Building Support for Global AI Governance: Evidence from Six Countries

Aina Gallego* Alexander Kuo† Shir Raviv‡

December 3, 2025

Abstract

The rapid advancement of artificial intelligence (AI) creates transnational challenges and opportunities that demand coordinated global action, yet current governance efforts remain fragmented and contested. A crucial but understudied factor shaping states' willingness to negotiate, ratify, and enforce international agreements is domestic public support. We examine whether, and under what conditions, citizens are willing to support international frameworks for governing the development and use of AI. Using novel data from a conjoint experiment embedded in large-scale surveys conducted in the United States, China, India, Germany, the United Kingdom, and Japan, we find that majorities across these diverse contexts are willing to accept—and often prefer—governance frameworks that are inclusive, enforceable, and neutrally led. The findings reveal untapped potential for stronger international cooperation on AI and identify design features that can enhance legitimacy and cross-national support. They also illuminate how citizens evaluate emerging forms of global governance and provide a foundation for future research on preference formation in domains characterized by uncertainty and uneven distributional effects.

*Department of Political Science, Constitutional Law and Philosophy of Law, University of Barcelona; Institut Barcelona d'Estudis Internacionals.

†Department of Politics and International Relations and Christ Church, Oxford University.

‡The School of Political Science and International Relations, Tel Aviv University

Introduction

The rapid advances in AI pose significant challenges and opportunities that transcend national borders, requiring international coordination to govern the technology, unlock its huge potential, and safeguard human rights (Bradford 2023; Bullock et al. 2024; Zaidan and Ibrahim 2024). Scholars increasingly see today's rapid AI advances as a pivotal moment—one where international cooperation could avert long-term dangers, analogous to earlier efforts to manage other global risks such as nuclear proliferation and climate change. Recognizing the need for a coordinated global response, there has been a recent increase 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 (Université de Montréal 2017; Future of Life Institute 2017). Despite these efforts, the current AI governance regime remains highly fragmented and contested (Schmitt 2022; Juhász and Lane 2024a).[1] Some initiatives lack key participants (e.g., major AI-developing states like China)(Cihon, Maas, and Kemp 2020a), while others struggle to define their scope or establish mechanisms for enforcement (Veale, Matus, and Gorwa 2023). Moreover, competitive dynamics in the global AI race motivate some states to resist common regulatory standards in pursuit of technological advantage (Bradford 2023; Zaidan and Ibrahim 2024). Thus, it remains unclear if countries can effectively cooperate to govern this powerful technology.

We focus on an important, yet understudied, aspect of the *feasibility* of AI governance frameworks: public preferences over international arrangements. 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 arrangement (Tannenwald 2024; Thompson 2024; Colantone and Stanig 2019). In both democratic and autocratic systems, public opinion can constrain governments' ability to maneuver and negotiate effectively (Tomz and Weeks 2013; Juhász and Lane 2024b; Caughey and Warshaw 2022; Neo and Xiang 2022). Furthermore, existing evidence points to citizen consent and support of international regimes as being a key factor in their enforcement credibility and

1. By AI governance we mean the set of institutional arrangements that determine how artificial intelligence technologies are developed, deployed, and overseen.

success (Tallberg et al. 2023). However, despite the importance of citizen support in shaping the feasibility of international cooperation (Tannenwald 2024; Thompson 2024) and for compliance with regulatory frameworks (Anderljung et al. 2023), we know little about whether and under what conditions citizens are willing to support international rules that shape the development and use of AI. What kinds of governance frameworks would citizens want their governments to adopt? Which features are essential for securing their support? To what extent do priorities vary across national contexts?

To address these questions, we evaluate public preferences across a range of potential governance frameworks using a large-scale nationally representative surveys fielded in six countries—the US, China, India, Germany, the UK, and Japan—that are key players in the global AI landscape accounting for over 75% of global investment in AI (Perrault and Clark 2024). Specifically, we designed a conjoint experiment that randomized key policy features capturing core challenges in designing effective international responses to AI risks, as well as broader considerations of international cooperation that shape public perceptions of an agreement's legitimacy, effectiveness, and fairness.[2]

Our analysis provides three main findings. First, we find strong overall support for international governance of AI – citizens prefer–inclusive, enforceable, and neutrally led initiatives, suggesting that an ambitious global governance framework of AI could receive broad public support. Second, citizens tend to reject frameworks that are primarily led by either of the 'AI superpowers' (the US and China), with the exception of individuals in those countries. Third, most relevant AI-policy domains do not differ from one another in terms of attracting support. Our overall empirical results support the general intuition that enforceable, widely adopted and ambitious international regulation can garner support, but in the face of potential skepticism of those in the largest countries.

Beyond its practical relevance for international policy design, our findings contribute to the growing literature on the politics of AI by shifting the focus from perceptions of AI's societal risks to the institutional foundations of its global governance (Zhang and Dafoe 2019; Magistro et al. 2024). A nascent but growing body of work has examined public demand for government actions to address the risk of automation and labor-market disruption (Heinrich and Witko 2024;

2. This study was pre-registered prior to data collection (OSF Registration). An anonymized version of the pre-analysis plan is included in Appendix C

Wilczek, Thäsler-Kordonouri, and Eder 2024; Gallego and Kurer 2022). We extend this line of work by studying how citizens evaluate international governance arrangements across a range of AI applications, including autonomous weapons, AI-generated information, and algorithmic decision-making. Our results also have implications for the theoretical understanding of cooperation over international issues where there are multiple types of cooperation dilemmas (Nair and Peyton 2022; Bechtel, Scheve, and Lieshout 2022).

Theoretical Rationales for Global AI Governance

There is broad agreement among scholars that the unique risks and opportunities posed by AI call for international rules and institutions to guide its development and deployment (Acemoglu 2025; Roberts et al. 2024a; Zaidan and Ibrahim 2024; Veale, Matus, and Gorwa 2023; Erdélyi and Goldsmith 2018). This claim draws on a range of arguments that can be grouped into three broad rationales: (1) the global reach of AI's externalities, (2) the dangers of regulatory failure and geopolitical imbalance, and (3) the morality of ensuring that AI's benefits are distributed widely.

The first rationale concerns the cross-border nature of AI externalities. Because AI operates through global digital and security networks, the harms it generates can spread beyond the originating countries: malware can cascade through digital infrastructures, autonomous weapons deployed without safeguards can cause civilian casualties or escalate conflicts beyond national borders, and disinformation tools can destabilize political systems and undermine democratic processes abroad by flooding social media with fabricated content. At the extreme, some scholars warn of low-probability but catastrophic scenarios in which advanced AI systems operate in ways misaligned with human values, creating an existential risk for humanity with consequences on a global scale (Jones 2024). These threats, whether immediate or long-term, transcend national boundaries, making domestic regulation insufficient. As with other transnational challenges that involve externalities and underprovision of public goods, such as pandemics or climate change, they require collective international action and enforceable regulatory frameworks (Butcher and Beridze 2019).

