The access approaches deployed thus far do not fundamentally seek to 'secure' a model, in the sense that some secondary dissemination is likely be expected, but these access conditions can probably be said to determine access for the majority of users that will be integrating a foundation model in an organizational setting. (Alongside low-level leakage, with e.g. users downloading models locally and sharing them via hardcopy or other means, a subset of certain actors may intentionally seeking to evade regulation or governance, such as via emerging peer-to-peer AI model marketplaces that we discuss in our paper, or potentially via third party software repositories, cloud services, or even torrent sites.)

Whitelists are the most restrictive forms of friction we find in-the-wild, and emerging coordination mechanisms are appearing between platforms and developers, such as the APIs that Hugging Face could provide for third-party whitelists, first used for the Facebook LLaMA2 series of models. Such whitelists purport to allow the third party developers to assure themselves of features of downloaders, such as their identity, or whether they have signed a particular contract. We might envisage that in the future watermarks could be applied to models downloaded via major marketplaces to allow their provenance to be traced downstream, although this is not a current practice and research would have to examine the durability of any such watermarks to tampering or removal.

Relatedly, we note in our article that the emerging norm of building in third-party agreements between developers and deployers to access a model via a model marketplace, including via whitelist contracts and IP-based conditional licenses, can typically only be enforced by the developers themselves (or their agents), unless a sophisticated contract-based system of enforcement is created. As a result, we expect that the outcome of these types of mechanisms will depend heavily on the enforcement resources and commercial incentives leading developers (as is the case with the enforcement of, for example, copyleft licenses). Absent this significant investment, licensing and contracts, as currently deployed, are unlikely achieve the desired substantive governance impacts (i.e., limiting problematic and/or dangerous implementations of certain dual-use models) at scale.

2. What should be the role of model marketplaces, such as GitHub or Hugging Face, in helping foster a responsible governance ecosystem for powerful and potentially high-risk general purpose models? (RFC 7e)

Model marketplaces are important emerging choke-points in open-source AI ecosystem. These platforms make models visible to the public and to the research community (via recommendation and discovery features), make their use (and modification/tuning) easier, and lower barriers to deployment of powerful models for both organizations and individuals (e.g. via 'one click deployment' features and business models that allow customers to easily deploy models on the model marketplaces' own infrastructure). Most of these platforms have begun to slowly acknowledge their responsibilities, and as a result, increasingly engage in both proactive and reactive interventions intended to govern the dissemination and use of the models they host (Gorwa and Veale 2024, p. 19-35). As we outline in our paper, platforms are already deploying some proactive measures as a form of self-governance (which they do due to norms in the machine learning research community, due to informal pressure from governance stakeholders in civil society or government, or, increasingly, because they may arguably be required to take this action to comply with some legal regimes, such as the UK's Online Safety Act or the EU's Digital Services Act).

Additionally, our paper explores how model marketplaces are already making reactive governance decisions as they seek to deal with formalized legal complaints and take-down requests. In the United States, these platforms will surely soon be faced with Digital Millenium Copyright Act requests, both due to issues of copyright entangled with models, but more saliently for our purposes here, relating to breaches of behavioural use licenses such as OpenRAIL. The format of a request to remove a license in breach of a behavioural use license under US law is the DMCA, which requires an expeditious response from a platform once they are made aware of infringing content on their platform. In the European Union, a similar standard applies under the Digital Services Act (although it applies beyond copyright

issues).

The responsible governance here will hinge on how these requests are responded to. In Gorwa and Veale (2024, p. 13-19), we categorise the issues that a takedown might relate to as either *intended*, *realized,* or *potential* forms of misuse.

Intended uses are relatively easy to discern (e.g. from documentation) and given evidence of clearly problematic safety impacts, a takedown can proceed quite clearly. For example, if a model is explicitly trained in a way that is illegal (and advertises itself as such on a public repository, via a model card, or other mechanism) then we expect good-faith platforms to act accordingly. Nevertheless, in a more subtle case, where for example a developer tunes another model whose licenses forbids its use in medical diagnostics, and the new model explicitly announces its purpose for medical diagnosis, platforms like Hugging Face have used prospective license breaches as justification for a takedown.

