Comments on the March 17, 2022 Initial Draft of the NIST AI Risk Management Framework

To: Elham Tabassi and the NIST Team

From: OpenAI

OpenAI is an artificial intelligence research and deployment company working to ensure that artificial general intelligence benefits all of humanity. We are grateful to NIST for the opportunity to review and provide feedback on the thoughtful and in-depth initial draft of the AI RMF. This draft serves as a solid foundation for the development and use of trustworthy and responsible AI. In this note, we respond to your request to identify issues that may be missing from the draft framework.

We recommend that NIST consider incorporating into its guiding principles **alignment with human values and intentions**.[1] Aligned AI systems reliably do what humans intend for them to do and the concept of alignment[2] is therefore directly relevant to the guiding principles' overarching concern of safeguarding "broader societal norms and values that indicate societal priorities". We believe that alignment is a crucial guiding principle for AI development, particularly for AI systems that will interact regularly with individuals, and over time will be increasingly integrated into our economy and communities. As AI systems become more powerful, risks stemming from alignment failures could be substantial.[3] NIST is well-placed to highlight this as an important challenge for AI risk management.

Alignment as a guiding principle serves to incentivize important discourse and research among AI developers on achieving systems' aligned with human values and intentions. The goal is to ensure that these aligned AI systems reliably do what humans intend them to do. Experimental training techniques that focus on integrating human feedback into the development life cycle of an AI system have already yielded early results in producing systems that are more aligned with human intent.[4] Given the rapid advancement of AI systems, continued progress in this area is needed. Human intention is a complex concept, and ultimately it will be necessary for AI systems to be able to grasp this complexity and integrate context, values, and preferences reliably.[5]

[1] NIST's Taxonomy of AI Risk notes alignment-related concepts from the OECD principles of "traceability to human values" and EU principle of "human agency and oversight," but NIST does not explicitly include alignment in its guiding principles.

[2] https://openai.com/alignment/

[3] For a discussion of some of the risks, see: [2109.13916] Unsolved Problems in ML Safety, The Alignment Problem: Machine Learning and Human Values

[4] For example: Aligning Language Models to Follow Instructions, [2204.05862] Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

[5] For a discussion of some of the challenges, see: [2001.09768] Artificial Intelligence, Values and Alignment