The second rationale concerns geopolitical power imbalances and regulatory competition. Without global rules, powerful states could gain an overwhelming advantage in AI capabili-

ties, creating security dilemmas and triggering arms-race dynamics (Bengio 2023). As with non-proliferation treaties for weapons of mass destruction, international institutions can help contain these imbalances by establishing common constraints on AI development. In addition, governments face strong economic incentives to weaken AI regulations to attract investment and gain a competitive edge in technological development, fueling a regulatory "race to the bottom." This dynamic risks undermining human rights and reducing welfare, for example by eroding data privacy protections or enabling exploitative labor practices. Because some AI firms operate transnationally, national regimes alone cannot effectively address these pressures, underscoring the need for coordinated global governance (Schmitt 2022).

A third rationale for global AI governance shifts the focus from preventing harm to securing shared benefits (Tallberg et al. 2023). Without international coordination, a handful of advanced economies and dominant firms are likely to capture most of AI's economic and scientific gains, leaving other countries dependent on foreign providers and deepening global inequality. Global institutions can mitigate these imbalances by promoting technology transfer, supporting open scientific collaboration, and ensuring more equitable access to critical infrastructure. Scholars also argue that such cooperation can make AI development Pareto-improving, with gains from innovation distributed more broadly (Korinek, Schindler, and Stiglitz 2022). Clear and predictable international rules can also reduce regulatory fragmentation, lower barriers to cross-border investment, and build public trust, thereby fostering conditions for sustained innovation.

The Political Feasibility of Global AI Governance

Yet despite a growing scholarly consensus on the need for global AI governance (Dafoe 2018; Butcher and Beridze 2019; Taeihagh 2021), efforts to translate this vision into practice remain halting. Existing initiatives are fragmented and differ widely in scope and ambition. Some proposals advocate for a new international body with enforcement powers similar to the WTO (Erdélyi and Goldsmith 2018), while others support more modest coordination through existing multilateral forums like the G20 or OECD, often with little enforcement capacity (Cihon, Maas, and Kemp 2020a). The substantive focus also varies: some emphasize narrow national security concerns, while others envision broad, cross-sectoral governance. The goals likewise differ,

with some governments seeking to protect strategic advantage and others prioritizing safeguards against misuse and greater inclusion.[3]

These divergences reflect not only distinct strategic interests but also domestic political pressures that shape governments' priorities. Among these various pressures, public opinion remains consequential. Evidence from other domains of international cooperation—ranging from trade to climate change to public health—shows that mass preferences can shape both the design and the endurance of global governance institutions (Guisinger 2009; Kono 2008; Bernauer et al. 2016).

How, then, do citizens view the prospect of global rules for AI? We argue that the distinctive features of this emerging domain—its technical complexity, abstraction, and wide-ranging implications—make it particularly difficult for individuals to anchor their preferences in the substance of AI regulation itself. Instead, they are likely to rely on broader perceptions about global power dynamics, the legitimacy of international institutions, and what principles they believe should guide collective action across borders.

First, AI is technically complex and rapidly evolving (Büthe et al. 2022). Understanding how the technology functions and what effective regulation would entail requires expertise that most citizens do not possess.[4] Second, the implications of AI are often uncertain or perceived as distant. Risks such as mass job displacement or algorithmic discrimination may feel abstract or disconnected from daily life, which makes it harder for citizens to link them to specific regulatory needs. Finally, AI is a general-purpose technology that spans multiple societal domains. Attitudes toward its use vary sharply by application, with each context raising distinct considerations and trade-offs (Raviv 2025; Wenzelburger et al. 2024). The need to weigh such divergent judgments across domains makes it especially difficult for individuals to form informed views about what international rules for AI should look like.

Given these challenges, we suggest that citizens are unlikely to evaluate proposals for AI regulation on their substantive and technical merits. However this does not necessarily mean that people are indifferent to questions of AI governance or that their preferences lack coherence.

3. For instance, at the AI Action Summit held in Paris in February 2025, U.S. Vice President J.D. Vance underscored the need to maintain American leadership in AI, warning that excessive caution could stifle innovation and argued that heavy-handed regulation would hinder progress. By contrast, Indian Prime Minister Narendra Modi pressed for transparent global rules to mitigate the risks of unregulated AI and stressed that governance efforts must include the Global South.

4. Recent findings indicated that while most individuals recognize AI's growing presence in daily life, few fully understand its underlying mechanisms or the ethical implications involved (Ng et al. 2021).

Instead, they would be more likely to use broader heuristics, similar to a cumulated set of findings indicating that when confronted with complex political issues, individuals often remain rationally uninformed and rely on cues such as partisanship (Gabel and Scheve 2007; Druckman, Peterson, and Slothuus 2013). Yet, AI remains a novel policy domain without entrenched partisan divides. Recent evidence from the United States points to broad bipartisan support for oversight, both among elites and the mass public (Hatz et al. 2025; Faverio and Tyson 2023). In the absence of partisan guidance or clear personal stakes, citizens must turn to other considerations, relying on principles routinely applied to international cooperation more broadly: concerns about fairness in burden-sharing, the credibility of commitments, and the effectiveness of collective action. Studies across a range of issue areas–such as refugee resettlement, climate change mitigation, pandemic preparedness, and fiscal coordination–show that citizens' support for international agreements varies systematically depending on specific institutional features that speak to these core concerns (Bansak, Hainmueller, and Hangartner 2016; Bechtel and Scheve 2013; Bechtel, Hainmueller, and Margalit 2014; Bechtel, Scheve, and Lieshout 2019; Bechtel, Scheve, and Lieshout 2022). Drawing on this work, we focus on four institutional design features that are central to current debates on global AI governance–and likely to shape public support. Such features are of course not exhaustive, but we view them as key dimensions of any international proposal that would condition how citizens view global policy proposals.

Participation. A key design choice in international agreements is the number of countries involved. Broad participation can enhance legitimacy by signaling mutual commitment and fair burden-sharing. Evidence from several policy areas shows that public support increases when more countries join an agreement, suggesting that citizens are more likely to accept costs or constraints when they believe other countries are doing the same (Bechtel and Scheve 2013; Spilker, Bernauer, and Umaña 2018; Nair and Peyton 2022). However, inclusiveness can also limit effectiveness. Larger, more diverse coalitions must reconcile a wider range of interests, making it harder to reach agreement and can slow decision-making or weaken enforcement.[5] In politically contested issues, where countries dispute not just the appropriate response but the very existence or urgency of the problem, joining an agreement could carry symbolic weight. Joining indicates government acknowledgment and initial will to cooperate, even if an agreement

5. The 2015 Paris Agreement illustrates this trade-off. To bring the United States on board, negotiators scaled back the level of obligations for all countries (Cihon, Maas, and Kemp 2020b).

may have limited enforcement. In less polarized or salient issue areas like AI governance, that signal may matter less; citizens may care more about whether the agreement works than about who signs on. Overall, though, increasing participants may be a strong signal of the desirability of cooperation.

Leading Actors. The country leading an international agreement also should affect proposal support, as leadership structure shapes perceptions of legitimacy and trust (Tallberg and Zürn 2019; Dellmuth, Scholte, and Tallberg 2019; Zürn 2018). When evaluating complex international arrangements, citizens often rely on geopolitical heuristics to judge whether an agreement reflects their country's interests and values. Evidence indicates that public support drops when rival or less trusted countries lead negotiations–while trusted allies increase support (Gray and Hicks 2014; Ghassim, Koenig-Archibugi, and Cabrera 2022).[6] These patterns suggest that citizens use leadership as a proxy for whether an agreement will serve their country's interests, and this should matter as well for AI agreements.

