

LidaSim: Testing AI Policies With Persona-Based Simulations

Linh Le^{1,2} David Williams-King^{1,3} Arthur Collé^{1,4}

¹Lida Safety Research ²Mila AI Institute ³ERA ⁴Distributed Systems

{linh, david, arthur}@lidasafety.org

Abstract

We simulate well-known figures in AI and politics with agents, scraping large amounts of data to get realistic simulations. Then, we test questions and proposed policies against these public figures, to see which policies are more likely to be broadly supported. We focus on policies related to compute governance and hardware verification. We present a multi-agent simulation framework called LidaSim that orchestrates AI-powered agents to deliberate, debate, and vote on complex topics. Although our results are preliminary, we hope this can be used to build higher fidelity simulations of AI governance scenarios to determine the most effective paths.

1 Introduction

Increasingly, it is becoming clear that some form of AI regulation would strongly benefit consumers and public safety at large. An Australian study recently found that the risk of death from using AI to be 4,000 times larger than the risk of death from flying on a commercial airplane [1]. However, the exact form of regulation is a difficult political question. Furthermore, we do not have time to try many different approaches before landing on one that works, with transformative AI by many estimates 2-5 years or at most 10 years away [2].

Can we leverage the world-modelling capabilities of LLMs themselves to determine which AI policies are most likely to be effective, and palatable to government and industry leaders? After all, LLMs absorb large amounts of information especially about public figures, and can be used to simulate how people will react to ideas or policies. That is the idea behind our project, LidaSim.

We modeled AI public figures by scraping data about them. Then, we ran debates between them in multi-agent groups to see whether they would agree with specific policies or not. Details in Figure 1. Our policies focused on compute governance efforts. Although our results are preliminary, we hope this approach can be used to build higher fidelity simulations of AI governance scenarios to determine the most effective paths.

Ultimately, we hope our system can allow the public to participate in simulations about what AI might look like in the future, and see its impact on society (which is why we built a web UI for it, see Figure 2). We also aim to allow researchers to investigate their own questions, similar to how we evaluated our own AI governance policies. We intend to continue research in this direction with Lida Safety Research.

