Written Opening Statement of Clement Delangue
Co-founder and CEO
Hugging Face

For the AI Insight Forum Kickoff
September 13, 2023

Senator Schumer, Members of the AI Insight Forum, and my fellow roundtable contributors, thank you for convening and bringing your insights to this important and urgent discussion. In order to ensure AI is developed safely and in U.S. interests, we must continue to coordinate across sectors and expertise. My name is Clement Delangue, and I am the co-founder and CEO of Hugging Face.

Hugging Face is a community-oriented company based in the U.S. with the mission to democratize good machine learning. We conduct our mission primarily through openness, which takes many forms, including public communication about the technology, open science, and open source. Our platform hosts machine learning models, datasets, and infrastructure that supports research and resource accessibility, aiming to lower the barrier for people from all backgrounds to contribute to AI.

At Hugging Face, we are seeing the active contributions and incredible potential for AI to improve American innovation across sectors, and benefit people and society. Amid the impressive technical achievements, there are also present-day risks such as harmful stereotypes, misinformation, threats to elections, and rising carbon emissions, which have led to our investment in ethics research. AI research and development is at a critical point where we need policymaker guidance.

This past June, I testified before the House Committee on Science, Space, and Technology to share how, together, we can ensure AI's development in the national interest. My first point to the committee highlighted the importance of fostering safe innovation via access and collaboration, and I was delighted with how closely our ongoing work paralleled the SAFE innovation framework announcement. Each aspect of the acronym will complement the next and we are eager to support your policy objectives in addition to sharing our perspective at the forefront of openness in AI. **I look forward to discussing with you the need for openness and transparency; measuring and mitigating harms to people and society; and investments in safeguards and controls.**

Openness and Transparency

Openness is a spectrum. Protections for more open systems and guidance for system transparency will benefit people who use AI, researchers, regulators, and the economy. Most advancements and notable AI systems in use today are based on open science and open source; popular large language models are largely based on accessible research, such as the 2017 "Attention Is All You Need" paper from Google. Open development allows us to pool resources, sharing ideas and tools that make AI as safe as possible, rather than more siloed and individualized approaches developed at each organization, which can create redundant work and hamper the ability to learn from others' mistakes.

More open systems empower perspectives from industry, academia, civil society, and independent researchers to contribute to research and risk mitigation, such as understanding and mitigating harmful biases against protected classes. When researchers are able to access not only models, but also artifacts such as datasets, they can better understand and evaluate those systems and build novel and beneficial applications. We see efforts such as red-teaming, recently exemplified at DEFCON with the AI Village, Rumman Chowdhury, and the White House's leadership, as successful examples of broader access making systems more secure.

Openness cultivates a thriving startup ecosystem and protecting access to systems, research, and tooling will help realize economic gains and improve American lives. AI applications are only at the beginning of their potential in many sectors, and sector-specific experts are starting to use AI for breakthroughs, such as medical researchers developing life-saving treatments for poorly understood human diseases. Open and collaborative development helps developers leverage expertise and inventiveness across sectors to optimize safety and performance. With these opportunities, we recognize that any system at any level of openness holds risks and misuse potential; all systems require controls. Our approach to openness balances tensions using policy and technical safeguards, from community moderation to gating mechanisms.

Openness enables the highest quality technology possible. For example, access to base models and training datasets makes it possible for scientists from diverse backgrounds to understand the complexity and scale of the mechanisms that lead AI systems to cause harm. **Accessible and transparent documentation makes it possible for consumers and legislators to better understand systems' strengths and weaknesses, making informed decisions about how they should and should not be used.** Grounding on openness as a key value can thus help meet priorities such as responsibility, accountability, and justice.

Toward this goal, Hugging Face provides tools and technical support to help test and inspect its hosted AI components, and promotes the creation and adoption of model and dataset cards as the first system information users see. Hugging Face's commitment to model and dataset documentation has led to hundreds of thousands of model cards and dataset documentation on our Hub, aiding users in making informed decisions between different options.