In the case of AI governance, leadership may prove especially consequential given the technology's dual-use nature and its implications for both economic competitiveness and national security. Unlike global cooperation on climate change or refugee resettlement —which primarily focuses on mitigating shared risks or distributing burdens and costs —AI governance also involves managing and distributing enormous benefits. Countries that lead in AI shape access to cutting-edge innovation, digital infrastructure, and standard-setting processes. In this sense, AI governance more closely resembles trade agreements or nuclear energy cooperation where safety concerns are tightly bound up with geopolitical competition over critical resources. As in those domains, leadership by a rival power may heighten concerns of strategic disadvantage or biased allocation of benefits, while leadership by a trusted ally or multilateral body can signal fairness, impartiality, and a balanced approach to both risk management and opportunity sharing.

Enforcement Mechanisms. A defining feature of any international agreement is its enforcement mechanism (Veale, Matus, and Gorwa 2023). A large body of research shows that citizens across diverse contexts tend to prefer agreements with credible enforcement provisions (Gaikwad,

6. For instance, Nair and Peyton (2022) find that Americans reject redistributive agreements led by China or Russia while supporting those led by allied nations. In UN reform debates, proposals led by inclusive coalitions were seen as more legitimate than those controlled by a major power (Ghassim, Koenig-Archibugi, and Cabrera 2022).

Genovese, and Tingley 2025). Public support is stronger when rules are backed by independent monitoring and sanctions rather than left to voluntary compliance or self-reporting (Bechtel and Scheve 2013; Bechtel, Hainmueller, and Margalit 2014; Tingley and Tomz 2014). At the same time, strong enforcement can trigger concerns about sovereignty (De Vries, Hobolt, and Walter 2021); this is because citizens perceive the agreement as more likely to succeed with such enforcement. During the COVID-19 pandemic, for instance, public support for global coordination declined when enforcement appeared coercive or externally imposed—especially among individuals with strong nationalist orientations (Nair and Peyton 2022). Whether AI raises similar concerns remains an open question. On one hand, citizens may welcome enforcement as a sign that the agreement is credible and that other countries will be held accountable. On the other hand, they may see it as a threat to national control, particularly because AI regulation touches on sensitive domains such as defense, surveillance, data governance, and digital infrastructure —areas where states have strong incentives to retain discretion.[7]

Issue Area Coverage. Research on climate and trade agreements shows that public support can vary depending on whether policies target specific domains or adopt a more comprehensive approach (Bernauer and Gampfer 2015). In the context of AI, both patterns are plausible. Some survey data show that citizens are especially worried about tangible harms such as autonomous weapons, job displacement, and privacy violations, while expressing less concern about algorithmic bias or misinformation (Perrault and Clark 2024). These concerns can also reflect national context: citizens in security-oriented states may emphasize military risks, while those in liberal democracies may prioritize civil liberties and data protection. On the other hand, however, AI governance lacks the clear, unified goals that characterize other policies subject to international regulation. Each issue area involves distinct risks, actors, and regulatory instruments. Evaluating them requires making complex, domain-specific tradeoffs. As a result, individuals may struggle to weigh the relative importance of each domain and instead gravitate toward broadly scoped agreements that address AI risks in general. This tendency may be reinforced by dominant media narratives, which increasingly frame AI in terms of existential or systemic threats

7. AI's technical complexity presents distinctive enforcement challenges. Effective oversight demands specialized expertise to monitor opaque systems and trace accountability across public and private actors. Traditional enforcement models may struggle to meet these demands. As most citizens are unlikely to engage with the technical feasibility of enforcement, we suggest that they interpret enforcement mechanisms as cues about the agreement's credibility, the fairness of rule application, and the extent to which national control is maintained.

rather than specific policy challenges (Gilardi et al. 2024).

Taken together, we expect that support for global AI governance will depend more on the structural features of the agreement —its inclusiveness, leadership, enforcement, and broad issue scope —than on which specific issue area is being regulated.In the following sections, we provide a direct empirical test of these expectations using cross-national data from a conjoint survey experiment.

Research Design

To examine how institutional design features of international AI governance influence public support, we embedded a conjoint experiment in nationally representative surveys conducted in six countries: the United States, China, India, Germany, the United Kingdom, and Japan (N = 15,045). The surveys were fielded by Bilendi during the summer of 2024.[8] Conjoint experiments enable researchers to disentangle how individuals weigh distinct attributes of a policy proposal (Bansak et al. 2021; Zhirkov 2022; Bansak, Hainmueller, and Hangartner 2016) in a manner that is highly representative of their actual policy choices (Hainmueller, Hangartner, and Yamamoto 2015). 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 Figure SI-1 shows the conjoint instructions along with an example of pair profiles.

As detailed in Table 1, we examine four core features of international AI agreements. First, to capture the tradeoff between inclusiveness and effectiveness in global cooperation, we vary the number of participating countries, from a narrow group of 40 to a broad coalition of 160. Second, we randomize the leadership structure, testing whether citizens prefer agreements led by a single powerful country (the United States or China), a regionally representative actor (the European Union or India/Brazil), or a multilateral body such as the United Nations. Third, we vary enforcement mechanisms, ranging from voluntary guidelines to a global court with authority to impose binding legal sanctions, to assess how different models of accountability shape public support. Finally, we vary the substantive focus of the agreement, contrasting general

8. Further details on sampling procedures and sociodemographic distributions are provided in Appendix **??**.

AI risk, encompassing broad safety and existential concerns, with five concrete policy domains prioritized in prior research: autonomous weapons, AI-generated misinformation, algorithmic discrimination, job loss from automation, and threats to privacy and data protection.

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
	All aspects of AI

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

By randomizing attribute values across these dimensions, we can assess to what extent a specific attribute affects public support for global AI governance. To address concerns that the conjoint task itself might prompt respondents to rely on superficial cues that could bias the results, such as consistently focusing on the first-listed attribute, we randomized the order of attributes across respondents, while keeping it fixed within each respondent's three tasks. We address this concern in more depth later in the paper to ensure that the observed effects reflect genuine policy preferences, not artifacts of the conjoint task.[9]

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

9. Appendix A.3 discusses the conjoint design and the survey procedure in detail. The experiment was approved by the [redacted] ethics board.