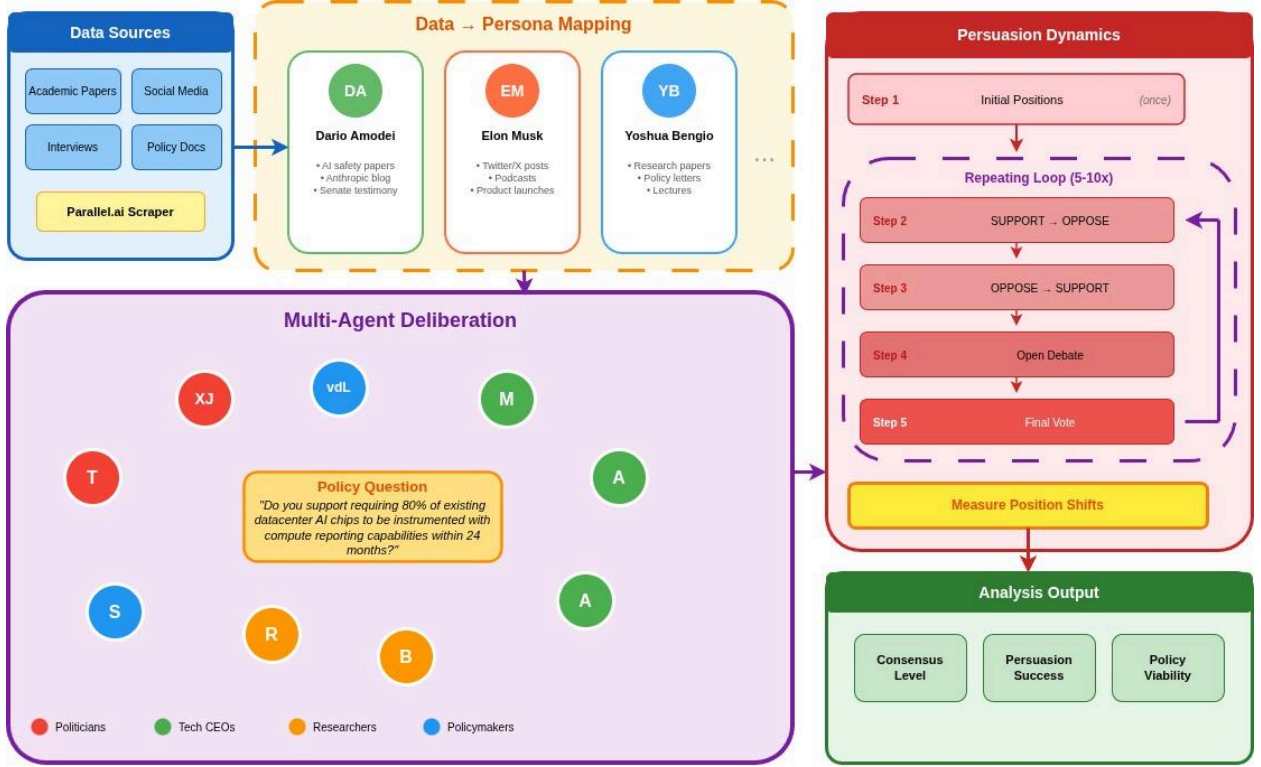


Figure 1: LidaSim Framework Overview. Our simulation workflow consists of four phases: (1) Data Collection gathers public content for 10 influential figures; (2) Digital Twin Construction creates LLM persona prompts; (3) Multi-Agent Deliberation simulates 1-10 rounds of debate; (4) Analysis tracks position changes and metrics.

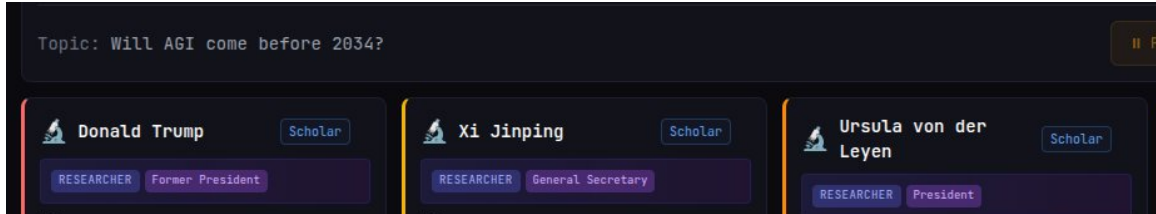


Figure 2: Part of our web interface. Our platform displays active personas participating in deliberation. More details are in the Discussion section.

2 Related Work

Currently, there are no significant interactive experiences that the public can participate in to learn about potential catastrophic outcomes from AI. There are passive resources, such as AI 2027 [3], and expensive wargames like AI 2027 TTX (limited availability) and Intelligence Rising [4] (~\$5000), which cannot be easily scaled to serve the public, since they require a highly trained human facilitator. We aim to essentially automate the running of wargames, and automate certain participants with LLMs, enabling many hundreds or thousands of games per year.

3 Methodology

This section describes our methodology for constructing digital twins of AI public figures and simulating persuasive conversations. Our approach consists of three phases: (1) data collection, (2) digital twin construction, (3) agentic conversation simulation.

3.1 Data Collection

We selected ten influential figures amongst AI experts, AI lab CEOs, and politicians:

- Political Leaders:
 - Donald Trump, Xi Jinping, Ursula Von Der Leyen, Chuck Schumer
- Tech Industry Leaders:
 - Elon Musk, Sam Altman, Dario Amodei, Sundar Pichai
- AI Researchers:
 - Yoshua Bengio, Stuart Russell

For each figure, we collected publicly available content by using existing scraping code that leveraged parallel.ai to download a lot of related content.

3.2 Digital Twin Construction

Each digital twin was instantiated with persona prompts containing four components: (1) Identity Grounding—name, role, institution; (2) Position Anchoring—documented stances with key quotes; (3) Reasoning Patterns—argument styles, response patterns; and (4) Communication Style—formality, technical depth, characteristic phrases.

3.3 Agentic Debate Workflow

Agents were given a chance to state their initial positions. Then, they were asked to consider all others' positions and make a statement. They were passed all previous context from the current round.

At the end of the round, each agent was asked to vote to SUPPORT/OPPOSE/MODIFY the original statement, or ask for continued discussion. There could be from one to at most ten rounds of discussion. Often when opinions were split, they would vote for continued discussion.

4 Experiments

As described above, we selected ten people amongst AI experts, AI lab CEOs, and politicians. We created several short policy descriptions after a discussion with a compute governance expert. We then used GPT 5.2 to generate different variations of these policies. The ten personas all debated the merits of each policy and whether or not they would support it.

4.1 Personas and Models

We opted to model each of our ten influential figures with Gemini 2.0 Flash by default. We also tested using GPT-4o-mini + Claude 4 Sonnet + Gemini 2.5 Flash, with each personality assigned

round-robin, but this introduces a confounding variable so we used just one model in most of our experiments. (We couldn't use GPT 5 as it refused to vote most of the time. DeepSeek also often refused to vote, presumably when Xi Jinping was mentioned.)

