PARTNERSHIP ON AI
to benefit people & society

Safety-Critical AI : Charter

Recent and rapid progress in artificial intelligence (AI) and machine learning (ML) raises the question: How can we ensure that these technologies are safe? This is an urgent short-term question, with applications in medicine, transportation, engineering, computer security, and other domains hinging on the ability to make AI systems behave safely despite uncertain, unanticipated, and potentially adversarial environments. It is also a pressing longer-term question. As technologies become more capable across various domains, we will need social and technical foundations for building AI technologies that are safe, predictable, and trustworthy; they must also align with the ethics, needs, and normative expectations of the individuals whom they affect in a variety of contexts, applications, and environments.

These objectives lead us to more specific questions: Who defines what it means for a technology to be safe? How do the creators of AI systems choose what safety constraints and values to build into such systems, and how should other stakeholders (i.e., users or regulators) influence or engage with these choices? What design and training processes can enable safe implementation of AI systems? What are sufficient ways to measure whether we have succeeded in developing safe systems, and can measurement be a tool to activate and motivate other stakeholders to work on safety? What role do safety factors, training, certification, or similar tools have to play? The Partnership on AI's Working Group on Safety-Critical AI must investigate the norms and institutions necessary for navigating these challenges, while thinking through how we can generalize successes and best practices in one domain to other areas.

This Working Group gathers representatives from technology companies, civil society, and academia who believe that these communities, along with the broader public, must think about how to design safe AI systems. In order to assist the AI community and society-at-large in answering those questions, our group will convene to address **topics including, but not limited to, the following:** defining and designing safety-critical AI, exploring attacks on deployed ML systems, investigating techniques to measure systems' safety and predictability characteristics, and evaluating computer security and AI; we will also promote data collection, labeling, and sharing and incorporate study of human factors in our considerations of AI safety.

There will be overlap and synergies across these topics, so the Working Group will deliberate and engage in a conversation around them and strive for delivering actionable output. Possible **deliverables** could include best practices and recommendations for those researching, designing, developing and/or deploying AI systems, as well as datasets, benchmarks, testbed-environments, platforms, and/or contests to help gauge AI safety in priority areas. We may also declare key principles, produce vision documents and risk models for specific categories of AI systems, as well as incubate a research community devoted to engaging the public through workshops or conferences.

Finally, these deliverables will be informed by an **execution roadmap** consisting of convenings in person and online, offsite meetings and virtual convenings for project teams, and published, promoted, and presented work in appropriate fora (online, press, and/or workshops).