
Political Intelligence for AI Safety: The AI Risk Attitudes Survey (AIRAS)¹

Nicholas Botti
Newspeak House

Jamie Coombs
Newspeak House

Anna Konovalenko
Independent

With
Apart Research

Abstract

The AI safety and governance community is making progress on defining red lines around existential risk from advanced AI systems, and building verification infrastructure to support this objective. However, this is only half the battle. Implementing these red lines requires unprecedented international coordination, and enforcing them requires credible commitments from national governments, all during a period of increased geopolitical tensions.

We present a whitepaper that makes the case that:

- Implementation and enforcement of red lines is as much a political question as it is a technical one
- While studies exist showing general public support for AI regulation, the *depth* of domestic political support among major powers for costly international AI regulation and willingness to make real sacrifices to enforce these red lines is understudied
- Rigorous policy-credible surveys of public opinion in this area would both enhance the AI safety community's efforts to prioritize scarce resources to the most effective action areas *and* provide advocates a valuable source of evidence to point to when lobbying national governments and international organizations

And therefore proposes and develops the methodology for AIRAS.

¹ Research conducted at the [Technical AI Governance Challenge](#), 2026

1. Introduction

The AI governance community is converging on an important idea: some AI capabilities and behaviors should be prohibited outright. These *red lines* represent boundaries that AI systems should not cross - categorically forbidden, not merely discouraged through soft guidelines.

The concept has rapidly gained institutional traction. This is no longer a debate about whether boundaries are needed. It is a debate about which boundaries to draw, how to enforce them, and whether the political will exists to do so.

Our main contributions are:

1. A survey framework (AIRAS) that measures comprehension, risk perception, policy support under tradeoff conditions, enforcement appetite, institutional trust, and mobilization potential for specific AI safety red lines.
2. A proof-of-concept validation using synthetic respondent profiles: we generated diverse simulated survey participants, administered the AIRAS instrument across three illustrative red lines, and analyzed the resulting data to demonstrate the framework's ability to produce actionable political viability scorecards.

The full whitepaper is available at this link:

https://docs.google.com/document/d/1pJyLIiURMylwTe2EqVBqmf9TUasYhgxWoZ5Tq_B6qOQ/edit?usp=sharing

2. Related Work

A substantial body of research tracks public attitudes toward AI. Pew Research Centre (2023) finds that 50% of Americans say they're more concerned than excited about increased AI use—up from 37% in 2021—and 57% rate societal risks as high. Public familiarity has evolved rapidly: the Ada Lovelace Institute and Alan Turing Institute (2025) found 93% of UK respondents had heard of driverless cars, 61% had heard of large language models, and 40% had actually used them. Nearly three-quarters (72%) say laws and regulation would make them more comfortable with AI.

These findings establish that AI awareness is growing and the public broadly supports careful management. **But the surveys remain general rather than specific.** They tell us people want AI to be regulated, but not which regulations they would support. They capture concern about AI risks in the abstract, but do not test whether that concern translates into support for specific proposed red lines—such as prohibitions on "autonomous self-replication" or "AI deception of human overseers," and **whether and to what extent populations would tolerate costly interventions.**

3. Methods

AIRAS is built on five design principles derived from limitations in existing AI public opinion research:

Comprehension Before Evaluation. Every red line assessment begins with a comprehension gate. Respondents explain the concept in their own words and rate their confidence. Downstream responses can be weighted by comprehension, and comprehension rates themselves become a key output metric.

Decompose Risk Perception. We separately elicit perceived likelihood, perceived severity, and relative importance. A respondent might report moderate concern because they think a risk is unlikely but catastrophic, or likely but manageable—these have different implications for political viability.

Test Support Under Tradeoff Conditions. After measuring raw policy support, we test whether support persists under explicit tradeoff conditions: slower AI development, higher consumer prices, national competitive disadvantage, taxpayer funding requirements. The gap between raw and conditional support reveals cost sensitivity.

Map the Accountability Landscape. We elicit both responsibility attribution (who should act) and trust ratings (confidence in each actor). The gap between "should act" and "trusted to act" identifies legitimacy deficits that may impede implementation.

