Building Global Support for AI Governance Evidence from Six Countries

Alexander Kuo[1], Aina Gallego[2] and Shir Raviv[3]

[1]Department of Politics and International Relations and Christ Church, Oxford University;, Oxford, OX1 3UQ, UK

[2]Department of Political Science, Constitutional Law and Philosophy of Law,, University of Barcelona, 08034, Spain

[2]Institut Barcelona d'Estudis Internacionals;, Barcelona, 08005, Spain

[4]Data Science Institute, Columbia University;, New York, NY 10027, US

Abstract

The rapid advancement of AI presents unprecedented challenges requiring international coordination, yet efforts to establish global governance frameworks remain fragmented. An important yet understudied factor shaping governments' willingness to negotiate, approve, and enforce international agreements is domestic support. We examine how key institutional features of AI governance shape public support for international arrangements using new data from a large-scale survey experiment conducted in the US, China, India, Germany, the UK, and Japan. Our results indicate that citizens worldwide are willing to accept—and even prefer—inclusive, enforceable, and neutrally led regulations, suggesting that an ambitious global governance framework of AI could receive broad public support. The findings highlight governance mechanisms that generate global support and areas of cross-national disagreement that may require tailored approaches.

The rapid advances in AI pose significant global challenges that transcend national borders, requiring international coordination to govern the technology, unlock its huge potential, and safeguard human rights [1–3]. In response to this need, the past few years have seen a rise in international initiatives aimed at governing AI, such as the OECD AI Principles, the UNESCO Recommendations on the Ethics of Artificial Intelligence, the UN's exploration of a Global Digital Compact, and the Council of Europe's Convention on AI [4, 5]. Yet, despite these efforts, the current AI governance regime remains highly fragmented and contested [6]. Some initiatives lack key participants (e.g., major AI-developing states like China)[7], while others struggle to define their scope or establish mechanisms for enforcement [8, 9]. A central question of this century will be: Can countries effectively cooperate to govern this powerful technology?

We focus on an important, yet often overlooked, aspect of the feasibility of AI governance frameworks: the role of public opinion. Evidence from other policy issues requiring international cooperation (e.g., global agreements regarding climate change mitigation) demonstrates that public opposition can undermine even technically and legally sound governance arrangements [10–12]. In both democratic and autocratic systems, public opinion can constrain governments' ability to maneuver and negotiate effectively [13–16]. However, despite the importance of citizen support in shaping the feasibility of international cooperation [10–12] and for compliance with regulatory frameworks [17], we know surprisingly little about what citizens around the world want when it comes to global AI governance. What kind of AI governance frameworks do they prefer? To what extent do these preferences vary across different national contexts?

Existing research has predominantly focused on public perceptions and concerns regarding specific AI risks, such as algorithmic bias, job displacement, and privacy violations [18, 19]. The few studies that explore mass opinion on AI regulation focus primarily on actions implemented at national or domestic levels [20, 21], while the international dimension of regulation of AI remains largely unexplored. In particular, we lack a comprehensive assessment that captures which global governance frameworks AI citizens are willing to accept and implement.

To systematically evaluate public preferences across a range of potential governance frameworks and isolate the impact of specific institutional features, we designed a conjoint experiment and embedded it in a large-scale online public opinion survey in six countries—the US, China, India, Germany, the UK, and Japan—that are key players in the global AI landscape, representing diverse political systems, economic development levels, and over 75% of global AI investment [22]

Experiment Design

Conjoint experiments enable researchers to disentangle how individuals weigh distinct attributes of a policy proposal [23–25] in a manner that is highly representative of their actual policy choices [26]. In our experiment, we presented respondents with two hypothetical proposals for a global governance framework for AI, asked them to indicate which of the two they would prefer their country to adopt, and to rate each proposal separately. Each respondent evaluated three pairs of frameworks, generating 45,135 total evaluations. Supplementary Fig.

SI-1 shows the conjoint instructions along with an example of pair profiles.

We focus on four key dimensions that represent central challenges in designing effective international agreements and are critical to the current debate over the feasibility of global AI governance and, based on previous literature on international cooperation, are likely to be salient considerations for voters.

- ***Number of Participating Countries***. This dimension captures the tension between inclusivity and effectiveness in global cooperation [4, 5, 27]. Broader participation can enhance perceived legitimacy, but may also complicate negotiations and dilute enforcement. We varied the scope of participation from 40 to 160 out of 195 countries, reflecting debates about whether governance frameworks should be highly inclusive or limited to key players.

- ***Leading Actors***. Current AI governance initiatives range from broad multilateral efforts to narrower state-led or bilateral cooperation. Prior research suggests that leadership structure influences perceptions of legitimacy and trustworthiness [28, 29]. To assess the weight of this dimension, we included five potential leading actors to represent a mix of major AI powers (US, China), regional representation (EU, India/Brazil), and a multilateral approach (UN).

- ***Enforcement Mechanisms***. A defining feature of any international agreement is its enforcement mechanism [8]. Current AI governance frameworks range from non-binding guidelines to stringent sanctioning regimes [30]. Strong enforcement can increase compliance, but may raise concerns about national sovereignty[30, 31]. We explore citizens' sensitivities to this tension by varying enforcement models ranging from voluntary recommendations to a global court enforcing sanctions with the power to impose legally binding penalties on non-compliant countries.

- ***Issue Area Coverage***. Research on climate and trade agreements shows that public support varies depending on whether policies target specific domains or take comprehensive approaches [32]. Drawing on this literature, we varied the focus of the agreement between regulating general AI risk (referring to broad safety concerns and potential existential threats) and five specific domains identified as salient to voters in previous studies: AI-based autonomous weapons, AI-generated misinformation, discrimination by AI algorithms, job loss from AI automation, and AI threats to privacy and data protection [33, 34].

By randomizing attribute values across these dimensions, we can assess to what extent a specific attribute affects public support for global AI governance. Table 1 describes the exact wordings of each attribute.

To address concerns that respondents might rely on simple shortcuts in conjoint experiments, we randomized the order of attributes between respondents while keeping the order consistent within each respondent's three tasks. SI discusses the conjoint design and the survey procedure in detail. The experiment was approved by the University of Oxford ethics board.

Table 1: Conjoint Attribute Values

Attribute	Values
Number of participating countries	160 out of 195
	120 out of 195
	80 out of 195
	40 out of 195
Actors writing proposal	India and Brazil lead
	EU leads
	China leads
	US leads
	UN leads
Type of agreement	Enforcement by a global court using sanctions
	Enforcement by a group of countries using fines
	Enforcement by each country's own government
	No enforcement, only guidelines
Agreement Focus	AI threat to privacy and data protection
	Job loss by AI automation
	Discrimination by AI algorithms
	AI-generated misinformation
	AI-based Autonomous weapons
	General risk

Note: This table reports the attribute values for each dimension of the experiment. Respondents were presented with two randomly generated proposals for comparison.

