CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Summary of *The Inigo Montoya Problem for Trustworthy AI*

Roughly 14 percent of AI papers include at least one of 13 trustworthy artificial intelligence keywords. The growth in the number of publications using these terms exceeded the growth of research on AI generally in the past five years. But not all keywords are used in the way policymakers think of them:

- The keywords ***reliability* and *robustness* were the most frequently mentioned trustworthy AI terms**. The frequency of these terms may be due, in part, to the fact that they are generally expected evaluation metrics.
- While **a majority of the papers using the terms *reliability*, *safety*, *security*, and *resilience* are aligned with policy concerns**, a significant minority of the papers used the terms in reference to how AI could improve the *reliability*, *safety*, *security*, and/or *resilience* of a non-AI system.
- **The term *bias* in AI papers has two main uses**, one is technical and describes the components of an algorithm, and one refers to unfair discrimination. Research using the word appears evenly split between the two meanings.
- *Explainability*, *interpretability*, *transparency*, and *accountability* reference how to develop AI models and systems that an end-user can trust, specifically in the context of the Explainable AI (XAI) research area. This is interesting because **while trustworthy AI is not currently considered a research area, XAI has developed into one**.

Since 2019, prestigious AI conferences have used more trustworthy AI keywords in their calls for papers. All but the keywords *reliability* and *resilience* are used in one of 11 top AI conferences.

Recommendations:

To keep making progress towards trustworthy AI, policymakers should understand how and why technologists use trustworthy AI terms and track the patterns of term use in research literature and in top AI conferences.

For more information:

- Download the report: https://cset.georgetown.edu/publication/the-inigo-montoya-problem-for-trustworthy-ai
- Contact: Emelia.Probasco@georgetown.edu