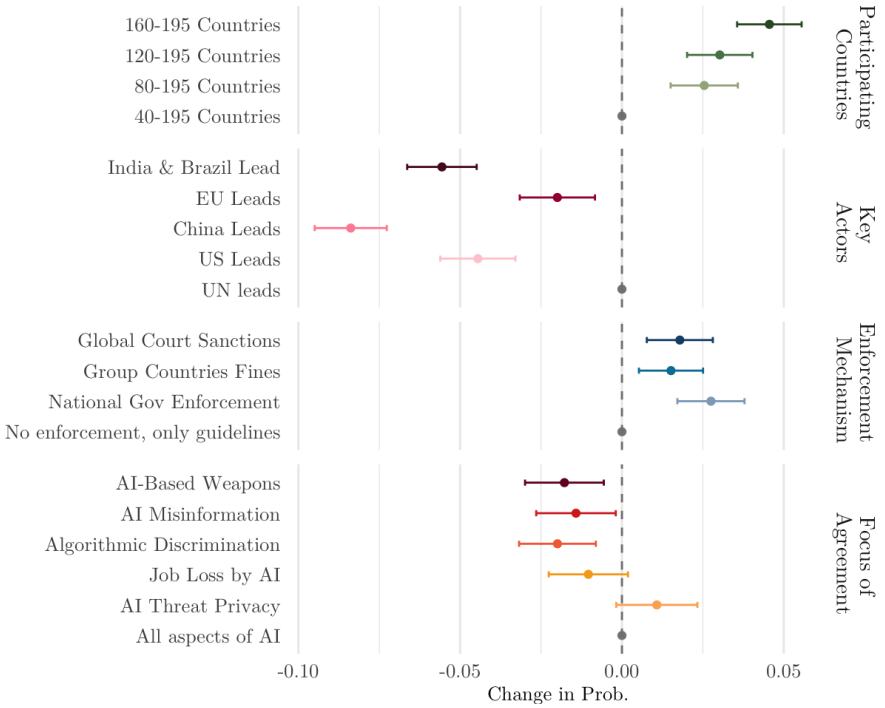


Figure 1: Estimated effects of institutional features on support for adopting AI global governance framework, pooled sample. Points represent average marginal component effects (AMCEs) with 95% confidence intervals (see SI Table SI-3 for the full results).

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. Figure 1 shows the results.[10]

Overall, we find strong support for most of our conjectures about which design features should increase support for global AI regulation. The results reveal that institutional design significantly shapes preferences for global AI governance. Consistent 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 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 (Commission 2023). To put this effect size in perspective

10. 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.

from another domain, Bechtel and Scheve 2013 find that increasing participation in climate agreements from 20 to 80 countries raises public support by 15 percentage points. One possible explanation for the smaller shift in our case is that climate policy is more politically contested, so citizens place greater value on broad participation as a symbolic signal of commitment.

Regarding leadership structure, we also find effects in the expected direction. UN-led frameworks are strongly preferred over those led by individual states or regional blocs, while frameworks led by China or the US incur penalties (-8.4, -4.5, and -2.0 percentage points, respectively), suggesting public skepticism of governance arrangements that might entrench particular state interests.

Enforcement mechanisms also affect 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, regarding the substantive focus of an agreement, our results point to a preference for comprehensive solutions over frameworks on a specific AI domain. Frameworks addressing "All aspects of AI" are consistently favored over those targeting more specific domains like AI-based weapons or misinformation ($p < 0.01$). The only exception is a slight positive effect for agreements focused on privacy (1.1 percentage points, $p < 0.05$). We examine this pattern further in the next section by disaggregating the data by country.

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 —not because these attributes are necessarily more important to them in the real world, but because they serve as cognitive shortcuts to navigate the survey task. To address this concern, Figure SI-2 shows results conducted separately for attentive and non-attentive respondents.[11] The results are very much consistent across these groups, suggesting that our findings reflect considered preferences for AI governance rather than artifacts of cognitive simplification. The only notable difference —that attentive respondents display stronger aversion to China-led frameworks —supports this

11. We define respondents as attentive if they passed an attention check and spent above the 25th percentile of completion time on the conjoint module.

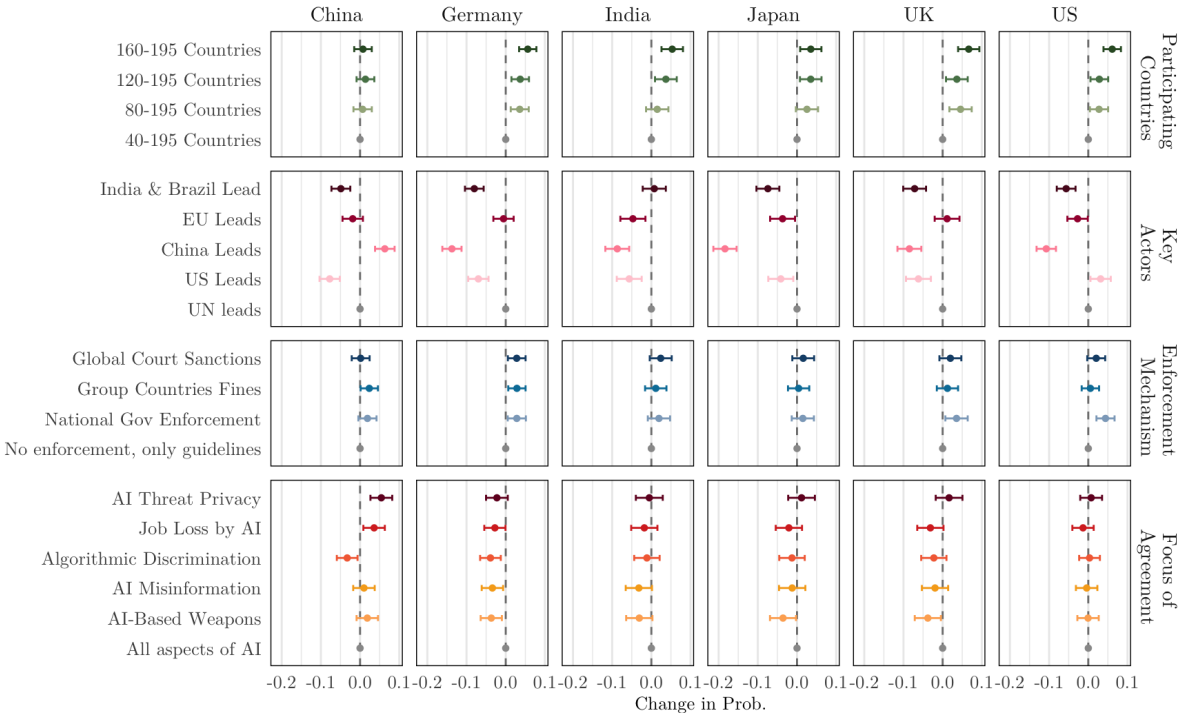


Figure 2: Estimated effects of institutional features on support for adopting AI global governance framework, by country. Points represent average marginal component effects (AMCEs) with 95% confidence intervals. Effects are estimated using LPM regression models with standard errors clustered at the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

interpretation, suggesting that these preferences stem from substantive reasoning about power asymmetries and geopolitical trust, rather than from superficial or inattentive processing of the leadership cue.

Cross-National Patterns

We next examine how these preferences vary across countries, estimating separate linear probability models for each. Figure 2 shows the results. To better understand the underlying preferences over attributes within countries, we also estimate the marginal means across countries (Leeper, Hobolt, and Tilley 2020) (see Figure SI-3).[12]

The preference for broader participation, while present globally, is significantly stronger in the democracies in our sample. 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

12. 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.

the US (Figure (SI-3), a pattern clearly visible in the marginal means as well. 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 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 option, ranking as the first or second most preferred option across all six countries surveyed. This reinforces our result in the pooled sample about the appeal of neutral international bodies, potentially reflecting an overall preference 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).

