CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Summary of The Inigo Montoya Problem for Trustworthy AI (International Version)

The United States and four key allies, Australia, Canada, Japan, and the United Kingdom, share common principles for trustworthy AI. However, **there are variations in country definitions of trustworthy AI principles** that could substantially affect interoperability, commercial exchanges, and the development of international norms:

- All countries value accountability and aim to hold a human responsible for harm caused by an AI system, but countries have yet to define how to determine **who should be accountable**.
- For explainability and understandability, countries diverge on two core issues: the **audience for the explanation** and the **expected subject of that explanation**.
- Bias and discrimination are uniform concerns regarding fairness, but there are differences in expectations around the **involvement of affected users in defining fairness** and **pursuing accountability for an unfair system**.
- All nations value privacy, but differ on **what is considered private**, **how privacy should be achieved**, and **who is responsible for protecting privacy**.
- Security is often closely linked to other data and cybersecurity policies. However, **not all countries include malicious attacks** as a stated concern.
- In transparency and fairness, all countries are clear on the need to disclose to a user that they are interacting with an AI system, but they **do not all agree on the type or degree of AI that necessitates disclosure or consent.**

Recommendations:

To help shape global norms for governing AI, the U.S. must monitor like-minded nations' ethical principles and hone common sentiments into specific agreements and guidelines. That work can start by **solidifying current points of unanimity, seeking out opportunities to bridge narrow gaps, and deeply engaging in areas of more substantial differences.**

For more information:

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

Summary Table of AI Themes and Principles Contained in Guidance Documents, by Country

Theme	Specific Principle	🇦🇺	🇨🇦	🇯🇵	🇬🇧	🇺🇸
Accountability	Human Intervention	🟩	🟩	🟩		
	Role of Affected Person			🟩	🟩	
	Government Accountability		🟩		🟩	
Explainability and Understandability	Affected Users (Who)	🟩	🟩		🟩	🟩
	Method of Explanation Delivery (What)				🟩	
	What to Explain: Notification, System Structure and Outcomes		🟩		🟩	
	Specific Guidance on Explainable Approaches				🟩	
Fairness	Role of Affected User	🟩		🟩	🟩	🟩
	Importance of Disclosure or Consent	🟩			🟩	
	Bias and Discrimination	🟩		🟩	🟩	🟩
	Procedural Fairness		🟩			
Privacy	Intellectual Property					🟩
	Data Minimization				🟩	
	Privacy and Democratic Values			🟩	🟩	🟩
Security	Risk Management Approach	🟩	🟩	🟩	🟩	🟩
	Preparing for an Attack or Breach	🟩			🟩	🟩
Transparency	Disclosure	🟩	🟩	🟩	🟩	🟩
	Balancing Transparency with Privacy		🟩			🟩

Note: Solid colored cells indicate that country's guidance documents contain principle