U.S. AI Safety Institute
National Institute of Standards and Technology
Department of Commerce
100 Bureau Drive
Gaithersburg, Maryland 20899

Re: AI Safety Institute Consortium Task Force 1.3: Development of a Virtual Reporting Template (VRT) for the NIST AI 600-1 Generative AI Profile

UL Research Institutes (ULRI) welcomes the opportunity to submit comments to the U.S. AI Safety Institute Consortium as part of its crucial Task Force 1.3 objectives, specifically the development of a Voluntary Reporting Template (VRT) for the NIST AI 600-1 Generative AI Profile. We believe that a successful such VRT stands to not only improve safety and common understanding of risks posed by generative AI systems but also bring the broader NIST AI Risk Management Framework into wider practice.

As a leading safety science organization with extensive experience in digital safety research, ULRI is well-positioned to contribute to this important endeavor. Our attached recommendations are the result of both the combined institutional expertise of UL Standards & Engagement (ULSE), UL Solutions, and ULRI, and the valuable insights gained from the task force's collaborative meetings over the past few months. During these meetings, ideas for developing the VRT were openly conceived, exchanged, and evaluated among members, leading to the identification of key findings and novel suggestions.

ULRI and its collaborator affiliates ULSE and UL Solutions all strongly support the AI Safety Institute's mission to advance the science, practice, and adoption of AI safety. We are committed to continuing our contributions to the AI Safety Institute through the Consortium and its Working Groups. If you require any additional information or clarification regarding our recommendations, please do not hesitate to reach out.

Thank you for the opportunity to provide input on this critical initiative.

Sincerely,

Kevin Paeth
Technical Lead, Digital Safety Research Institute
UL Research Institutes

Nicholas C. Judd, PhD
Lead Research Engineer, Digital Safety Research Institute
UL Research Institutes

Suggestion 1: Align the VRT to use existing internal organizational controls and AI governance artifacts in order to reduce burden of reporting and maximize quantity of reporting entities.

Many organizations that would consider reporting compliance with NIST AI 600-1 have already developed internal AI governance processes that seek to follow best practices from the NIST AI RMF (published March 30, 2023) as well as ready compliance with the European Union's AI Act (published July 12, 2024; entered into partial force August 1, 2024). These organizations might be considered "candidate reporters," and are likely to have invested not only in purely meeting standards recommendations or requirements of legislation but *also in producing processes and artifacts that generalize to multiple frameworks*. This includes creating institutional AI governance processes that generate many artifacts intended for review, either internally (for example, with trust and safety teams) or externally (for example, with clients). ULRI observes this trend among partner organizations in industry.

In addition, from discussion and presentations among members the Task Force has learned that a great number of the actions proposed in the Generative AI Profile may correspond to common AI development-type artifacts (such as model cards or data sheets) or related governance-type artifacts (such as model use policies or impact assessments).

A successful VRT will empower candidate reporters to bootstrap compliance with emerging AI safety controls on the basis of these existing artifacts.

As such, **we recommend that the VRT should:**

1. Guide candidate reporters by identifying which such artifacts can be used to demonstrate compliance with particular controls in the Generative AI Profile and allow these artifacts to be shared in the template, and
2. Specify that such artifacts demonstrate compliance with particular controls in the Generative AI Profile **if and only if** they satisfy clear, specific requirements included in the VRT.

The use of these artifacts is consistent with some suggested actions of the Generative AI Profile, such as that for *MG-3.1-005* which suggests that AI actors "review various transparency artifacts (e.g., system cards and model cards) for third-party models." On the other hand, examples of even the most well-known artifacts, such as model cards or dataset data sheets, are inconsistent in their coherence and quality. We conjecture that by introducing a difference between "a model card," for instance, and "a model card conforming to VRT guidelines," NIST will create an opportunity for reporters to bring existing practice up to a more consistent level of completeness and comprehension. We further conjecture that this would achieve the same ends as a wholly new VRT without obliging reporters to duplicate efforts, thus increasing the likelihood of adoption.

In effect, this would make reporting more *accessible and less burdensome*; candidate reporters will be more likely to voluntarily report application of the Generative AI Profile on account of already having produced artifacts helpful for measuring compliance and knowing to which suggested actions they might map.

Suggestion 2: Prioritize particular suggested actions by relevant GAI risk to increase reporting and make reports more useful to downstream AI actors.

Candidate reporters using the VRT for the Generative AI Profile should be confident that their reporting is meaningful to consumers of this information and that it will not be lost to prospective purchasers, deployers, or other AI actors due to complexity or size of reports (the Profile contains over 200 suggested actions). Additionally, while the most competitive or well-resourced actors may be able to exhaustively report and identify progress across all controls, other candidate reporters may not be able to, nor might they all be relevant to the system, product, or practices at hand.

Similarly, the VRT should result in reports that are usable by these actors, allowing the most important information to be most easily communicated regardless of how exhaustive reports may be.

A potential way to encourage this result **is to explicitly prioritize particular suggested actions in the Generative AI Profile for purposes of the VRT and allow omitting certain suggested actions.** This would set the expectation that reporting using the VRT does not require being exhaustive for all controls, preventing would-be reporters from dropping out. Similarly, prioritized suggested actions could be emphasized to report consumers, making the VRT more usable.

One way to potentially prioritize suggested actions is by using the set of risks in the Profile. For example, if the generative AI system(s) in question are deemed to have risks relevant to chemical, biological, radiological, or nuclear (CBRN) capabilities according to the reporter, the VRT could allow emphasizing (for reporting and reading) the set of actions where this is identified as a relevant family of risks. Alternatively, the same sections could be de-emphasized (made optional or reordered in the report) if the training data and/or deployment system of the model are so constrained that such a set of risks are irrelevant. For example, a deployed system that is not trained on PII or private data nor requires interaction or data collection by human subjects obviates multiple suggested actions in the MEASURE 2.2 group; such elements could be re-ordered to the back of a VRT report.

Suggestion 3: Allow reporters to identify planned actions.

Candidate reporters and users of the Generative AI Profile will be in various stages of the AI system life cycle across different projects but should always be concerned with governance and management of resulting AI systems even when they are only just being designed or developed. Accordingly, the most competitive AI actors intending to bring a generative AI system to deployment or market might be interested in using the VRT to characterize products or systems still under development and testing. At these points, it will not yet be possible for the reporter to satisfy many of the suggested actions.

If it is envisioned that the VRT is used not only for finalized products but as well as for characterizing developing ones, **we recommend that the VRT allow for explicitly recording the intention of the AI actor to meet a suggested action that is not yet realizable.** For example, Action ID *MG-4.2-001* suggests that AI actors "conduct regular monitoring of GAI systems and publish reports detailing the performance, feedback received, and improvements made," which may not be possible while the system remains undeployed. Instead, a reporter should optionally retain the ability to indicate that this is a planned capability, and the VRT should accommodate this.

This suggestion complements the prior Suggestion 2, which allows for re-ordering or emphasizing different controls.