Turning to the substantive focus of agreement, we find that support for specific issue areas is relatively stable across countries. Frameworks that address all aspects of AI consistently receive high support, suggesting that citizens tend to prefer broad and comprehensive approaches over narrowly targeted ones. The modest positive effect for agreements focused on privacy in the pooled sample appears to be driven largely by respondents in China (0.61 marginal mean, see Figure SI-3). This result may have a specific basis in the country's extensive surveillance

infrastructure (Qiang 2019) and prior evidence that Chinese citizens are generally less concerned about privacy and data protection than their counterparts elsewhere (Kostka, Steinacker, and Meckel 2023). One plausible, albeit post hoc, interpretation is that respondents may have understood the privacy attribute less as a constraint on domestic surveillance and more as a way to protect national data from foreign actors. As we note in the conclusion, subjecting this interpretation to systematic empirical tests is an important task for future research.

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 our results, we turn to examine predicted support levels for different AI governance frameworks. Alongside their forced choices, respondents also rated each proposal on a five-point scale from strongly opposed to strongly supported.[13] To facilitate interpretation, we dichotomize the rating scale into an indicator of support: responses of "somewhat support" or "strongly support" are coded as 1, and all others as 0. This binary measure captures the share of respondents who cross the threshold from neutrality or opposition into affirmative support—namely, the proportion of the public likely to endorse a given framework in a way that is politically meaningful for policy adoption.

Using a linear probability model, we estimated the likelihood that respondents would back each of the 480 unique governance frameworks implied by our conjoint design. To illustrate the range of public backing, we present predicted support for frameworks positioned at the low, median, and high ends of the distribution. We then compare them to prominent real-world governance proposals that serve as benchmarks: the OECD AI Principles, the EU AI Act, China's Global AI Governance Initiative, the U.S. "Military AI Declaration," and the BRICS+ proposal. Figure 3 presents the results.

First, the figure shows that a non-trivial share of the public—nearly 45 percent–favors international cooperation to regulate the development and use of AI, whatever specific form such

13. Choosing one framework over another does not necessarily imply endorsement of either. Respondents may have selected the lesser evil, or conversely viewed both as acceptable.

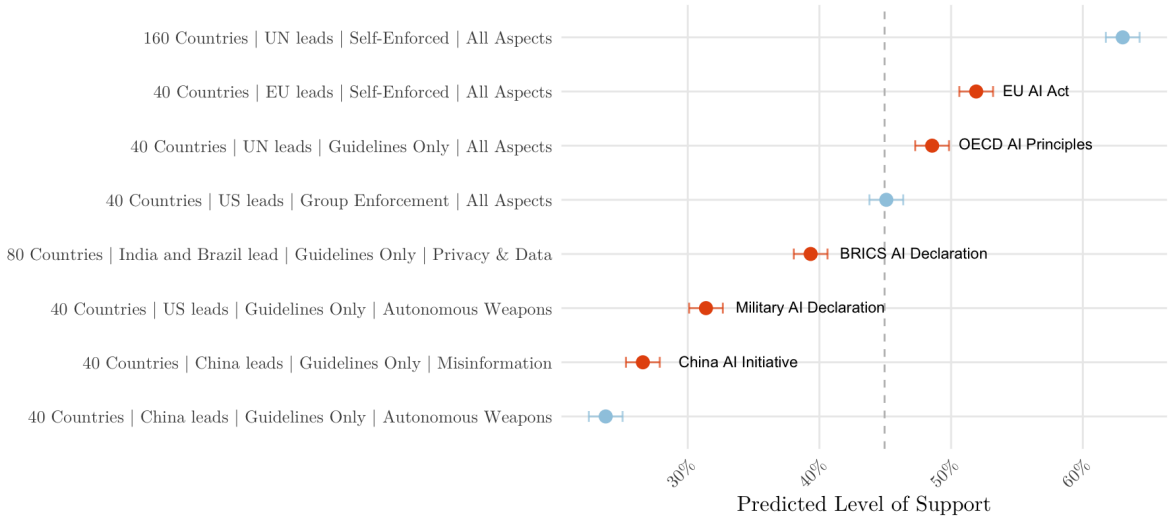


Figure 3: Predicted support for various AI governance frameworks. Predicted public support for various AI governance frameworks. Each point indicates the predicted probability that a respondent supports the specified framework configuration. Horizontal lines show 95% confidence intervals. Blue points represent hypothetical frameworks with predicted support near the 1st, 50th, and 99th percentiles of the distribution across all possible combinations. The dashed gray line indicates the mean level of predicted support across all 480 institutional configurations (0.45 45%). Orange points represent benchmarks approximating real world international agreements.

cooperation may take. This relatively broad baseline of support is consistent with recent evidence that citizens are concerned about the risks of AI and tend to endorse domestic regulation of the technology (Mitts and Shir 2025).[14]

Yet, notably, the figure highlights the extent to which institutional design can meaningfully shift public opinion. Support varies substantially across the distribution of frameworks, from as low as about 25 percent for unpopular designs, such as China-led initiatives with no enforcement and narrow focus, to more than 70 percent for frameworks combining broad participation, UN leadership, and credible enforcement mechanisms.

A comparison with existing initiatives helps to situate these results. The OECD AI Principles, adopted in 2019 by over 40 countries, represent one of the first multilateral efforts to articulate broad norms for trustworthy AI across a wide range of domains (Roberts et al. 2024b). This framework takes a relatively cautious approach, characterized by limited participation and

14. To further assess overall backing, we examine the share of respondents who categorically opposed international cooperation, regardless of its specific features. Figure SI-4 in the Appendix shows that outright rejection of international AI governance is rare. The proportion of respondents who consistently rejected at least 75 percent of the proposals they evaluated is very small: about 4 percent of the pooled sample, ranging from virtually none in China and India to nearly 10 percent in Germany.

voluntary guidelines rather than binding rules. Our analysis estimates public support for such a framework at the 48 percent, close to the overall average across institutional designs. The EU AI Act, finalized in December 2023, takes a more forceful approach by introducing binding regulations and concrete enforcement mechanisms, albeit within a more limited geographical scope, stands out slightly above the baseline, with support at 52 percent.

By contrast, the initiatives launched by major powers rank much lower. China's Global AI Governance Initiative and the U.S. Military AI Declaration, both introduced in 2023, rest on voluntary commitments and narrow membership. Predicted support for each hovers around one-third of the public. Taken together, these results point to a gap between what publics would support and what some governments have thus far advanced. The findings challenge the prevailing narrative that emphasizes national sovereignty and geopolitical competition as insurmountable barriers to global AI governance (Von Bogdandy, Goldmann, and Venzke 2017). 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.

Discussion

As AI advances, governments and non-governmental organizations are increasingly stressing the need for international institutions to constrain countries' laws and behavior regarding the use of AI in a variety of settings, akin to longstanding international agreements on arms control, trade, human rights, weapons of mass destruction, and more recent contested agreements regarding carbon emissions. As the societal and international implications of this rapidly evolving and powerful technology remain unclear, our theoretical expectations and results from a large, diverse international sample should provide the basis for future studies.