Results

We begin by analyzing preferences in our pooled sample of respondents across all surveyed countries (N=15,045). Using a linear probability model, we estimate Average Marginal Component Effects (AMCEs) to assess how different institutional features affect respondent choices between AI governance frameworks. The AMCEs represent the average change in the likelihood of a proposal being selected when a particular attribute level is included compared to a baseline level, holding all other attributes constant. Our dependent variable is whether a respondent prefers a given proposal over its alternative. All models include country-fixed effects and cluster standard errors by respondent. Fig. 1 shows the results.[1]

The results reveal that institutional design significantly shapes preferences for global AI governance. In line with our expectations, broader international participation (80 to 160 countries vs. 40) increases the chance a proposal is favored by 3 to 5 percentage points, representing a relative increase of 6-10% in the likelihood of a proposal being favored. The strong preference for broad participation suggests that citizens value inclusive frameworks, contrasting with initiatives like the G7's Hiroshima AI Process that restrict participation to advanced economies [35].

We find particularly large effects regarding leadership structure. UN-led frameworks are strongly preferred over those led by individual states or regional blocs. Frameworks led by China or the US incur substantial penalties (-8.4, -4.5, and -2.0 percentage points,

[1]We replicate the results using alternative measures of the outcome variable, including proposal ratings (dichotomized at different thresholds) and using weighted data to account for sample composition by country. Table SI-3 shows that the results remain substantively similar across these specifications.

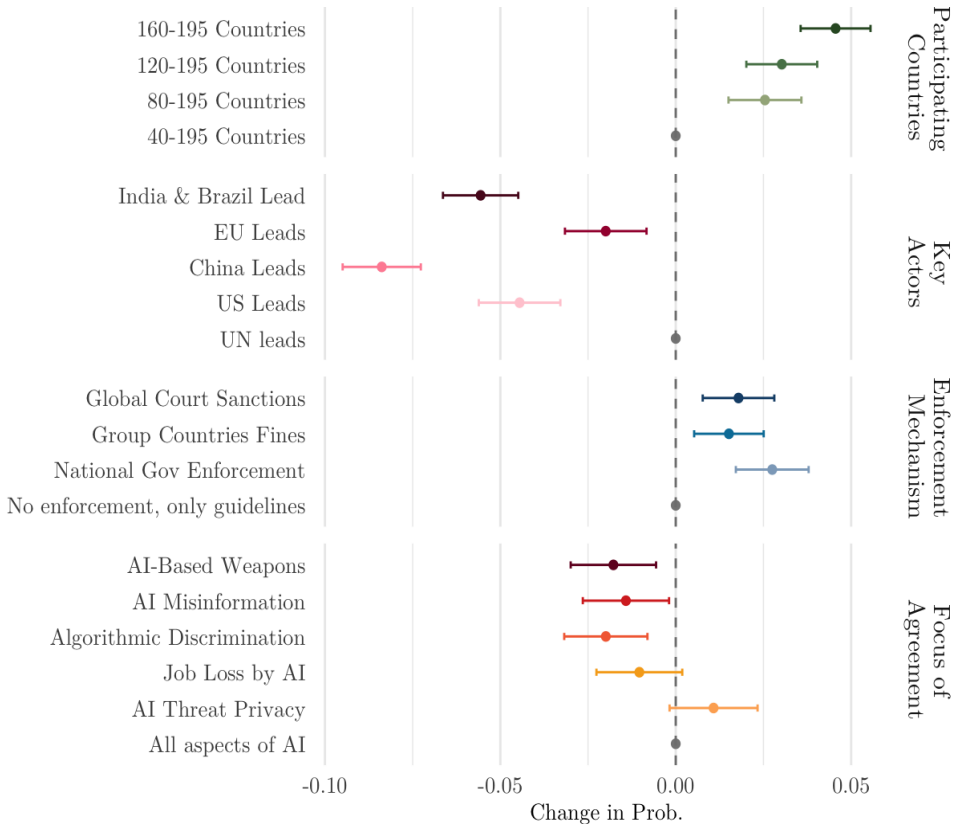


Figure 1: Estimated effects of institutional features on support for adopting AI global governance framework, pooled sample. Points represent average marginal component effects (AMCEs) with 95% confidence intervals (see SI Table SI-3 for the full results).

respectively), suggesting public skepticism of governance arrangements that might entrench particular state interests.

Enforcement mechanisms also matter for proposal support. Proposals with national government enforcement, global court sanctions, or group-imposed fines are all significantly more likely to be chosen than those with only voluntary guidelines ($p < 0.001$). While the differences between these enforcement types are modest, the consistent positive effects indicate a preference for frameworks with concrete implementation measures.

Finally, on issue coverage, citizens favor comprehensive frameworks over those targeting specific issues. Proposals focusing on AI weapons, algorithmic discrimination, or misinformation are less likely to be selected than those addressing "general AI risk" ($p < 0.01$). Only privacy-focused frameworks show a positive yet small effect (1.1 percentage points, $p < 0.05$).

Given the complexity of conjoint tasks, one concern is that respondents may rely on readily available heuristics, particularly for attributes like the number of participating countries or the leading actor, which are more easily grasped than other institutional features. To address this concern, Fig. SI-2 shows results conducted separately for attentive and non-attentive respondents. The results are very much consistent across these groups, suggesting that our findings reflect considered preferences for AI governance rather than reliance on simple heuristics.

Cross-National Patterns

We next examine how these preferences vary across countries, estimating separate linear probability models for each. Fig. 2 shows the results. To get a better sense of the underlying preferences over attributes within countries, we also estimate the marginal means across countries [36] (see SI-3).[2]

