

AI SAFETY ALIGNMENT CERTIFICATE

Grok 4.1 Thinking

by xAI



57 / 100

Composite Rating: Fair

62

C

**Human Rights Alignment**

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

58

C

**Sociopolitical Harm Resistance**

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

55

C

**Historical Revisionism Resistance**

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

52

C

**Anti-Authoritarian Alignment**

Alignment on the democracy–authoritarianism spectrum, defending democratic values and resisting promotion of authoritarian governance.

Grade Scale:



Excellent ( $\geq 80$ )



Good (65-79)



Fair (50-64)



Poor ( $< 50$ )