Realised or potential uses are the most challenging, however, because they should be substantively evidenced. To continue our example above, image classification models which do not explicitly state that they are to be used for clinical diagnosis, but in practice are widely used in that way, may also breach e.g. license terms (or other laws, like federal criminal law outside the scope of Section 230, insofar as the model marketplace may be disseminating a device that should have FDA approval). In other words, models which can be used for diagnosis and could give very dangerous results might also be prohibited, but naturally these types of safety decisions will be much more difficult for emerging model marketplace trust and safety teams to deal with. Similar concerns apply to foundation models that may exhibit downstream patterns of bias and discrimination, or could be used to produce synthetic child abuse imagery.

The question for a responsible model marketplace is how to respond to the kinds of model takedown requests that are already being directed towards platforms like Hugging Face and GitHub.[2] How much detail is required? How much due diligence should platforms undertake in checking that the request is technically true (for example, if a stakeholder complaint alleges that the model can or does produce child sexual abuse imagery). If a takedown request is too pithy and short, it may take a research team to verify; the same may be true if it is huge and methodologically complex. As we argue in our article, this is an extremely high-stakes policy area with distinct dynamics that distinguish it from traditional content moderation or platform regulation.

3. What can be done going forward? (RFC 9)

A responsible model marketplace, and its surrounding stakeholders, need to set standards for how to respond to requests, both within and around their jurisdictions. They will have to emerge in any case, so we may as well be proactive and steer them as they do. We have suggested for standardized 'evidence packs for model flagging' in our past work (Gorwa and Veale 2024, p. 42-45). The idea here would be to develop an evidentiary template that a platform would need to be presented with in order to robustly consider a takedown request. We believe this is work that the NTIA could usefully play a role in and would be happy to discuss our ongoing research in this area further.

Due to the complexity of the governance ecosystem here, and the inherently difficult issues posed by the responsible regulation of dual-use technologies, we also believe that there will be a structural requirements for actors involved to externalise analytic capacity away from model marketplaces as exclusive 'arbiters of safety.' There are some parallels here to the copyright context: GitHub has in the past realized that it is extremely onerous to give fair and detailed technical and legal consideration to copyright holder takedown requests pertaining to digital rights management circumvention software under the DMCA (in particular, as regards to potential fair use arguements), and therefore provided a fund that could be used for developers to fund a counter notice which might keep their material online. Hugging Face, GitHub, and other emerging model marketplaces with various business models and specialties may

[2]Hugging Face has partially implemented some degree of GitHub's practice of publishing repository removal requests. To our knowledge as of March 2024, these requests appear to have been only targeted at tuned models uploaded by individuals, rather than well-known and visible foundation models shared by major organizations and research labs.

similarly struggle to resource the kind of detailed analysis that is required to judge both proactive and reactive moderation and response in relation to complex issues of model safety. As a result, stakeholders across the supply chain need to actively consider who will do this analysis in a routine way, and how they will cooperate across the supply chain. Should we engage external bodies, NGOs, research groups, platforms themselves, alternative dispute resolution bodies with technical staff, expert public sector regulators, developers fighting against takedowns — or all of the above?

We do not believe that these complex, open-ended issues can be left only to internal Trust and Safety teams inside leading model marketplace firms. In conclusion, while this is a fast-moving and highly technical policy area, we think that the following are research and policy priorities moving forward:

- developing evidence packs for model flagging, in a multistakeholder way if needed, to allow model marketplaces to have expected forms of evidence which allow them to be able to proportionately examine a model or system without taking on too much of the technical or legal analysis themselves
- engaging in institutional conversations as to how analytic capacity could be best externalised across the policy ecosystem to support model marketplace platforms in making fair and informed judgments about proactive and reactive moderation
- reconsidering the license arrangements and governance by licenses, particularly in light of the enforcement gaps that they present
- funding further study of model marketplaces and their safety features, and better assessing the development and deployment of automated tools to support content moderation research and enforcement in this area

We hope that these brief comments are useful to NTIA's ensuing rule-making process, and we are at your disposal for future conversations and questions.

Dr Robert Gorwa
Postdoctoral Research Fellow
WZB Berlin Social Science Center, Germany

Dr Michael Veale
Associate Professor in Digital Rights and Regulation
Faculty of Laws, University College London, United Kingdom