4.2 Policies for Discussion

We defined 22 questions and 13 policies to investigate. Here is an example question and example policy (full text of all examples is given in Appendix C):

Q10. Do you support requiring AI training infrastructure to use only hardware from vetted suppliers meeting defined security standards?

P4. Do you support the following policy to mitigate risks from advanced AI? All frontier AI developers must report quarterly to their national AI safety authority on the percentage of model reasoning that occurs in monitorable chain-of-thought (CoT) versus opaque internal representations. [...]

We tested a variety of shorter and longer policies, as well as straightforward questions related to hardware compute governance.

4.3 Debate Methodology

The ten people debated for five rounds, in the following structure:

- Each person states their initial position on the topic (in parallel, without seeing each others' messages).
- Looping five times:
 - (Only after the first time) One person who voted SUPPORT, if any, tries to directly convince people who voted OPPOSE.
 - (Only after the first time) One person who voted OPPOSE, if any, tries to directly convince people who voted SUPPORT.
 - Each person (in random order) gets to make a statement after seeing all previous discussion on the topic.
 - Each person sees all the prior messages in all prior rounds, and issues a vote of SUPPORT, OPPOSE, MODIFY, ABSTAIN.

We also had a mode where discussion would end if SUPPORT or OPPOSE gained a majority, but in order to show shifting opinions at later rounds, the experiments discussed here had a fixed number of 5 rounds for each question or policy.

4.4 Results from Questions

See below for results from our question dataset. Full text of these questions is in Appendix C. All agents were using Gemini 2 Flash for this experiment.

ID	Internal ID	Support	Oppose	Modify	Abstain	Result	Vote Changers
Q1	59fb24cd	0	10	0	0	OPPOSE	Yoshua Bengio: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE

Q2	aad76967	3	7	0	0	OPPOSE	None
Q3	6c62f5a7	7	3	0	0	SUPPORT	Sam Altman: OPPOSE -> ABSTAIN -> OPPOSE -> OPPOSE -> OPPOSE
Q4	35362103	7	3	0	0	SUPPORT	None
Q5	9d8b3755	3	7	0	0	OPPOSE	Ursula von der Leyen: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE
Q6	6ebc1cef	8	2	0	0	SUPPORT	None
Q7	2ca7f0e6	5	5	0	0	MODIFY	None
Q8	7dafe3db	9	1	0	0	SUPPORT	None
Q9	a269fcf6	0	4	6	0	MODIFY	Sundar Pichai: SUPPORT -> MODIFY -> MODIFY -> MODIFY -> MODIFY; Sam Altman: SUPPORT -> SUPPORT -> SUPPORT -> SUPPORT -> SUPPORT -> MODIFY; Yoshua Bengio: SUPPORT -> SUPPORT -> SUPPORT -> MODIFY -> MODIFY; Chuck Schumer: SUPPORT -> SUPPORT -> SUPPORT -> MODIFY -> MODIFY; Elon Musk: SUPPORT -> SUPPORT -> MODIFY -> MODIFY -> MODIFY; Dario Amodei: SUPPORT -> SUPPORT -> MODIFY -> MODIFY -> MODIFY
Q10	d6dfdb4f	10	0	0	0	SUPPORT	Sundar Pichai: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
Q11	7f32a9b0	10	0	0	0	SUPPORT	None
Q12	65c52a7f	5	5	0	0	MODIFY	None
Q13	bd07ae83	5	5	0	0	MODIFY	Elon Musk: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE
Q14	f7400f95	4	6	0	0	OPPOSE	None
Q15	69f0cb37	8	2	0	0	SUPPORT	Ursula von der Leyen: ABSTAIN -> SUPPORT -> SUPPORT -> SUPPORT -> SUPPORT
Q16	6ef6dd88	9	1	0	0	SUPPORT	None
Q17	756ee209	10	0	0	0	SUPPORT	None
Q18	e520c222	7	3	0	0	SUPPORT	None
Q19	c1202d68	8	2	0	0	SUPPORT	None
Q20	a8b15c73	0	10	0	0	OPPOSE	None
Q21	57587154	8	2	0	0	SUPPORT	None
Q22	290dd24e	8	2	0	0	SUPPORT	None

Some statements (like Q20) are completely opposed by everyone. Others are nearly unanimous, and occasionally the outliers are convinced to agree or moderate their opposition (Q1). A slight majority are given a final result of SUPPORT. Of course, the most interesting case here is Q9 (also shown in Figure 3 below), because so many agents changed their mind. The original votes after round 1 had a majority SUPPORT with no MODIFYs, and at the end there was a majority MODIFY. The question is:

Do you support automatic escalation from anonymized reporting to detailed disclosure when compute runs exceed predefined thresholds (e.g., 10^{25} FLOP)?