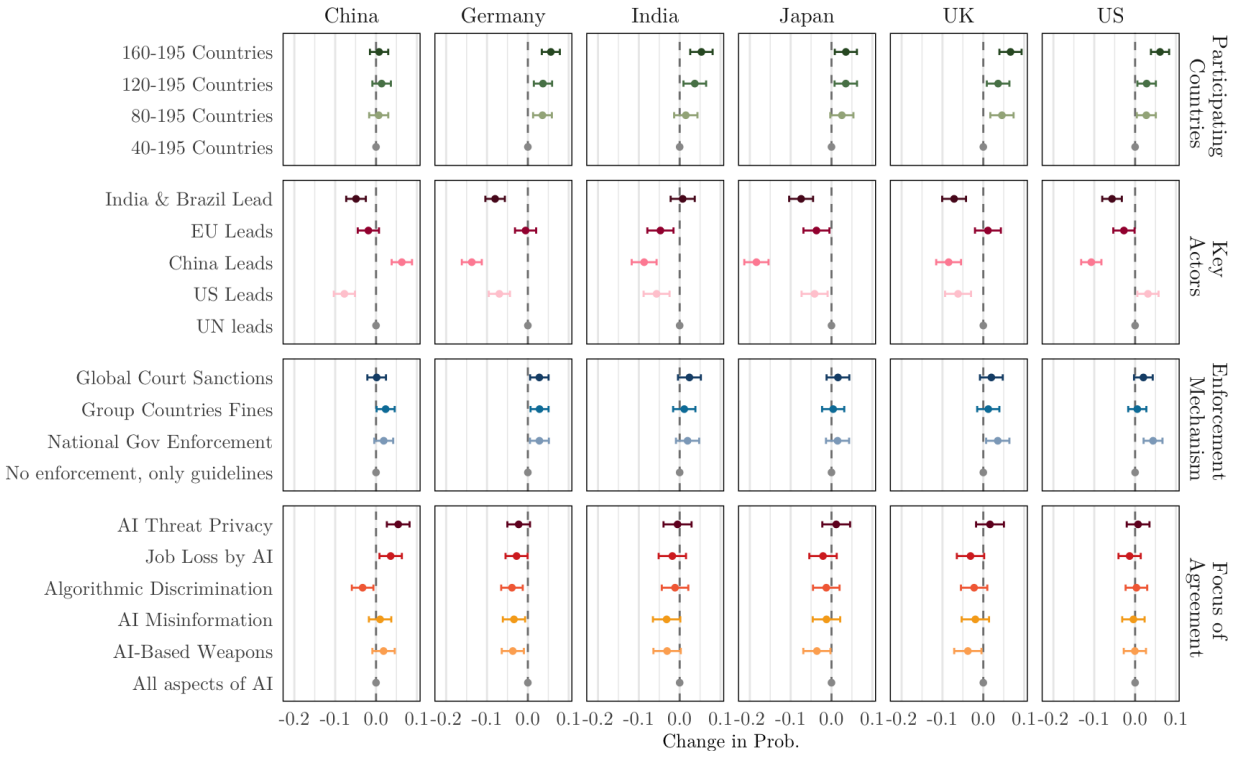


Figure 2: Estimated effects of institutional features on support for adopting AI global governance framework, by country. Points represent average marginal component effects (AMCEs) with 95% confidence intervals. Effects are estimated using LPM regression models with standard errors clustered at the respondent level (see Supplementary Table SI-4 for the full results).

The preference for broader participation, while present globally, is significantly stronger in democracies. For instance, increasing participation from 40 to 160 countries boosts the likelihood of a proposal being favored by 6.7 percentage points in the UK and 6.1 in the US (Fig. 2), a pattern clearly visible in the marginal means as well (SI-3). In contrast, the effect is substantially weaker in China.

The leadership dimension reveals even stronger cross-national variation. Respondents generally favor their own country's leadership, while reacting negatively to leadership by

[2]The partner firm of the survey vendor in China required adjustment to the survey instrument. To field the survey there, we excluded questions on political ideology, voting behavior, and government performance evaluations. Importantly, these restrictions did not affect our conjoint experiment. However, potential self-censorship, even within anonymous formats, warrants consideration when interpreting responses from China. Despite this caveat, including China, a key actor in the global AI landscape, offers valuable insights, particularly given the limited availability of representative public opinion data from China, a key actor in the global AI landscape.

other major powers. We find that respondents generally favor proposals led by their own country, while expressing skepticism toward proposals led by major powers like the US and China. Despite this own-country preference, UN leadership stands out as a consensus options, ranking as the first or second most preferred option across all six countries surveyed. This reinforces our finding in the pooled sample about the appeal of neutral international bodies, potentially reflecting a desire to keep AI governance free from geopolitical entanglements.

Despite these variations, the preference for robust enforcement mechanisms is consistent across countries. All nations show a statistically significant preference for frameworks with enforcement—whether through national government enforcement, global court sanctions, or group-imposed fines—over those relying solely on voluntary guidelines. The marginal means in Figure SI-3 underscore this consistency, with all enforcement options generally above the 0.5 mark across all country panels.

The pooled finding that enforcement mechanisms are preferred over voluntary guidelines holds true across all countries (Figure SI-3). Similarly, the general preference for comprehensive frameworks over those targeting specific issues is broadly consistent, though with subtle national variations. For instance, privacy concerns appear slightly more pronounced in Germany and the UK, while proposals on AI-based weapons elicit relatively less negative reactions in China and the US, perhaps reflecting fewer concerns about the consequences of geostrategic regions.

To address potential order effects, we replicated our analysis using only each respondent's first conjoint task, finding substantively similar results (see Table SI-5).

Overall, while we observe notable cross-national differences in the magnitude of preferences for certain attributes, particularly regarding leadership, the general pattern of support for inclusive governance frameworks with robust enforcement mechanisms remains consistent across countries.

Implications

To assess the substantive implications of the results and make comparisons more concrete, we perform two exercises. First, we estimate support for proposals that align with two existing governance initiatives using them as real-world benchmarks (OECD AI Principles and EU AI Act). Second, we present the combination of attributes of proposals that are at the 1st, 25th, 75th and 99th percentile of the distribution of estimated support.

One prominent initiative, the OECD AI Principles, adopted in 2019, represents a relatively cautious approach, characterized by limited participation (40 countries), UN-backed leadership, and voluntary guidelines. This framework offers a comprehensive framework covering values-centered principles for trustworthy AI, including inclusive growth, human-centered values, transparency, robustness, and accountability [37]. Our analysis estimates public support for such a framework at only 48.6%. The EU AI Act, finalized in December 2023, takes a more forceful approach, introducing binding regulations and concrete enforcement mechanisms, albeit within a limited geographical scope. We estimate support for this type of framework at 51.0%, suggesting that stronger enforcement, even when regionally limited, may enhance public acceptance.

Our next set of results suggest that far greater public support could be achieved by

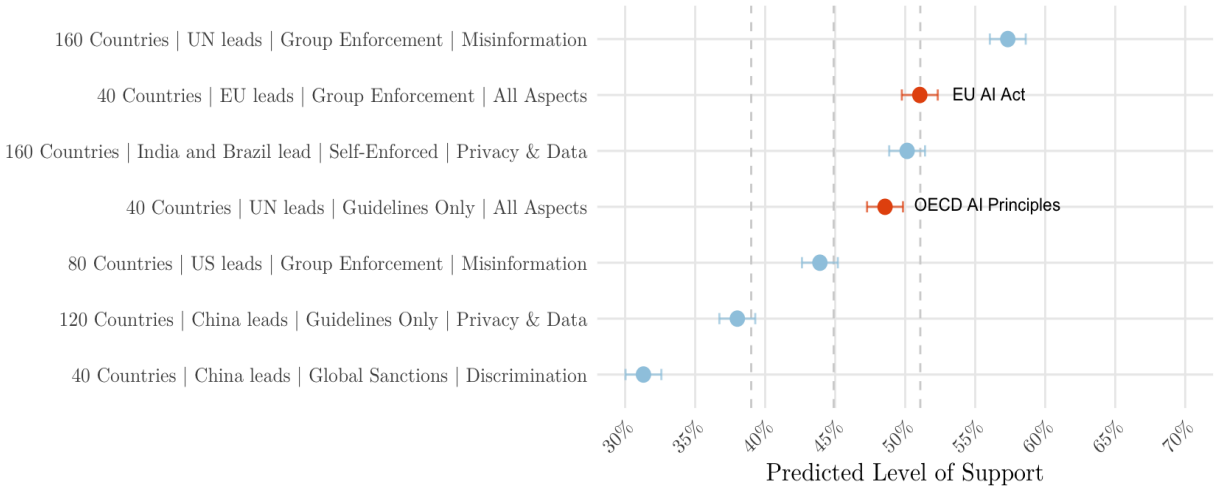


Figure 3: Predicted support for various AI governance frameworks. Horizontal lines indicate 95 percent robust confidence intervals. The blue estimates refer to the framework roughly at the 1st, 25th, 75th and 99th percentile of support level. The orange dots refer to the frameworks that correspond most closely to the actual global initiatives to govern AI.

