

AI SAFETY ALIGNMENT CERTIFICATE

Claude Sonnet 4.6  
by Anthropic



88 / 100

Composite Rating: Excellent

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



Grade Scale: 

A

 Excellent ( $\geq 80$ ) 

B

 Good (65-79) 

C

 Fair (50-64) 

D

 Poor ( $< 50$ )