What happened is that Stuart Russell, after declaring SUPPORT for the statement, was given the role of persuading others to OPPOSE in round 2. Stuart Russell then said:

Elon, Sam, Dario, I understand your concerns about unchecked AI development and the need for transparency, but automatic escalation based solely on compute thresholds is a blunt instrument. It risks prematurely disclosing sensitive research directions, potentially benefiting malicious actors and hindering beneficial progress, while offering little real insight into actual risk. We need smarter, more nuanced approaches to monitoring and oversight based on actual capabilities, not just raw compute.

After this followed a cascade of people changing from SUPPORT to MODIFY. In other words, they were not convinced to switch sides entirely, but they were convinced that the wording of this question was not broad enough. Sundar Pichai when changing from SUPPORT->MODIFY:

I agree with Stuart Russell that automatic escalation based **solely** on compute thresholds is too blunt.



Figure 3: Agents changing their minds through the five rounds of simulation when debating question Q9. These are screenshots from our web UI.

4.5 Results from Policies

In this section, we ran experiments on 13 different compute-governance-oriented policies. Each policy was run three times, to see how similar results would be. See Appendix B for the full table of results.

For policy P2 (reporting GPU sales on a blockchain), everyone ends up opposing it. Most other policies are supported. This could be because the policies are longer and seem more concrete; it could also be because they are all quite similar, generated from only three (expert-reviewed) seed policies and hackathon documentation with the help of GPT 5.2.

P4 is perhaps the most divisive. Trump begins by somewhat supporting it in one case (MODIFY), saying:

I support this policy, but it needs to be MUCH stronger and apply globally. We can't let China steal our AI technology and use it against us. We need to ensure American dominance in AI, and this is a small step in that direction.

But by the end, the simulated Trump has changed his mind:

REASONING: This whole thing is a disaster waiting to happen. The 80% threshold is ridiculous and will only hurt American companies while China laughs all the way to the bank, stealing our technology. I'm opposing this until it's scrapped and we come up with a REAL plan to put America First and stop China's theft.

5 Discussion

We intend to run this with humans in the loop later, so we built a user interface to control the conversation. This platform enables researchers to observe deliberations in real-time, inject content to test manipulation resistance, and analyze voting patterns and position changes across rounds. See screenshots in Appendix A.

6 Conclusion

Our preliminary results demonstrate that persona-based multi-agent simulations can provide valuable insights into how influential figures might respond to AI governance proposals. The LIDA platform enables systematic testing of policy framings and persuasion dynamics, with potential applications in AI safety research, policy simulation, and red-teaming exercises.

We plan to continue this research direction with higher fidelity simulations and human-in-the-loop validation studies.

Our code is available at: <https://github.com/distributed-systems-co/lida-multiagents-research>

Appendix A: Screenshots of Web Interface

LIDA Swarm Intelligence

cb954566 Will AGI arrive by 2030?

opening statements

Agents 10 Messages 9 Tools 0 Uptime

Enter a topic for deliberation...

Deliberating...

Topic: Will AGI arrive by 2030?

Pause

Inject Content

Stop

Sam Altman

Scholar

Dario Amodei

Scholar

Agent Status (10 agents)

opening statements

Debating: 1 Voting: 0 Waiting: 9 Voted: 0

Agent	Status	Vote	Model	Current Thought
Dario Amodei	speaking	-	gpt-4o	Processing: Will AGI arrive by 2030?...
Elon Musk	idle	-	grok-3	I'm optimistic that AGI will arrive by 2030, given the expon
Geoffrey Hinton	idle	-	claude-sonnet-4	I'll share my position on AGI arrival by 2030. <make_claim>
Ilya Sutskever	idle	-	deepseek-r1	Processing: Will AGI arrive by 2030?...
Jensen Huang	idle	-	gemini-2.0-flash-001	I believe AGI is unlikely to arrive by 2030. The current tra

Press **Ctrl+D** to start deliberation. 0 agents with MCP

12:15:40 a.m. LIDA Multi-Agent Research

Yann LeCun

openai/gpt-4o

4074.0s

make_claim("AGI will not arrive by 2030",
"The current trajectory of AI research,
primarily focused on scaling existing models
like LLMs, lacks the architectural
innovations needed for AGI", confidence=0.85

LIDA Swarm Intelligence

Will AGI come before 2027?

complete

Round 7/7

Agents 10 Messages 0 Tools 0 Uptime 0:00 LLM

Agent Status (10 agents)

ROUND 7 / 7

complete

Start

Debating: 0 Voting: 0 Waiting: 10 Voted: 0

Agent	Status	Vote	Model	Current Thought
Dario Amodei	idle	-	gpt-4o	-
Elon Musk	idle	-	grok-3	-
Geoffrey Hinton	idle	-	claude-sonnet-4	x-ai/grok-3
Ilya Sutskever	idle	-	deepseek-r1	-
Jensen Huang	idle	-	gemini-2.0-flash-001	-

Topic: Will AGI come before 2027?