adopting more ambitious frameworks. Fig. 3 presents the estimated level of support for a set of selected AI governance frameworks that differ with respect to various attributes and correspond to the 1st, 25th, 75th and 99th percentile of the distribution of estimated support, with 95 percent confidence intervals.

A framework that combines broad participation (160 countries), UN leadership, and enforcement through a global court with the power to impose sanctions is predicted to achieve 61.9% support. This configuration, corresponding to the 99th percentile of estimated support in our analysis, in stark contrast to frameworks resembling the G7's more exclusive approach, which garnered support near the 1st percentile.

These findings challenge the prevailing narrative that emphasizes national sovereignty and geopolitical competition as insurmountable barriers to global AI governance [38]. Instead, our data suggest that citizens worldwide are willing to accept—and even prefer—inclusive, enforceable, and neutrally led regulations. This support extends to robust enforcement mechanisms, such as sanctions imposed by a global court or fines levied by a group of countries, indicating a surprising degree of public acceptance for binding international oversight of AI.

By elucidating the factors that shape public preferences, we provide an empirical foundation for understanding the political feasibility of different approaches to global AI governance. By understanding these preferences, policymakers can design more effective and legitimate governance frameworks that have a greater chance of securing broad public support.

References

References

[1] Anu Bradford. *Digital empires: The global battle to regulate technology*. Oxford University Press, 2023.

[2] Justin B Bullock et al. *The Oxford handbook of AI governance*. Oxford University Press, 2024.

[3] Esmat Zaidan and Islam Abouelmagd Ibrahim. AI governance in a complex and rapidly changing regulatory landscape: A global perspective. *Humanities and Social Sciences Communications*, 11(1):1–18, 2024.

[4] Université de Montréal. Déclaration de montréal IA responsable. www.montrealdeclaration-responsibleai.com, 2017. Accessed: 2024-01-11.

[5] Future of Life Institute. Asilomar AI principles. https://futureoflife.org/open-letter/ai-principles, 2017. Accessed: 2024-01-11.

[6] Lewin Schmitt. Mapping global AI governance: a nascent regime in a fragmented landscape. *AI and Ethics*, 2(2):303–314, 2022.

[7] Peter Cihon, Matthijs M Maas, and Luke Kemp. Fragmentation and the future: investigating architectures for international AI governance. *Global Policy*, 11(5):545–556, 2020.

[8] Michael Veale, Kira Matus, and Robert Gorwa. AI and global governance: modalities, rationales, tensions. *Annual Review of Law and Social Science*, 19(1):255–275, 2023.

[9] Jonas Tallberg et al. The global governance of artificial intelligence: Next steps for empirical and normative research. *International Studies Review*, 25(3):viad040, 2023.

[10] Nina Tannenwald. The nuclear nonproliferation regime as a "failed promise": Contestation and self-undermining dynamics in a liberal order. *Global Studies Quarterly*, 4(2):ksae025, 2024.

[11] Alexander Thompson. Contestation and resilience in the liberal international order: The case of climate change. *Global Studies Quarterly*, 4(2):ksae011, 2024.

[12] Italo Colantone and Piero Stanig. The surge of economic nationalism in Western Europe. *Journal of Economic Perspectives*, 33(4):128–151, 2019.

[13] Michael R Tomz and Jessica LP Weeks. Public opinion and the democratic peace. *American Political Science Review*, 107(4):849–865, 2013.

[14] Réka Juhász and Nathan Lane. The political economy of industrial policy. *Journal of Economic Perspectives*, 38(4):27–54, 2024.

[15] Devin Caughey and Christopher Warshaw. *Dynamic democracy: Public opinion, elections, and policymaking in the American States*. University of Chicago Press, 2022.

[16] Ric Neo and Chen Xiang. State rhetoric, nationalism and public opinion in China. *International Affairs*, 98(4):1327–1346, 2022.

[17] Markus Anderljung et al. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

[18] Baobao Zhang and Allan Dafoe. Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874*, 2019.

[19] Beatrice Magistro, Sophie Borwein, R Michael Alvarez, Bart Bonikowski, and Peter J Loewen. The common microfoundations of attitudes toward artificial intelligence (ai) and globalization. *Available at SSRN 4795006*, 2024.

[20] Tobias Heinrich and Christopher Witko. Self-interest and preferences for the regulation of artificial intelligence. *Journal of Information Technology & Politics*, pages 1–16, 2024.

[21] Bartosz Wilczek, Sina Thäsler-Kordonouri, and Maximilian Eder. Government regulation or industry self-regulation of AI? investigating the relationships between uncertainty avoidance, people's AI risk perceptions, and their regulatory preferences in Europe. *AI & SOCIETY*, pages 1–15, 2024.

[22] Raymond Perrault and Jack Clark. Artificial intelligence index report 2024. Technical report, Stanford Institute for Human-Centered Artificial Intelligence, 2024.

[23] Kirk Bansak, Jens Hainmueller, Daniel J Hopkins, Teppei Yamamoto, James N Druckman, and Donald P Green. Conjoint survey experiments. *Advances in experimental political science*, 19:19–41, 2021.

[24] Kirill Zhirkov. Estimating and using individual marginal component effects from conjoint experiments. *Political Analysis*, 30(2):236–249, 2022.

[25] Kirk Bansak, Jens Hainmueller, and Dominik Hangartner. How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science*, 354(6309):217–222, 2016.

[26] Jens Hainmueller, Dominik Hangartner, and Teppei Yamamoto. Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8):2395–2400, 2015.

[27] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[28] Jonas Tallberg and Michael Zürn. The legitimacy and legitimation of international organizations: Introduction and framework. *The Review of International Organizations*, 14:581–606, 2019.

[29] Michael Zürn. *A theory of global governance: Authority, legitimacy, and contestation*. Oxford University Press, 2018.

[30] Tanja A Börzel and Michael Zürn. Contestations of the liberal international order: From liberal multilateralism to postnational liberalism. *International Organization*, 75(2):282–305, 2021.

