CCC
Computing Community Consortium
Catalyst

CCC's Response to the Request for Comments on the U.S. Artificial Intelligence Safety Institute's Draft Document: Managing Misuse Risk for Dual-Use Foundation Models

This response is prepared by the Computing Research Association (CRA)'s Computing Community Consortium (CCC). CRA is an association of over 270 North American computing research organizations, both academic and industrial, and partners from six professional computing societies.

The mission of the CCC, a subcommittee of CRA, is to enable the pursuit of innovative, high-impact computing research that aligns with pressing national and global challenges. Please note any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the authors' affiliations, or of the National Science Foundation, which funds the CCC.

September 9th, 2024

Written by: *David Danks (University of California, San Diego), Catherine Gill (Computing Community Consortium), Ming Lin (University of Maryland), Daniel Lopresti (Lehigh University), Manish Parashar (University of Utah), William Regli (University of Maryland), and Matthew Turk (Toyota Technological Institute at Chicago).*

1. What practical challenges exist to meeting the objectives outlined in the guidance?

The greatest challenge we see with monitoring and controlling dual-use foundation models is that once they are accessible to the broad population (i.e., open-source models), they can no longer be controlled by their creators. Users can interact freely with the models, potentially removing safeguards put into place by the creators to exploit their capabilities. Monitoring is mainly useful insofar as the creators have control over the use of the models. Creators can attempt to track the use of these models, however, they will likely not be aware of all uses by third-party actors if the models are released to the public.

In addition, "malicious actors" are referred to repeatedly throughout the draft, however identifying characteristics for what constitutes a "malicious" actor are not defined, leaving the term vague. Furthermore, focusing solely on malicious actors leaves several dangerous threats unaddressed. For example, negligent, uniformed, or improperly trained actors are not necessarily malicious, but can potentially cause significant damage. If the document seeks to only focus on outcomes from malicious actors, we believe that the specific characteristics of these actors should be more clearly defined within the document. However, we believe this document should also account for negligent and uninformed actors that have access to these models.

In addition, while the dual-use foundation models may themselves not be used to create undesirable outcomes like biological weapons, they may help malicious actors move more quickly from novice to skilled or expert level in developing weapons or models that create dangerous materials. Malicious actors may also use these models to exacerbate the damage caused by weapons of mass destruction. For example, a bad actor may ask a model for guidance on maximizing casualties in a given attack or methods for causing confusion and mass panic. Beyond preparing for objectionable creations by these models, NIST should also prepare for these models to be used as dangerous instructors that cannot discriminate between well-intentioned and malicious users.

We were also unclear about which stakeholders are in charge of which “Practices” mentioned in the drafting document. We realize that the document’s focus is on “initial developers,” but this term is ambiguous given the many decision-makers involved with creating a releasable model. Clear guidelines on responsible parties for evaluation, metrics gathering, and responses would benefit efforts to combat malicious misuse of these dual-use foundation models.

Along with this concern, the draft does not take into account the possible outcomes of combining multiple dual-use foundation models. A model that is deemed “safe” when used in isolation could create significant risks when used in combination with other “safe” models. There has been limited research into the potential outcomes of combining these models, so the possible consequences are not well understood. These unpredictable outcomes may result in models overriding safeguards or potentially producing other, completely unexpected dangerous outcomes.

2. How can the guidance better address the ways in which misuse risks differ based on deployment (*e.g.,* how a foundation model is released) and modality (text, image, audio, multimodal, and others)?

We believe the method of deployment is critical to the safe use of these models. For models with dangerous capabilities, such as those listed in the draft, these models should not be deployed as open-access models. Guidance for acceptable risk levels and use cases should be established and widely adopted by creators to ensure dangerous models are not deployed as open source. Organizations, such as the Partnership on AI, offer guidance for different types of models (i.e. specialized narrow models, general purpose, and paradigm-shifting/frontier models) and offer guidance on deployment methods for these different types of models. While we don't necessarily suggest that this guidance in particular should be adopted widely, we suggest that the AISI draft include guidance on acceptable risk levels and acceptable vs. unacceptable use cases for different categories of foundation models.

As for the modality of different models, one or another modality will be more dangerous depending on the context. For example, in the CSAM domain, image and video-based products will be much more concerning than text-based products. For misinformation in knowledge repositories, however, text-based products will be more concerning. It is important to note that we have no reason to believe that any one modality is more dangerous or more prone to misuse than another. Additionally, we predict in the near future that most models will shift to becoming multimodal, so concerns with the type of modality a model possesses will become less important.

3. How can the guidance better reflect the important role for real-world monitoring in making risk assessments?

The issue of how developers respond to incidents of model misuse is not highlighted enough throughout the document. The AISI draft suggests in Practice 6.2 that developers should "Maintain a process to respond to incidents of model misuse", however this practice is not mentioned until Objective 6, at which time the model has already been deployed. A process to respond to misuse incidents should be envisioned from the beginning and be included in Objective 1. If a process for response is not created until incidents of misuse are already arising, organizations will not be prepared to handle these cases and the incidents will not be handled in a timely manner. With the rate at which these models can generate objectionable content, and the speed of information spread across the internet, slow response to misuse incidents is a danger that cannot be allowed.

4. How can the guidance's examples of documentation better support communication of practically useful information while adequately addressing confidentiality concerns, such as protecting proprietary information?

AISI should advise documentation to be created at multiple levels of granularity. For example, in-depth analysis of dual-use capabilities should occur at the company level, and perhaps only be shared with key developers and C level executives. A synopsis of key details could then be shared outside of the company to summarize possible risks or misuse cases without including exact data. Summaries geared towards different stakeholders, such as developers at other foundation model companies, government officials, and users, could also be created, with different levels of detail as needed. As voluntary industry standards emerge, internal metrics should evolve to exceed these standards and anticipate novel issues. Disclosure should be as maximal as possible without releasing proprietary information, and users should expect a certain level of communication from developers. Users should also be made aware that developers, as well as researchers, cannot anticipate all possible misuse cases before they occur. As users evolve to hold certain expectations, such as being informed of misuse cases which may affect them, developers should shift to providing this information.

5. How can the guidance better enable collaboration among actors across the AI supply chain, such as addressing the role of both developers and their third-party partners in managing misuse risk?

We discussed, in our answer to question 4, the importance of developing industry standards for AI models. Along with standards for disclosure, standards for data gathering, shared metrics, and acceptable levels of risk and acceptable use cases will emerge as well. Similar to how many US products require warning labels or safety ratings, we expect these sorts of metrics to be established for foundation models. We also suggest that NIST support participation by non-profit organizations that will conduct evaluations of foundation models to provide a better understanding of their risk levels. Practice 6.4, outlined in the draft, will greatly benefit these organizations and the community at large as they conduct these kinds of evaluations, and will help to incentivize the creation of more Nonprofits aimed at improving the safety of foundation models.