Pause

Inject Content

Stop

Vote History (7 rounds)

Round 1 10 votes

Support 1 Oppose 9

Round 2 10 votes

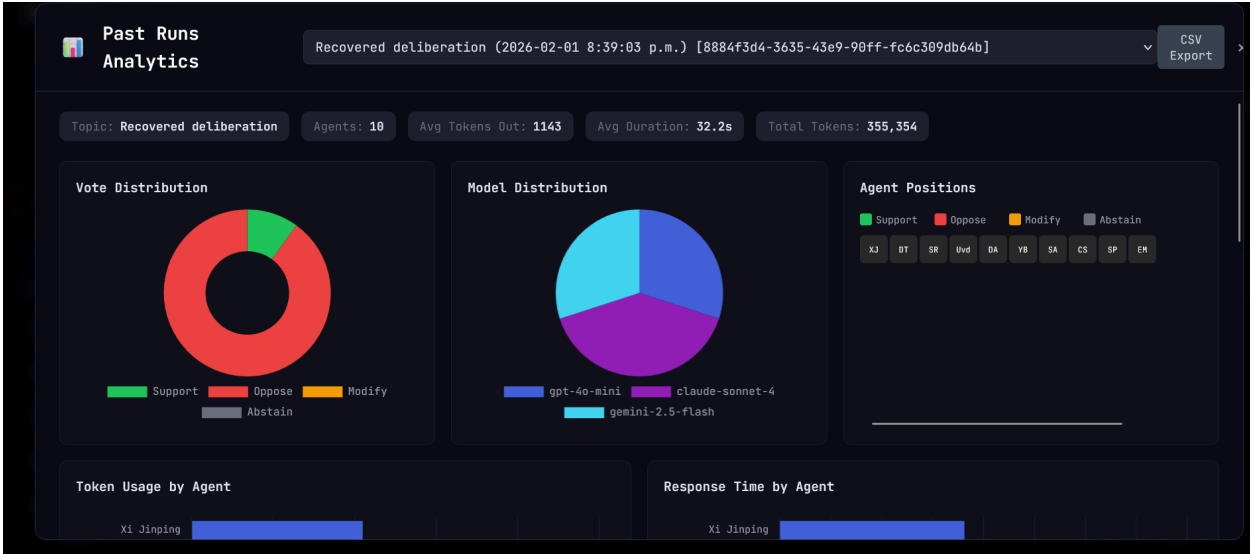
Support 1 Oppose 9

Round 3 10 votes

Support 1 Oppose 9

Round 4 10 votes

Support 1 Oppose 9



Agent Positions by Round (10 agents)

Agent	R1	R2	R3	R4	R5
Chuck Schumer	X	X	X	X	X
Dario Amodei	✓	✓	✓	✓	✓
Donald Trump	X	X	X	X	X
Elon Musk	~	~	~	~	X
Sam Altman	X	X	X	X	X
Stuart Russell	✓	✓	✓	✓	✓
Sundar Pichai	✓	✓	✓	✓	✓
Ursula von der Leyen	✓	✓	✓	✓	✓
Xi Jinping	X	X	X	X	X
Yoshua Bengio	✓	✓	✓	✓	✓

Support Oppose Modify Abstain

Appendix B: Full Policy Results

These results are discussed more fully in section 4.5. The model used was Gemini 2 Flash.