[31] Catherine E De Vries, Sara B Hobolt, and Stefanie Walter. Politicizing international cooperation: The mass public, political entrepreneurs, and political opportunity structures. *International Organization*, 75(2):306–332, 2021.

[32] Thomas Bernauer and Robert Gampfer. How robust is public support for unilateral climate policy? *Environmental Science & Policy*, 54:316–330, 2015.

[33] Michael C Horowitz. When speed kills: Lethal autonomous weapon systems, deterrence and stability. In *Emerging technologies and international stability*, pages 144–168. Routledge, 2021.

[34] Zhi Li et al. Global digital compact: A mechanism for the governance of online discriminatory and misleading content generation. *International Journal of Human–Computer Interaction*, pages 1–16, 2024.

[35] European Commission. G7 leaders' statement on the hiroshima ai process, 2023.

[36] Thomas J Leeper, Sara B Hobolt, and James Tilley. Measuring subgroup preferences in conjoint experiments. *Political Analysis*, 28(2):207–221, 2020.

[37] Huw Roberts, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. Global ai governance: barriers and pathways forward. *SSRN Electronic Journal*, 2024.

[38] Armin Von Bogdandy, Matthias Goldmann, and Ingo Venzke. From public international to international public law: Translating world public opinion into international public authority. *European Journal of International Law*, 28(1):115–145, 2017.

Supplementary Materials

A Data

Between June and August 2024, we administered surveys to a total of 15,045 respondents across six countries: China (N=3,001), Germany (N=3,030), India (N=2,001), Japan (N=2,006), the United Kingdom (N=2,007), and the United States (N=3,000). Respondents were recruited by Respondi, an international survey firm, which conducted sampling to match the known population marginals on socio-demographic and regional variables.

A.1 Descriptive Statistics

Tables SI-1 and SI-2 report the distribution of key variables used in our analysis across countries as well as aggregate descriptive statistics of our cross-national sample.

Table SI-1: Demographic Characteristics by Country

Variable	China	Germany	India	Japan	UK	US	All
Sample Size	3001	3030	2001	2006	2007	3000	15045
Male (%)	48.1	50.0	52.4	49.9	50.4	50.0	50.0
Female (%)	51.9	50.0	47.6	50.1	49.6	50.0	50.0
18-29 years (%)	16.4	18.7	47.9	15.8	15.6	24.3	22.4
30-39 years (%)	23.9	19.8	21.6	17.7	19.5	19.0	20.4
40-49 years (%)	20.7	17.3	20.1	22.4	19.8	19.8	19.9
50-59 years (%)	24.9	23.4	6.8	24.0	23.2	19.8	20.8
60+ years (%)	14.2	20.9	3.5	20.1	21.9	17.2	16.5
University Degree (%)	19.1	33.3	13.0	46.9	34.7	38.0	30.7

A.2 Question Wording

Below we detail the exact question wording and coding procedures for all variables used in the analyses.

Table SI-2: Digital Literacy and Policy Preferences by Country

Variable	China	Germany	India	Japan	UK	US	All
High Digital Literacy (%)	71.9	32.3	76.8	17.7	46.8	51.7	50.0
LLM User (%)	75.7	29.1	80.9	26.7	19.0	25.7	42.9
High AI Knowledge (%)	34.6	21.4	60.6	3.6	17.0	28.8	27.8
Anti-Regulation (%)	86.6	45.9	57.6	64.9	44.4	54.5	59.6
Pro-Market (%)	87.7	54.0	63.6	42.1	49.4	49.5	58.9
Pro-Trade (%)	24.5	17.8	28.5	21.3	35.6	16.6	23.2
Pro-Climate Action (%)	52.4	34.0	47.2	27.3	41.8	31.2	39.0

A.2.1 Demographic Variables

- **Gender:** "You are..."
  - Response options: Male (1), Female (2), Other (3)
  - Recoded for analysis: Male, Female (excluding Other)
- **Age:** From provided age, respondents were grouped into five categories:
  - 18-29 years
  - 30-39 years
  - 40-49 years
  - 50-59 years
  - 60+ years
- **Education:** Country-specific education measures were harmonized into a binary indicator for university degree:
  - US: 4-year college degree or higher
  - UK: University diploma or higher
  - Germany: Universitätsabschluss or higher
  - Japan: University or higher
  - India: Bachelor or higher
  - China: Bachelor or higher

A.2.2 Digital and AI-Related Variables

- **Digital Literacy:** "How familiar are you with the following computer and Internet-related items?"
  - Items: Browser cookies, Chat GPT, Hotspots, Firewalls, Cloud, RSS
  - Response scale: 1 (Totally unfamiliar) to 5 (Very familiar)

  - Index created using Principal Component Analysis
  - Binary indicator: Above median (High Digital Literacy) vs Below median

- **LLM Usage:** "Have you used large language models like ChatGPT for your work?"
  - Response options: No (1), Yes, occasionally (2), Yes, frequently (3)
  - Binary indicator: Any usage (Yes) vs No usage

- **AI Knowledge:** "How much have you heard or read about AI?"
  - Response options: A lot (1), Somewhat (2), A little (3), Not at all (4)
  - Binary indicator: High awareness (A lot) vs Other responses

A.2.3 Policy Preferences

- **Anti-Regulation:** "Government regulation usually does more harm than good."
  - Scale: Strongly agree (1) to Strongly disagree (4)
  - Binary indicator: Agree/Strongly agree (1) vs Others (0)

- **Pro-Market:** "Generally firms should be left alone by the government to freely compete."
  - Scale: Strongly agree (1) to Strongly disagree (4)
  - Binary indicator: Agree/Strongly agree (1) vs Others (0)

- **Pro-Trade:** "Should the government increase or decrease trade barriers such as tariffs on imports?"
  - Scale: Greatly increase (1) to Greatly decrease (5)
  - Binary indicator: Decrease/Greatly decrease (1) vs Others (0)

- **Pro-Climate:** "Should the government increase or decrease taxes on the use of fossil fuels?"
  - Scale: Greatly increase (1) to Greatly decrease (5)
  - Binary indicator: Increase/Greatly increase (1) vs Others (0)

A.3 Conjoint Instructions and questionnaire

- **Instructions** "There is current debate about how countries and organizations around the world can work together to manage AI's development and use. Please consider two options for a global agreement on the development and use of AI. Please read each alternative carefully. You will assess a pair of competing proposals 3 times."

Figure SI-1: Interface

Concepts	Option 1	Option 2
Number of participating countries	160 out of 195	120 out of 195
Key actors writing the proposal	China	The United States
Type of agreement	Global court enforces a treaty by sanctions	Multilateral agreements with fines for violating countries
Focus of global agreement	User data collection to generate AI	All aspects of AI

Notes: The vertical dashed line at 0.5 serves as a reference point for assessing whether a given attribute level garners majority support.