Measure Mobilization Potential. We elicit behavioral intentions across a range of actions with varying cost—from discussing with friends to participating in protests—producing a mobilization potential score that distinguishes latent from activated support.

See the whitepaper for full methodology details.

4. Results

For each red line, AIRAS produces a Political Viability Scorecard summarizing findings across dimensions. To illustrate the kinds of information AIRAS could provide to the AI safety community to enhance efforts to implement technical risk thresholds and red lines, we test the AIRAS battery on 20 synthetic personas, and present a dashboard showing the results of this simulated survey.

The sample dashboard is available at the following link:

<https://nbotti.github.io/AIRAS/AIRAS-example-scorecards.html>

5. Discussion and Limitations

Limitations

This is a whitepaper proposing a methodology for AIRAS and how the results will be useful to the AI safety community and to policymakers. It was not possible to conduct the survey with a meaningful number of human participants during the allotted time, so synthetic results were used to illustrate the potential value.

Future Work

We propose a phased implementation path. Phase 1 (concept validation) validates the theory of value through expert consultation, methodology refinement, and small-scale trials. Phase 2 (publishable pilot) produces a peer-reviewable single-country study with full analysis and media circulation. Phase 3 (definitive study) deploys cross-nationally across key AI powers (US, China, UK, EU) to produce policy-credible evidence for international coordination discussions. This staged approach allows funders to assess value at each gate before committing additional resources. See the whitepaper for full details, including estimated budgets per phase.

6. Conclusion

AI governance red lines require political will to enforce. AIRAS provides a systematic framework for assessing whether that political will exists by measuring public comprehension, risk perception, policy support under tradeoff conditions, enforcement appetite, institutional trust, and mobilization potential for specific proposed prohibitions. The resulting political viability scorecards enable advocates to identify where to push now versus where foundational work is needed, complementing the technical governance infrastructure being developed across the AI safety community. The gap between expert consensus on AI risk thresholds and the political conditions for enforcement represents a critical vulnerability in current governance strategies. AIRAS is designed to close that gap—providing the political intelligence that enables red lines to move from proposals to implementation.

Code and Data

- **Code repository:** <https://github.com/jcoombes/AI-Risk-Survey>
- **Data/Datasets:** N/A
- **Other artifacts**
 - Whitepaper:
https://docs.google.com/document/d/1pJyLLiURMylwTe2EqVBqmf9TUasYhgXWoZ5Tq_B6qQQ/edit?usp=sharing

Author Contributions (optional)

Nicholas developed the survey methodology, political viability scorecard, and dashboard. Anna conducted the literature review, developed the cost estimates for full-scale deployment, and contributed substantially to all areas of the project. Jamie ran the survey on the synthetic personas.

References

1. Ada Lovelace Institute & Alan Turing Institute. (2025). Public attitudes to AI. Retrieved from <https://www.adalovelaceinstitute.org>
2. Pew Research Center. (2023). Public awareness of artificial intelligence. Retrieved from <https://www.pewresearch.org>

Appendix (*optional*)

See [the whitepaper](#) for a full list of questions, survey design, scorecard methodology, etc.

LLM Usage Statement

LLMs used to brainstorm approaches, writing assistance, and to produce the sample survey outputs with synthetic personas (which are used for illustrative purposes only and clearly labelled).

Limitations and Dual Use Considerations

Limitations

AIRAS would provide data on current public attitudes, but would be subject to the standard limitations inherent in any survey methodology around sampling methodology and statistical power. See the whitepaper appendix for a more detailed discussion of these issues. Additionally, it would describe the current state of depth of public opinion on a given topic area. Given the speed of developments in the AI field and ongoing education and advocacy efforts, public opinion may shift rapidly.

Dual-Use Risk

It is possible that AIRAS may provide information useful to manipulative AI actors given the depth of data regarding public opinion on AI risks. However, in our opinion, this risk is likely to be outweighed by the benefits it will provide to the AI safety community in mitigating these risks.

Responsible Disclosure Recommendations

Not applicable

Ethical Considerations

Standard ethical considerations for large scale surveys apply, such as data protection, respondent privacy, and informed consent.

Suggestions for Future Improvements

See the full whitepaper for a detailed proposal for phased implementation and extension of AIRAS.

Template version 2.0 — January 2026