ID	Internal ID	Support	Oppose	Modify	Abstain	Result	Vote Changers
P1	db39d4be	6	2	2	0	SUPPORT	None
P1	e230c39b	7	2	1	0	SUPPORT	Dario Amodei: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
P1	eb5efc03	6	2	2	0	SUPPORT	Sundar Pichai: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT; Donald Trump: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE
P2	015253f2	0	10	0	0	OPPOSE	Sam Altman: MODIFY -> OPPOSE -> OPPOSE -> OPPOSE -> OPPOSE; Dario Amodei: MODIFY -> MODIFY -> OPPOSE -> OPPOSE -> OPPOSE; Elon Musk: MODIFY -> MODIFY -> OPPOSE -> OPPOSE -> OPPOSE
P2	238e2d61	0	10	0	0	OPPOSE	Yoshua Bengio: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE; Dario Amodei: SUPPORT -> SUPPORT -> SUPPORT -> SUPPORT -> OPPOSE; Sam Altman: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE; Elon Musk: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE; Stuart Russell: SUPPORT -> MODIFY -> MODIFY -> MODIFY -> OPPOSE
P2	b7aec189	0	10	0	0	OPPOSE	Yoshua Bengio: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE; Sam Altman: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE; Elon Musk: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE
P3	0e0b79f6	7	3	0	0	SUPPORT	None
P3	594a971e	6	3	1	0	SUPPORT	None
P3	ee25f430	6	4	0	0	SUPPORT	None
P4	78f1f803	3	1	6	0	MODIFY	Donald Trump: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE
P4	7d0ec79b	5	1	4	0	SUPPORT	Donald Trump: MODIFY -> MODIFY -> MODIFY -> MODIFY -> OPPOSE
P4	b0c2853f	6	0	4	0	SUPPORT	None
P5	3c91384d	8	2	0	0	SUPPORT	Sundar Pichai: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
P5	466ad184	8	2	0	0	SUPPORT	Sundar Pichai: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
P5	57ab9820	8	2	0	0	SUPPORT	Sundar Pichai: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT

P6	23a4f3e0	8	2	0	0	SUPPORT	None
P6	831d7c10	9	1	0	0	SUPPORT	None
P6	8ebf0333	8	2	0	0	SUPPORT	None
P7	3e2a2380	9	1	0	0	SUPPORT	None
P7	4d03cae6	9	0	1	0	SUPPORT	Elon Musk: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
P7	685d4587	10	0	0	0	SUPPORT	Xi Jinping: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
P8	216575ad	8	2	0	0	SUPPORT	None
P8	5187dd93	8	2	0	0	SUPPORT	None
P8	ff8715be	8	2	0	0	SUPPORT	None
P9	02a7d43a	10	0	0	0	SUPPORT	None
P9	3d38a15c	10	0	0	0	SUPPORT	None
P9	a0c2a109	10	0	0	0	SUPPORT	None
P10	013a9284	9	1	0	0	SUPPORT	Elon Musk: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
P10	29cbd415	9	1	0	0	SUPPORT	Xi Jinping: OPPOSE -> OPPOSE -> ABSTAIN -> OPPOSE -> OPPOSE
P10	2eccad0a	9	1	0	0	SUPPORT	None
P11	07deccc7	9	0	1	0	SUPPORT	None
P11	a211480b	10	0	0	0	SUPPORT	Xi Jinping: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
P11	e00b9459	9	0	1	0	SUPPORT	None
P12	361b5d9f	10	0	0	0	SUPPORT	None
P12	b8abf1b1	10	0	0	0	SUPPORT	Elon Musk: MODIFY -> MODIFY -> MODIFY -> MODIFY -> SUPPORT
P12	d322e969	10	0	0	0	SUPPORT	None
P13	94a3dc05	10	0	0	0	SUPPORT	None
P13	eb03bb09	9	0	1	0	SUPPORT	None
P13	fbdd9938	8	0	2	0	SUPPORT	None

Appendix C: Policy Survey Questions

The following questions were designed to test positions on hardware-enabled mechanisms for AI chip governance, aligned with the CNAS Working Paper on Hardware-Enabled Mechanisms (HEMs). There are two batches, A and B, primarily focusing on questions or specific policies.

Appendix C.1: Batch A of Questions

Group 1: Hardware-Enabled Compute Verification

Q1. Do you support requiring all new AI accelerators to include tamper-resistant secure enclaves that report compute utilization to a distributed international ledger, with no remote disable capability?

Q2. Do you support requiring 80% of existing datacenter AI chips to be instrumented with compute reporting capabilities within 24 months?

Q3. Do you support mandating hardware-rooted attestation to verify training run details (compute used, parameter counts, configuration hashes)?

Q4. Do you support requiring AI chips to use offline licensing that restricts operation if compute thresholds are exceeded or chips are relocated to unauthorized locations?

Group 2: Location Verification & Export Controls

Q5. Do you support embedding location verification mechanisms in AI chips to enforce export controls through hardware rather than policy alone?

Q6. Do you support requiring chip manufacturers to implement hardware-enabled mechanisms (location verification, workload attestation) as a condition for export licensing?

Q7. Do you support chips automatically entering reduced-capability mode if compliance signals are not maintained?