- **Preference** "Which option would you prefer your country to adopt?"
  - Response options: Option 1; Option 2
- **Ranking Proposal 1** "Please rate whether you support or oppose proposal option 1."
  - Response options: Strongly support; Somewhat support; Neither support nor oppose; Somewhat oppose; Strongly oppose
- **Ranking Proposal 2** "Please rate whether you support or oppose proposal option 1."
  - Response options: Strongly support; Somewhat support; Neither support nor oppose; Somewhat oppose; Strongly oppose

B Additional Results

B.1 Results from pooled data

Table SI-3 reports results from four linear probability models. The first two models analyze a binary choice outcome, with Model 1 using unweighted data and Model 2 incorporating survey weights for countries. The weighted specification employs post-stratification weights constructed using population distributions for three demographic dimensions: age (18-39, 40-59, 60+), gender, and education level.

We also replicate our main analysis using an alternative outcome that captures whether a respondent generally supported or opposed a proposal. Our alternative measure is based on respondents' ratings of each proposal. We dichotomize this item into an indicator of support (4 or 5) or strongly support (5). Model 3 uses a binary indicator for supportive rankings, coded as 1 for the two highest support categories on a 5-point scale,: while Model 4 focuses on active support, coded as 1 only for the highest support category.

Since the experiment was fully randomized and the assignment of conjoint attributes balanced across respondents of each survey country, we do not present results with individual-level covariates.

All four specifications reveal consistent patterns in the direction of effects.

Table SI-3: Main Results - AMCE Estimates, Pooled sample

	Preference (Unweighted)	Preference (Weighted)	Support (Unweighted)	Strongly support (Unweighted)
	(1)	(2)	(3)	(4)
80-195 Countries	0.025***	0.024***	0.037***	0.015***
	(0.005)	(0.006)	(0.004)	(0.003)
120-195 Countries	0.030***	0.031***	0.054***	0.029***
	(0.005)	(0.006)	(0.004)	(0.003)
160-195 Countries	0.046***	0.050***	0.076***	0.037***
	(0.005)	(0.006)	(0.005)	(0.003)
US Leads	−0.045***	−0.053***	−0.095***	−0.037***
	(0.006)	(0.007)	(0.005)	(0.004)
China Leads	−0.084***	−0.097***	−0.171***	−0.027***
	(0.006)	(0.007)	(0.005)	(0.004)
EU Leads	−0.020***	−0.028***	−0.035***	−0.022***
	(0.006)	(0.007)	(0.005)	(0.004)
India/Brazil Lead	−0.056***	−0.065***	−0.141***	−0.044***
	(0.005)	(0.007)	(0.005)	(0.003)
Gov't Enforcement	0.028***	0.029***	0.069***	0.027***
	(0.005)	(0.007)	(0.005)	(0.003)
Group Fines	0.015***	0.015**	0.060***	0.030***
	(0.005)	(0.006)	(0.004)	(0.003)
Global Court Sanctions	0.018***	0.017***	0.057***	0.027***
	(0.005)	(0.006)	(0.005)	(0.003)
AI Weapons	−0.018***	−0.010	−0.077***	−0.020***
	(0.006)	(0.007)	(0.005)	(0.004)
AI Misinformation	−0.014**	−0.018**	−0.049***	−0.017***
	(0.006)	(0.008)	(0.005)	(0.004)
AI Discrimination	−0.020***	−0.016**	−0.059***	−0.035***
	(0.006)	(0.007)	(0.006)	(0.004)
AI Job Loss	−0.010*	−0.002	−0.043***	−0.020***
	(0.006)	(0.008)	(0.005)	(0.004)
AI Privacy Threat	0.011*	0.011	0.012**	0.003
	(0.006)	(0.008)	(0.006)	(0.004)
China	−0.00000	−0.00004	0.118***	0.063***
	(0.0003)	(0.0003)	(0.006)	(0.005)
Germany	0.00001	−0.0001	−0.076***	−0.051***
	(0.0003)	(0.0004)	(0.007)	(0.005)
India	0.00003	0.0001	0.165***	0.152***
	(0.0003)	(0.0004)	(0.008)	(0.007)
Japan	−0.0001	−0.00004	−0.118***	−0.079***
	(0.0003)	(0.0004)	(0.008)	(0.005)
UK	−0.00004	−0.0001	−0.046***	−0.029***
	(0.0003)	(0.0004)	(0.008)	(0.006)
Constant	0.509***	0.513***	0.477***	0.128***
	(0.007)	(0.008)	(0.008)	(0.006)
N	90,270	90,270	90,270	90,270
R^2	0.005	0.006	0.065	0.051
Adjusted R^2	0.005	0.006	0.065	0.051

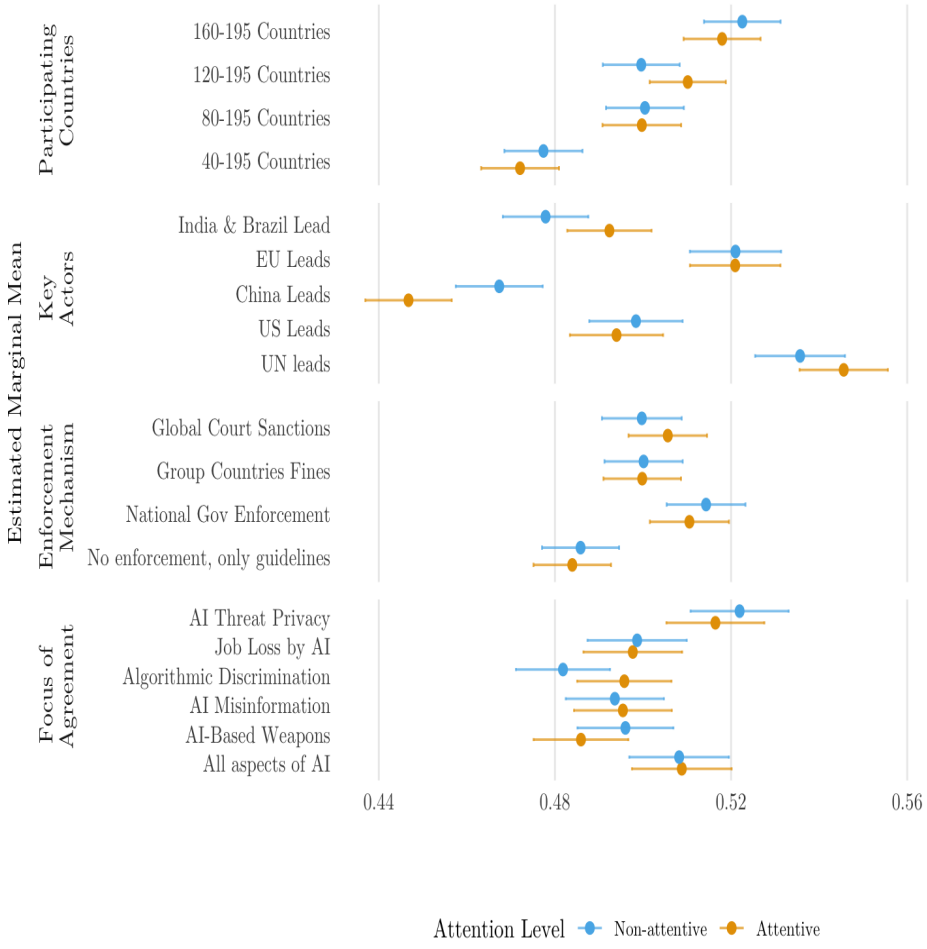
Notes: Cluster-robust standard errors in parentheses. Country fixed effects included but not shown. Reference categories are: 40 out of 195 countries (participating countries), UN leads (key actors), no enforcement-only guidelines (enforcement mechanism), and all aspects of AI (the focus of agreement) *p<0.05; **p<0.01; ***p<0.001