Policymaker guidance on transparency requirements for key system components, such as pretraining data, fine-tuning data, and models, can help create a standard for disclosure. In particular, attention to data subjects can protect and amplify the voices of less visible people involved in creating data, from annotators to artists whose work is incorporated into training data. Our hosted initiatives to build systems openly is pioneering documentation approaches that focus on the people in AI and has proved to excel in existing regulatory proposal compliance research. **Policy guidance on mandatory disclosure standards can also enable stronger personally identifiable information (PII) and intellectual property (IP) protection**; requiring human-centered risk documentation related to sensitive information in training data, inputs, and prompts, can help prevent privacy and IP violations.

Measuring and Mitigating Present Day Harms

The landscape of harms from AI systems deployed today requires better understanding. The personal and societal impacts of AI are emerging and will continue to emerge as systems are increasingly integrated into daily life, but how to fully evaluate these impacts and conduct robust impact assessments requires significantly more investment. In order to address present day harms, such as negative stereotypes of protected class demographics in outputs and adverse economic and labor impact, **we encourage policymakers to shape risk taxonomies and guide prioritization**. This includes specifying sectors and use cases that are most likely out-of-scope or highest-risk for a given system, which is especially needed for systems that are developed without a specific use case and are sometimes referred to as general-purpose.

The most important risk factors for an AI system are its scale and context of use. **We see proportional requirements for systems by use, sector, risk, and breadth of impact as the most effective means for mitigating present day harms while encouraging innovation.** We hold ourselves accountable by prioritizing and documenting our ethical work throughout all stages of AI research and development.

The best approaches for measuring and mitigating risk will differ by system modality and use. For example, biases in image generation manifest differently than those in language. While complex social impacts like bias are not technically solvable, better evaluations can shape out-of-scope uses and mitigation research agendas. **Evaluations and measurements complement disclosure**; tools to examine PII in training data can aid in notifying data subjects and assessing the legality of the processed data. Better carbon emission measurement and mandated disclosure can incentivize more environmentally friendly practices, and help ensure that the carbon footprint of the technology is commensurate with its utility. These overarching impact areas will each need enumeration and more policy guidance.

Sectoral policies and regulation should be updated to consider AI contexts, particularly to protect marginalized groups from AI exacerbating inequality or unfair outcomes. Evaluation and audits are best performed in the context of a given sector and context. Policy guidance for test & evaluation, validation & verification (TEVV) experts, such as requirements for in-house and

external actors, should be clarified. TEVV experts should be community-oriented, represent relevant computer and social science disciplines, and be protected with fair work conditions. Further clarification is needed for tests by system component through a system lifecycle, from dataset collation and governance, to training processes, to model deployment.

Investing in Safeguards and Reliability

The need to build mechanisms and infrastructure for building safe and transparent AI is only growing. In order to ensure leading AI development is in the national interest and the U.S. continues its technological leadership, **the U.S. government should support more research infrastructure and incentives for building and applying safeguards and safety validation.** Ensuring safety and security in system development and for people and society is not solely a technical effort and will require input from many disciplines. **We strongly support more funding for the National Institute of Standards and Technology (NIST) and funding the National AI Research Resource (NAIRR).**

Similarly, while Hugging Face is based in the U.S., our contributors from around the world strengthen research and the community. As the fastest-growing open-source library of pre-trained models in the world, we have seen openness and access be key to making systems work better for more people. Supporting broader perspectives will require building accessible infrastructure for lower resource researchers and more tailored resourcing for social impact and safety research. **Coordination with our democratic allies will bring needed perspectives on human value alignment and shape open questions on tradeoffs between what constitutes high system performance and safety.**

Conclusion

In order to bolster safety research and mitigate critical risks in AI today, we must encourage openness and disclosure as is appropriate. Researchers need more access to models and infrastructure to build better evaluations and research that addresses present-day harms. And ongoing, agile investments are needed for safeguards to evolve with the AI landscape. We stand ready to support and be a resource to the Forum and again thank you for this initiative.