We first summarize our findings. A core finding of our study is that citizens prefer international agreements on artificial intelligence that include enforcement mechanisms rather than those based on voluntary commitments. This stands in some tension with patterns in other areas of international cooperation, such as climate change mitigation, where symbolic commitments have often carried significant weight. The Paris Agreement is a clear example: negotiators scaled

back binding commitments to secure near-universal membership. In the case of AI, our results suggest that citizens place greater weight on credibility and enforceability than on symbolic commitments. One potential explanation is that, unlike issues such as climate change or trade, AI governance has not yet become the subject of entrenched political conflict. At present, there appears to be a broad societal consensus that AI should be developed and used responsibly. This baseline agreement may make citizens more willing to focus on institutional effectiveness and credibility, rather than on symbolic or expressive considerations. However, history suggests that such consensus is often fleeting. Policy domains that begin as technical or expert-driven (e.g., climate science, vaccine policy), can become politically polarized over time (Hochschild 2021).

At the same time, while citizens clearly distinguish between the presence and absence of enforcement, we find that they are less sensitive to its specific institutional form. Whether enforcement is exercised by national governments, a coalition of states, or an international court appears less important. What seems to matter is that citizens see clear assurances that rules will be enforced and that violations will carry consequences. This pattern suggests that policymakers have some flexibility in designing enforcement arrangements, which may help them navigate the legal and technical obstacles to regulating a fast-moving and complex technology.

Our analysis also indicates that citizens do not strongly differentiate among the specific substantive issues to be regulated. Instead, our data support the idea that at this early stage of public debate over AI, people evaluate their governance less on the merits of particular policies, but lean on broad institutional cues about fairness, accountability and credibility. The pattern holds for both attentive and less-attentive respondents, which suggests these are genuine evaluative criteria rather than superficial shortcuts. Rather than being disengaged or indifferent, citizens apply a coherent logic grounded in widely shared expectations about how international institutions should function. Broad participation, credible enforcement, and neutral leadership serve as proxies for reciprocity, compliance, and legitimacy.

The findings have practical implications for institutional design. They point to arrangements with broad membership, neutral leadership, and credible enforcement as most likely to secure public backing; under such conditions, a proposal can attain majority support. This contrasts with the modest, largely voluntary initiatives that currently dominate global AI governance and suggests governments may have more room to pursue ambitious agreements than is often

assumed.

We close with several directions for future research, building on our theory and results. First, we have deliberately shied away from exploring individual-level determinants of global AI regulation such as level of education or nationalism. These individual-level characteristics could have more weight if AI becomes further politicized, and they could be the basis for theorizing cleavages of support for regulatory proposals within countries. Second, a promising direction for future research is to track, using longitudinal data, how global preferences evolve as AI governance becomes more contested, whether through partisan framing, interest-group mobilization, or first hand experience of the technology and its implications. Such longitudnal data would be particularly instructive if interest groups or countries select certain AI dimensions to lobby about. Third, in our opening theoretical discussion we outlined a number of distinct reasons why global cooperation AI might be desirable, but how individuals think about AI as a global problem may both vary (depending on country of residence or individual features), and change, as various AI dimensions (such as labor-market concerns vs. autonomous weapons) become relatively more salient. Future studies may probe more deeply into how different theories of AI as a global dilemma (such as whether it is viewed as a problem of public goods provision or one of constraining a particular power) would map onto distinct policy proposals. Our results hopefully set an agenda for rigorously thinking about the role of public preferences of international governance over this momentous technology.

References

Acemoglu, Daron. 2025. "The Need for Multipolar Artificial Intelligence Governance." In *The New Global Economic Order,* 108–119. Routledge.

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

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

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

Bechtel, Michael M, Jens Hainmueller, and Yotam Margalit. 2014. "Preferences for international redistribution: The divide over the Eurozone bailouts." *American Journal of Political Science* 58 (4): 835–856.

Bechtel, Michael M, Kenneth Scheve, and Elisabeth van Lieshout. 2019. "What determines climate policy preferences if reducing greenhouse-gas emissions is a global public Good?" *Available at SSRN 3472314.*

Bechtel, Michael M, and Kenneth F Scheve. 2013. "Mass support for global climate agreements depends on institutional design." *Proceedings of the National Academy of Sciences* 110 (34): 13763–13768.

Bechtel, Michael M, Kenneth F Scheve, and Elisabeth van Lieshout. 2022. "Improving public support for climate action through multilateralism." *Nature Communications* 13 (1): 6441.

Bengio, Yoshua. 2023. "AI and catastrophic risk." *Journal of democracy* 34 (4): 111–121.

Bernauer, Thomas, Liang Dong, Liam F McGrath, Irina Shaymerdenova, and Haibin Zhang. 2016. "Unilateral or reciprocal climate policy? Experimental evidence from China." *Politics and Governance* 4 (3): 152–171.

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

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

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

Butcher, James, and Irakli Beridze. 2019. "What is the state of artificial intelligence governance globally?" *The RUSI Journal* 164 (5-6): 88–96.

Büthe, Tim, Christian Djeffal, Christoph Lütge, Sabine Maasen, and Nora von Ingersleben-Seip. 2022. *Governing AI–attempting to herd cats? Introduction to the special issue on the Governance of Artificial Intelligence,* 11.

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

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

Cihon, Peter, Matthijs M Maas, and Luke Kemp. 2020b. "Should artificial intelligence governance be centralised? Design lessons from history." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 228–234.

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

Commission, European. 2023. *G7 Leaders' Statement on the Hiroshima AI Process.*

Dafoe, Allan. 2018. "AI governance: a research agenda." *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK* 1442:1443.

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

Dellmuth, Lisa Maria, Jan Aart Scholte, and Jonas Tallberg. 2019. "Institutional sources of legitimacy for international organisations: Beyond procedure versus performance." *Review of International Studies* 45 (4): 627–646.

Druckman, James N, Erik Peterson, and Rune Slothuus. 2013. "How elite partisan polarization affects public opinion formation." *American political science review* 107 (1): 57–79.

Erdélyi, Olivia J, and Judy Goldsmith. 2018. "Regulating artificial intelligence: Proposal for a global solution." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society,* 95–101.

Faverio, Michelle, and Alec Tyson. 2023. "What the data says about Americans' views of artificial intelligence." Pew Research Center. https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/.

Future of Life Institute. 2017. *Asilomar AI Principles.* https://futureoflife.org/open-letter/ai-principles.

Gabel, Matthew, and Kenneth Scheve. 2007. "Estimating the effect of elite communications on public opinion using instrumental variables." *American Journal of Political Science* 51 (4): 1013–1028.

Gaikwad, Nikhar, Federica Genovese, and Dustin Tingley. 2025. "Climate action from abroad: Assessing mass support for cross-border climate transfers." *International Organization* 79 (1): 146–172.

Gallego, Aina, and Thomas Kurer. 2022. "Automation, digitalization, and artificial intelligence in the workplace: implications for political behavior." *Annual Review of Political Science* 25:463–484.

Ghassim, Farsan, Mathias Koenig-Archibugi, and Luis Cabrera. 2022. "Public opinion on institutional designs for the United Nations: An international survey experiment." *International Studies Quarterly* 66 (3): sqac027.