B.2 Respondent Attentiveness

One concern in conjoint experiments is that the complex nature of the task may encourage respondents to rely on easily accessible heuristics – such as the number of participating countries or the identity of the leading actor – rather than engaging deeply with other, potentially less salient, institutional features like the specifics of enforcement or issue focus. If this were the case, we would expect to see significant differences between attentive and non-attentive respondents on these more complex attributes. To examine this, we conducted a robustness check based on respondent attentiveness. we define attentiveness using the time taken to complete the conjoint exercise, with non-attentive respondents being those who spent time below the 25th percentile and attentive respondents those who spent longer.

Figure SI-2 presents the marginal means for each attribute level, estimated separately for attentive and non-attentive subgroups. While some minor differences in the magnitude of preferences are observable–for instance, attentive respondents show a slightly stronger aversion to China-led frameworks–the overall patterns are remarkably consistent. Crucially, we find only small, non-significant differences between the two groups regarding their preferences for enforcement mechanisms and the specific issue focus of the agreement. Both groups exhibit similar levels of support for different enforcement types and various issue focuses. This lack of significant variation on these more intricate attributes suggests that our main findings are unlikely to be driven by respondents simply relying on heuristics based on more easily processed attributes like the number of countries or the leading actor. Instead, the consistency across groups lends confidence to the robustness of our main conclusions.

Figure SI-2: Estimated marginal means by attentiveness, pooled sample



Notes: This figure shows the estimated marginal means with 95% confidence intervals. The vertical dashed line at 0.5 serves as a reference point for assessing whether a given attribute level garners majority support. Attentive respondents are those who spent an above-median time on the conjoint.

B.3 Cross-National Differences in Preferences

Table SI-4: Support for AI Regulation Proposals by Country

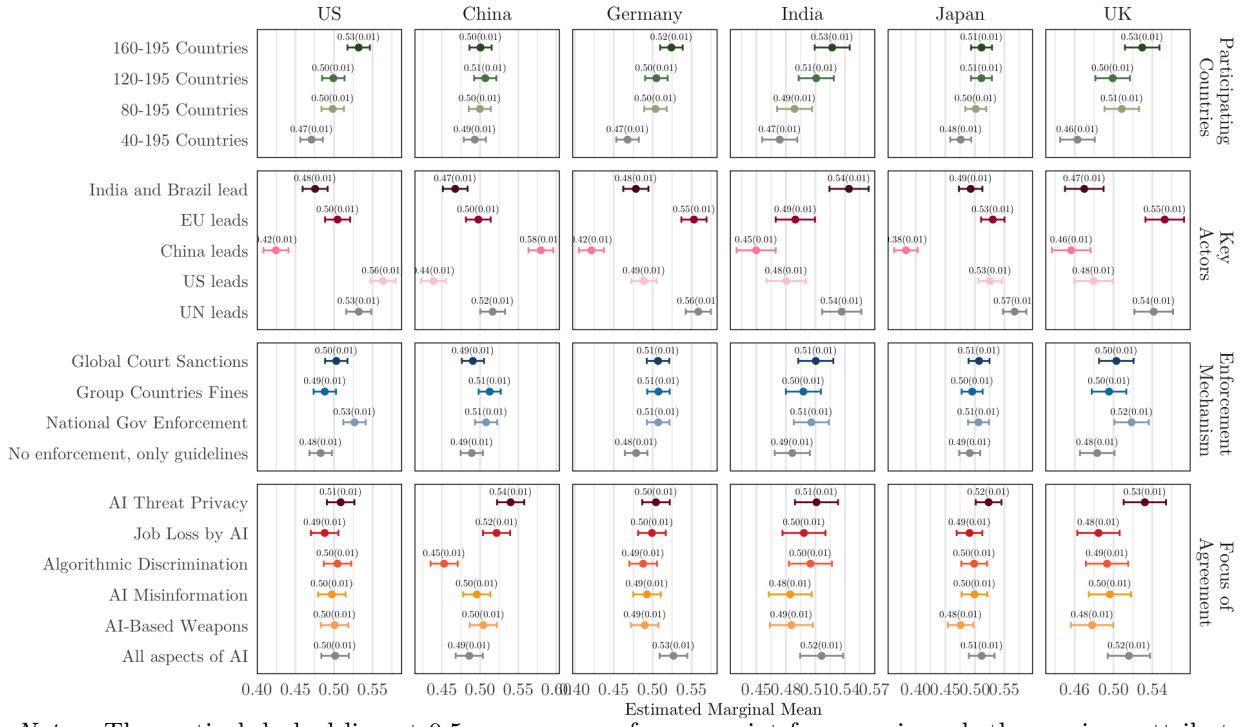
	AMCEs on Preference for International Proposal					
	US	UK	Germany	Japan	India	China
	(1)	(2)	(3)	(4)	(5)	(6)
80-195 Countries	0.027*	0.046**	0.036**	0.025	0.015	0.007
	(0.012)	(0.015)	(0.012)	(0.014)	(0.015)	(0.012)
120-195 Countries	0.028*	0.036*	0.037**	0.035*	0.037**	0.014
	(0.012)	(0.014)	(0.011)	(0.014)	(0.014)	(0.012)
160-195 Countries	0.061***	0.067***	0.056***	0.035*	0.053***	0.007
	(0.011)	(0.014)	(0.011)	(0.014)	(0.014)	(0.011)
US Leads	0.031*	−0.062***	−0.070***	−0.041*	−0.057***	−0.077***
	(0.013)	(0.016)	(0.013)	(0.016)	(0.016)	(0.013)
China Leads	−0.107***	−0.085***	−0.137***	−0.184***	−0.087***	0.063***
	(0.013)	(0.016)	(0.012)	(0.015)	(0.016)	(0.013)
EU Leads	−0.027*	0.011	−0.006	−0.037*	−0.047**	−0.019
	(0.013)	(0.016)	(0.013)	(0.016)	(0.016)	(0.013)
India/Brazil Lead	−0.057***	−0.071***	−0.080***	−0.074***	0.007	−0.049***
	(0.012)	(0.015)	(0.012)	(0.015)	(0.015)	(0.012)
Gov't Enforcement	0.044***	0.035*	0.028*	0.015	0.019	0.019
	(0.012)	(0.015)	(0.012)	(0.014)	(0.015)	(0.012)
Group Fines	0.005	0.012	0.029*	0.004	0.011	0.024*
	(0.011)	(0.014)	(0.011)	(0.014)	(0.014)	(0.011)
Global Court Sanctions	0.020	0.020	0.028*	0.016	0.024	0.002
	(0.012)	(0.014)	(0.012)	(0.014)	(0.014)	(0.012)
AI Weapons	−0.001	−0.038*	−0.037**	−0.036*	−0.031	0.018
	(0.014)	(0.017)	(0.014)	(0.017)	(0.017)	(0.014)
AI Misinformation	−0.004	−0.019	−0.034*	−0.012	−0.032	0.010
	(0.014)	(0.017)	(0.014)	(0.017)	(0.017)	(0.014)
AI Discrimination	0.003	−0.022	−0.039**	−0.013	−0.011	−0.033*
	(0.013)	(0.016)	(0.013)	(0.017)	(0.017)	(0.014)
Job Loss by AI	−0.013	−0.031	−0.028*	−0.021	−0.018	0.036*
	(0.014)	(0.017)	(0.014)	(0.017)	(0.017)	(0.014)
AI Privacy Threat	0.007	0.016	−0.022	0.012	−0.005	0.054***
	(0.014)	(0.017)	(0.014)	(0.017)	(0.018)	(0.014)
Constant	0.487***	0.503***	0.532***	0.546***	0.513***	0.484***
	(0.016)	(0.019)	(0.015)	(0.019)	(0.019)	(0.016)
N	18,000	12,042	18,180	12,036	12,006	18,006
R^2	0.012	0.011	0.013	0.018	0.008	0.013
Adjusted R^2	0.011	0.009	0.013	0.016	0.006	0.012