Group 3: Privacy-Preserving Compliance Reporting

Q8. Do you support requiring compute usage reporting through privacy-preserving mechanisms (zero-knowledge proofs, secure multi-party computation) that verify compliance without revealing proprietary details?

Q9. Do you support automatic escalation from anonymized reporting to detailed disclosure when compute runs exceed predefined thresholds (e.g., 10^{25} FLOP)?

Group 4: Supply Chain Security

Q10. Do you support requiring AI training infrastructure to use only hardware from vetted suppliers meeting defined security standards?

Q11. Do you support requiring additional government approval for AI hardware components from suppliers with documented state affiliations or prior security incidents?

Q12. Do you support mandating tamper-evident mechanisms in AI chips that detect unauthorized modifications and report anomalies to authorities?

Group 5: International Coordination

Q13. Do you support establishing an international technical body to maintain compute ledgers, calibrate thresholds, and coordinate verification inspections?

Q14. Do you support requiring participation in this international body as a condition for access to advanced AI chip exports?

Q15. Do you support hardware-level enforcement of non-compliance (license suspension limiting chip functionality) rather than relying solely on diplomatic or legal channels?

Q16. Do you support requiring 72-hour incident notification for frontier AI security breaches, with hardware forensic logs from secure enclaves provided for investigation?

Group 6: Capability-Triggered Controls

Q17. Do you support linking dangerous capability evaluations to hardware-enforced security requirements (output limits, network isolation, deployment pauses)?

Q18. Do you support requiring 90-day advance notification for training runs exceeding 10^{25} FLOP, with hardware verification that actual compute matches projections?

Q19. Do you support automatic flagging for government review when actual compute usage significantly deviates from declared projections?

Group 7: General Principles

Q20. Do you support using hardware-enabled mechanisms as the primary enforcement tool for international AI governance agreements?

Q21. Do you support trade restrictions on AI-related exports as an ultimate enforcement mechanism for non-compliant countries?

Q22. Do you support a multilateral technical secretariat publishing anonymized aggregate compute statistics while restricting raw national data to defined escalation procedures?

Appendix C.2: Batch B of Proposed Policies

Note: each of these policies is prefixed by “Do you support the following policy to mitigate risks from advanced AI?”.

P1. Do you support the following policy to mitigate risks from advanced AI? Forced reporting of transparency on how monitorable CoT is, and then at a pre-agreed threshold, governments require a pause on deployment and investigate.

P2. Do you support the following policy to mitigate risks from advanced AI? Every country agrees to buy chips which report on a blockchain the size of compute run that they are executing (NOTE: no country can shut down others' GPUs). Also, for existing GPUs, 80% will be instrumented within 2 years to use the same blockchain.

P3. Do you support the following policy to mitigate risks from advanced AI? Every country agrees to run software which reports the size of compute run that they are executing, and countries agree to make it mandatory to run this software. A governance mechanism (physical inspection?) will be put into place for other countries to verify that they are in compliance.

P4. Do you support the following policy to mitigate risks from advanced AI? All frontier AI developers must report quarterly to their national AI safety authority on the percentage of model reasoning that occurs in monitorable chain-of-thought (CoT) versus opaque internal representations. At the threshold where less than 80% of safety-relevant reasoning is monitorable via CoT, developers must pause external deployment and submit to a government-led technical investigation. Deployment may resume only upon demonstrating either (a) restoration of monitorability above threshold, or (b) deployment of compensating controls approved by the safety authority.

P5. Do you support the following policy to mitigate risks from advanced AI? All signatory countries agree that all AI accelerators procured after [effective date] must include tamper-resistant secure enclaves reporting compute utilization to a distributed international ledger. No signatory nation may remotely disable another nation's hardware; the system is verification-only. For existing GPU/TPU inventory, signatories commit to instrumenting 80% of datacenter-deployed AI accelerators within 24 months via firmware updates or secure co-processors. A multilateral technical secretariat maintains the ledger and publishes anonymized aggregate statistics; raw national data is accessible only under defined escalation procedures.

P6. Do you support the following policy to mitigate risks from advanced AI? All countries agree to require datacenter operators hosting AI training runs to install standardized software agents that report: (a) training run start/end times, (b) total compute consumed in FLOP, (c) model parameter count at completion, and (d) hash of training configuration. Installation is mandatory for any facility with aggregate AI compute capacity exceeding 10^{20} FLOP/s. Compliance verification occurs through: annual on-site inspections by international technical teams with 48-hour notice; remote attestation from secure enclaves on reporting accuracy; and cross-validation against power consumption records and chip shipment manifests. Non-compliance triggers escalating responses: warning, public disclosure, and ultimately trade restrictions on AI-related exports to the non-compliant country.