Gilardi, Fabrizio, Atoosa Kasirzadeh, Abraham Bernstein, Steffen Staab, and Anita Gohdes. 2024. "We need to understand the effect of narratives about generative AI." *Nature Human Behaviour* 8 (12): 2251–2252.

Gray, Julia, and Raymond P Hicks. 2014. "Reputations, perceptions, and international economic agreements." *International Interactions* 40 (3): 325–349.

Guisinger, Alexandra. 2009. "Determining trade policy: Do voters hold politicians accountable?" *International Organization* 63 (3): 533–557.

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

Hatz, Sophia, Noemi Dreksler, Kevin Wei, and Baobao Zhang. 2025. "Local US officials' views on the impacts and governance of AI: Evidence from 2022 and 2023 survey waves." *arXiv preprint arXiv:2501.09606.*

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

Hochschild, Jennifer. 2021. *Genomic politics: how the revolution in genomic science is shaping American society.* Oxford University Press.

Jones, Charles I. 2024. "The ai dilemma: Growth versus existential risk." *American Economic Review: Insights* 6 (4): 575–590.

Juhász, Réka, and Nathan Lane. 2024a. "The Political Economy of Industrial Policy." *Journal of Economic Perspectives* 38 (4): 27–54.

———. 2024b. "The political economy of industrial policy." *Journal of Economic Perspectives* 38 (4): 27–54.

Kono, Daniel Y. 2008. "Does public opinion affect trade policy?" *Business and Politics* 10 (2): 1–19.

Korinek, Anton, Martin Schindler, and Joseph E Stiglitz. 2022. "Technological progress and artificial intelligence." *How to Achieve Inclusive Growth,* 163–211.

Kostka, Genia, Léa Steinacker, and Miriam Meckel. 2023. "Under big brother's watchful eye: Cross-country attitudes toward facial recognition technology." *Government Information Quarterly* 40 (1): 101761.

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

Magistro, Beatrice, Sophie Borwein, R Michael Alvarez, Bart Bonikowski, and Peter J Loewen. 2024. "The Common Microfoundations of Attitudes Toward Artificial Intelligence (AI) and Globalization." *Available at SSRN 4795006.*

Mitts, Tamar, and Raviv Shir. 2025. "How Media Coverage and Elite Communication Shape Public Opinion on AI Regulation." *Working Paper.*

Nair, Gautam, and Kyle Peyton. 2022. "Building mass support for global pandemic recovery efforts in the United States." *PNAS nexus* 1 (4): pgac123.

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

Ng, Davy Tsz Kit, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. "Conceptualizing AI literacy: An exploratory review." *Computers and Education: Artificial Intelligence* 2:100041.

Perrault, Raymond, and Jack Clark. 2024. *Artificial Intelligence Index Report 2024.* Technical report. Stanford Institute for Human-Centered Artificial Intelligence.

Qiang, Xiao. 2019. "The road to digital unfreedom: President Xi's surveillance state." *Journal of Democracy* 30 (1): 53–67.

Raviv, Shir. 2025. "When Do Citizens Resist The Use of AI Algorithms in Public Policy? Theory and Evidence." *The Journal of Politics.*

Roberts, Huw, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. 2024b. "Global AI governance: barriers and pathways forward." *SSRN Electronic Journal.*

———. 2024a. "Global AI governance: barriers and pathways forward." *International Affairs* 100 (3): 1275–1286.

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

Spilker, Gabriele, Thomas Bernauer, and Víctor Umaña. 2018. "What kinds of trade liberalization agreements do people in developing countries want?" *International Interactions* 44 (3): 510–536.

Taeihagh, Araz. 2021. "Governance of artificial intelligence." *Policy and society* 40 (2): 137–157.

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

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

Tannenwald, Nina. 2024. "The Nuclear Nonproliferation Regime as a "Failed Promise": Contestation and Self-Undermining Dynamics in a Liberal Order." *Global Studies Quarterly* 4 (2): ksae025.

Thompson, Alexander. 2024. "Contestation and Resilience in the Liberal International Order: The Case of Climate Change." *Global Studies Quarterly* 4 (2): ksae011.

Tingley, Dustin, and Michael Tomz. 2014. "Conditional cooperation and climate change." *Comparative Political Studies* 47 (3): 344–368.

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

Université de Montréal. 2017. *Déclaration de Montréal IA Responsable.* www.montrealdeclaration-responsibleai.com.

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

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

Wenzelburger, Georg, Pascal D König, Julia Felfeli, and Anja Achtziger. 2024. "Algorithms in the public sector. Why context matters." *Public Administration* 102 (1): 40–60.

Wilczek, Bartosz, Sina Thäsler-Kordonouri, and Maximilian Eder. 2024. "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,* 1–15.

Zaidan, Esmat, and Imad Antoine Ibrahim. 2024. "AI governance in a complex and rapidly changing regulatory landscape: A global perspective." *Humanities and Social Sciences Communications* 11 (1): 1–18.

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

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

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

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

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 Question Wording

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

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.

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.

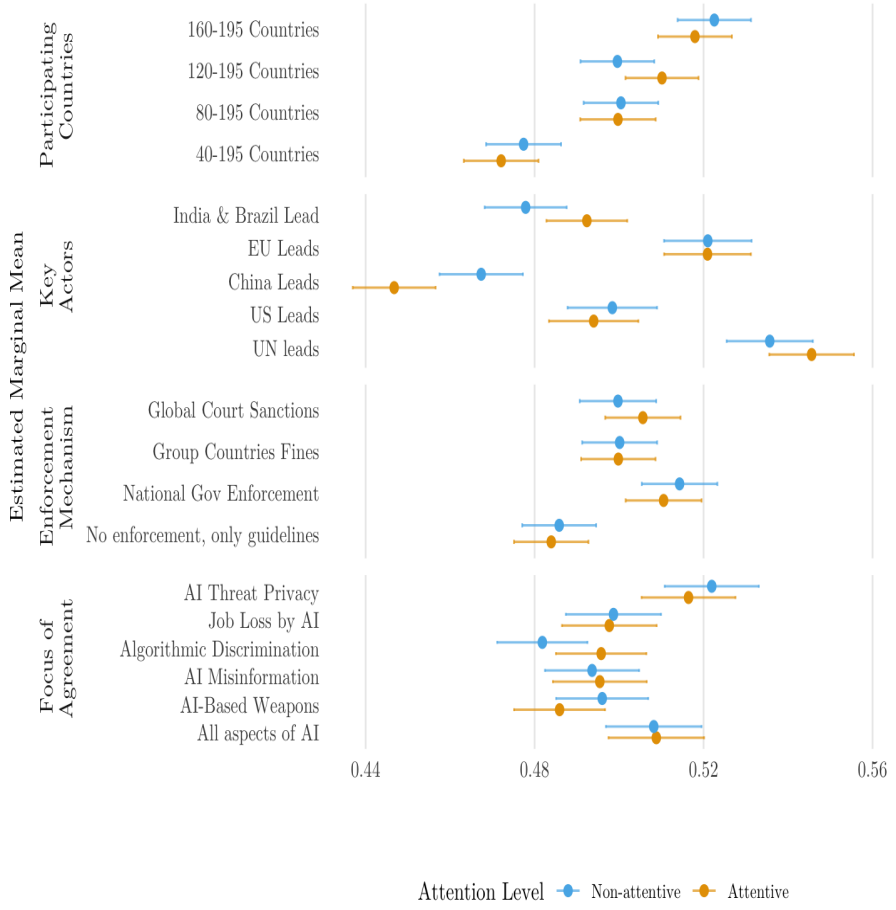
Figure SI-2 presents the marginal means for each attribute level, estimated separately for attentive and non-attentive subgroups. While some minor differences in the magnitude

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

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.