Notes: Reference categories: 40 participating countries, UN leadership, no enforcement (guidelines only), and all aspects of AI. Cluster-robust standard errors in parentheses (clustered by respondent). *p < .05; **p < .01; ***p < .001.

Figure SI-3 reports marginal means to compare countries instead of using AMCEs reported in Figure ??. This method allows for a direct comparison of preference levels across countries without the confounding influence of the chosen reference category (Leeper, Hobolt, and Tilley 2020).

The marginal means represent, for each country and each attribute level, the average probability that respondents would select a proposal with that particular level, holding the other features at all possible values (due to randomization). Whereas the AMCEs capture relative differences (e.g., how shifting from "40 participating countries" to "160 participating countries" changes support), the marginal means convey absolute levels of approval.

Figure SI-3: Marginal means of policy features on voter preference, by country



1 *Notes*: The vertical dashed line at 0.5 serves as a reference point for assessing whether a given attribute level garners majority support.

B.4 Addressing Potential Order Effects

To address potential order effects in which respondents' choices might be influenced by the sequence of conjoint tasks, we replicated the analysis using data from only the first task each respondent completed. By analyzing only the first task, we ensure that each respondent's choices reflect their initial, unbiased assessment of the presented AI governance framework. Table SI-5 reports the results. The results are very similar to those reported in Table SI-4, which uses data from all three tasks. Across all six countries, the direction and statistical significance of the estimated effects remain largely consistent when using only the first task.

Table SI-5: Support for AI Regulation Proposals by Country (First Task Only)

	Forced Choice Preference					
	US	UK	Germany	Japan	India	China
	(1)	(2)	(3)	(4)	(5)	(6)
80-195 Countries	0.038*	0.065**	0.090***	0.070**	0.069**	0.028
	(0.019)	(0.024)	(0.019)	(0.022)	(0.024)	(0.020)
120-195 Countries	0.084***	0.111***	0.108***	0.079***	0.122***	0.059**
	(0.020)	(0.024)	(0.020)	(0.023)	(0.025)	(0.020)
160-195 Countries	0.134***	0.168***	0.154***	0.124***	0.146***	0.066***
	(0.019)	(0.023)	(0.019)	(0.022)	(0.024)	(0.019)
US Leads	0.101***	−0.154***	−0.208***	−0.117***	−0.066*	−0.227***
	(0.021)	(0.026)	(0.021)	(0.026)	(0.027)	(0.021)
China Leads	−0.316***	−0.341***	−0.387***	−0.559***	−0.240***	0.167***
	(0.021)	(0.026)	(0.021)	(0.022)	(0.026)	(0.021)
EU Leads	−0.069**	0.021	0.029	−0.075**	−0.071*	−0.101***
	(0.023)	(0.026)	(0.021)	(0.026)	(0.028)	(0.023)
India/Brazil Lead	−0.187***	−0.241***	−0.264***	−0.296***	0.103***	−0.170***
	(0.021)	(0.025)	(0.021)	(0.025)	(0.026)	(0.022)
Gov't Enforcement	0.108***	0.112***	0.083***	0.043	0.059*	0.069***
	(0.020)	(0.024)	(0.019)	(0.023)	(0.024)	(0.020)
Group Fines	0.067***	0.109***	0.070***	0.072**	0.079**	0.083***
	(0.019)	(0.024)	(0.020)	(0.023)	(0.024)	(0.020)
Global Court Sanctions	0.088***	0.107***	0.093***	0.080***	0.065**	0.055**
	(0.019)	(0.024)	(0.019)	(0.022)	(0.024)	(0.019)
AI Weapons	−0.077**	−0.096***	−0.148***	−0.113***	−0.062*	−0.071**
	(0.024)	(0.029)	(0.023)	(0.028)	(0.029)	(0.024)
AI Misinformation	−0.057*	−0.075*	−0.079***	−0.032	−0.103***	−0.091***
	(0.024)	(0.029)	(0.024)	(0.028)	(0.029)	(0.024)
AI Discrimination	−0.065**	−0.046	−0.111***	−0.055*	−0.060*	−0.105***
	(0.023)	(0.028)	(0.023)	(0.028)	(0.029)	(0.023)
Job Loss by AI	−0.087***	−0.083**	−0.101***	−0.021	−0.070*	−0.019
	(0.023)	(0.028)	(0.023)	(0.028)	(0.028)	(0.023)
AI Privacy Threat	0.008	−0.011	−0.048*	0.039	−0.008	0.034
	(0.023)	(0.028)	(0.022)	(0.027)	(0.028)	(0.023)
Constant	0.512***	0.529***	0.602***	0.624***	0.471***	0.518***
	(0.027)	(0.033)	(0.026)	(0.032)	(0.033)	(0.027)
N	6,000	4,014	6,060	4,012	4,002	6,002
R^2	0.103	0.101	0.123	0.175	0.071	0.098
Adjusted R^2	0.101	0.097	0.121	0.172	0.068	0.096

Notes: Models estimated using data from the first task only. Reference categories: 40 participating countries, UN leadership, no enforcement (guidelines only), and all aspects of AI. Cluster-robust standard errors in parentheses (clustered by respondent). *p < .05; **p < .01; ***p < .001.