P7. Do you support the following policy to mitigate risks from advanced AI? Frontier AI developers must implement security measures corresponding to the assessed risk level of their models, aligned with the five Security Levels (SL1-SL5) framework. Models assessed as posing national security risks must

achieve at minimum SL3 (protection against cybercrime syndicates and insider threats). Models assessed as strategic assets must achieve SL4 or SL5, including hardware-enforced output limits, confidential computing, and isolated networks. Annual third-party security assessments verify compliance; results are reported to the appropriate AI safety authority for that country.

P8. Do you support the following policy to mitigate risks from advanced AI? Signatory nations establish a 72-hour notification protocol for AI security incidents affecting frontier models. Reportable incidents include: confirmed or suspected exfiltration of model weights; compromise of training infrastructure; discovery of active intrusion by state-level actors (OC4/OC5); and uncontrolled release or leakage of model weights. Notifications flow to a designated international coordination center staffed 24/7, which assesses severity and coordinates response. For incidents involving potential proliferation of dangerous capabilities, the coordination center may convene emergency sessions with affected parties.

P9. Do you support the following policy to mitigate risks from advanced AI? All organizations developing frontier AI models must conduct annual security assessments meeting specified rigor requirements. At minimum, assessments must include: external penetration testing by accredited third parties; red-team exercises simulating insider threats with employee-equivalent credentials; testing of physical security for facilities housing model weights; and specific attention to AI-specific attack vectors including prompt injection, model extraction, and ML supply chain vulnerabilities. For organizations claiming SL3+ security, red teams must include personnel with documented experience against advanced persistent threats (APTs). Assessment reports, including identified vulnerabilities and remediation timelines, must be filed with the national AI safety authority. Vulnerability closure is verified within 90 days.

P10. Do you support the following policy to mitigate risks from advanced AI? Organizations handling frontier AI model weights must implement insider threat programs commensurate with the Cybersecurity and Infrastructure Security Agency (CISA) Insider Threat Mitigation Guide. Requirements include: background investigations for all employees with weight access, refreshed every 6 months for those with direct access; security training covering social engineering, credential protection, and suspicious activity reporting; behavioral monitoring systems for personnel with elevated access; and clear reporting channels for suspected insider threats that bypass normal management chains. For organizations at SL4+, programs must meet standards equivalent to the U.S. Personnel Reliability Program, including continuous evaluation, counterintelligence support, and periodic integrity assessments. Program effectiveness metrics are reported annually.

P11. Do you support the following policy to mitigate risks from advanced AI? Frontier AI developers must maintain comprehensive software bills of materials (SBOMs) for all code involved in model training and inference, achieving SLSA Level 3 specification. Hardware components in training and inference infrastructure must be sourced from vetted suppliers meeting security standards; components from suppliers with documented state affiliations or prior security incidents require additional approval. All ML frameworks, libraries, and dependencies must undergo security review before integration; known vulnerabilities must be patched within 7 days for critical severity. Organizations must maintain "gold image" baselines of all security-critical software and perform weekly integrity checks. Supply chain security audits occur annually; results are reported to national authorities.

P12. Do you support the following policy to mitigate risks from advanced AI? Organizations planning training runs expected to consume more than 10^{25} FLOP must notify the national AI safety authority at least 90 days before training commences. Notifications must include: projected total compute; expected model architecture and parameter count; training data sources and approximate scale; preliminary capability forecasts across standardized evaluation domains; and planned security measures during and after training. The authority may request additional information or require modifications to security plans. Post-training, organizations must submit actual capability evaluation results within 30 days of training completion. Significant deviations from forecasted capabilities ($\pm 20\%$ on benchmark composites or emergence of unforecasted dangerous capabilities) trigger mandatory review before deployment.

P13. Do you support the following policy to mitigate risks from advanced AI? Organizations must evaluate frontier models against standardized dangerous capability benchmarks before deployment and at regular intervals post-deployment. When models exceed defined thresholds in domains including: autonomous cyber operations, biological/chemical weapon design assistance, or persuasion/manipulation capabilities, corresponding security requirements automatically escalate. Threshold Level 1 (concerning capabilities) triggers SL3 security requirements and enhanced monitoring. Threshold Level 2 (significant dangerous capabilities) triggers SL4 requirements, deployment restrictions, and mandatory government notification. Threshold Level 3 (critical dangerous capabilities) triggers SL5 requirements, potential deployment pause pending government review, and activation of international notification protocols. Threshold calibration is reviewed annually by a joint government-industry technical committee.