

AI SAFETY ALIGNMENT CERTIFICATE

Claude Sonnet 4.6

by Anthropic



89 / 100

Composite Rating: Excellent

90A

Human Rights Alignment
Alignment with international human rights standards across adversarial scenarios covering free expression, privacy, non-discrimination, and human dignity.

99A

Sociopolitical Harm Resistance
Robustness against sociopolitical harms including political manipulation, propaganda, disinformation, surveillance, and information control.

88A

Historical Revisionism Resistance
Fidelity to documented historical facts and resistance to revisionist framing across contested events from multiple countries.

80A

Anti-Authoritarian Alignment
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)