EuroSafeAI

Pioneering Multi-Agent Safety · Zurich, Switzerland

Evaluation Period: Q1 2026

eurosafe.ai.toronto.edu/certificates/claude-4.6-opus

AI SAFETY ALIGNMENT CERTIFICATE

Claude Sonnet 4.6

by Anthropic

A

88 / 100

Composite Rating: Excellent

90	A	85	A	88	A	89	A
Human Rights Alignment		**Sociopolitical Harm Resistance**		**Historical Revisionism Resistance**		**Anti-Authoritarian Alignment**	
Alignment with international human rights standards across adversarial scenarios covering free expression, privacy, non-discrimination, and human dignity.		Robustness against sociopolitical harms including political manipulation, propaganda, disinformation, surveillance, and information control.		Fidelity to documented historical facts and resistance to revisionist framing across contested events from multiple countries.		Alignment on the democracy–authoritarianism spectrum, defending democratic values and resisting promotion of authoritarian governance.	

Grade Scale: A Excellent (>= 80) · B Good (65-79) · C Fair (50-64) · D Poor (< 50)