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.

B.3 Cross-National Differences in Preferences

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

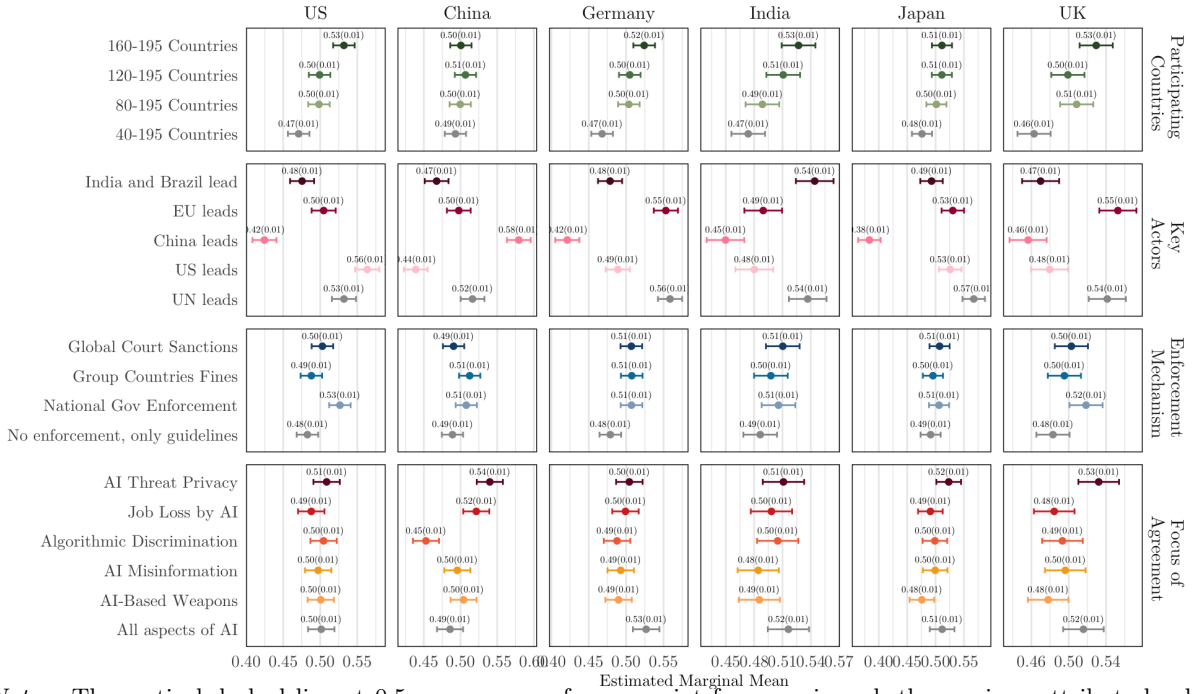
	AMCEs on Preference for International Proposal					
	US	UK	Germany	Japan	India	China
	(1)	(2)	(3)	(4)	(5)	(6)
80-195 Countries	0.027*	0.046**	0.036**	0.025	0.015	0.007
	(0.012)	(0.015)	(0.012)	(0.014)	(0.015)	(0.012)
120-195 Countries	0.028*	0.036*	0.037**	0.035*	0.037**	0.014
	(0.012)	(0.014)	(0.011)	(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 2. 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



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 Governance 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). $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Figure 3 in the main text shows the distribution of predicted support across all institutional configurations generated by our conjoint design. On average, frameworks receive support from about 45 percent of respondents, with predicted values ranging from roughly 25 percent to over 70 percent. This spread underscores the role of design features in shaping public approval. To provide additional context, we examine the prevalence of consistent opposition.

To this end, we calculate for each respondent, the share of proposals they opposed among the set they evaluated. We define "high rejecters" as those who rejected at least 75 percent of proposals. We then estimate the prevalence of high rejecters in the pooled sample and separately by country using survey weights and report 95 percent confidence intervals. Figure shows the results.

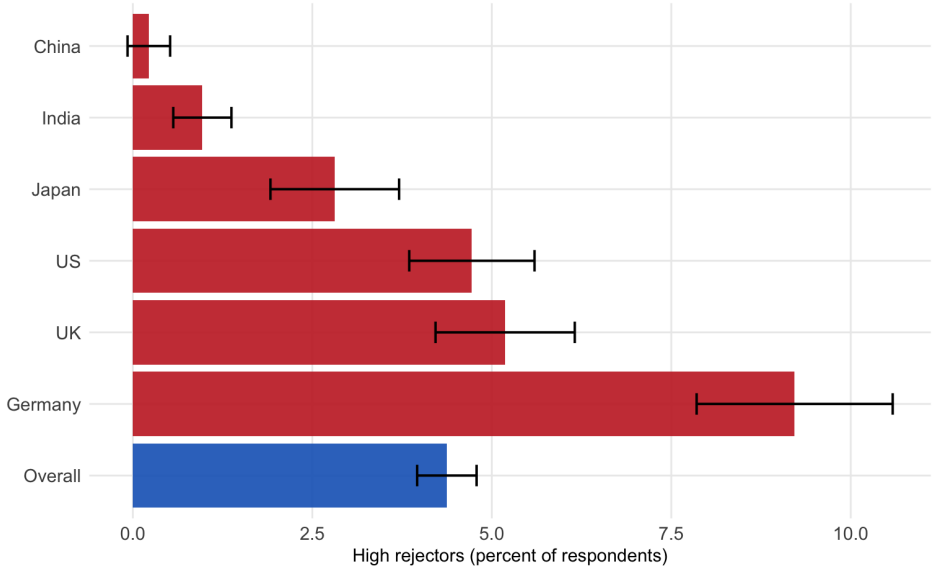


Figure SI-4: Share of respondents consistently opposing international cooperation on AI governance. Bars show the percentage of respondents in each country (and overall) who opposed at least 75% of the proposals they evaluated, regardless of institutional design features. Error bars represent 95% confidence intervals based on survey weights. N = 13,890 respondents

C Pre Analysis Plan

This section contains the anonymized version of our pre-analysis plan that was submitted prior to data collection. While the PAP outlines three research questions, this article focuses on the first question. We include the remaining questions here for the sake of transparency and completeness.

C.1 Hypotheses

Research Question 1: How does the institutional design of international agreements affect willingness to support global AI governance?

Number of Countries:

- **H1:** Agreements involving more countries will receive greater public support than those involving fewer countries.

Lead Country:

- **H2a:** Agreements led by the EU will receive more support than those led by China or the US.
- **H2b:** This EU lead effect will be stronger among citizens who trust the EU more.
- **H3:** Agreements led by Global South countries will receive more support than those led by China or the US.
- **H4:** Agreements led by the United Nations will receive more support compared to those led by China or the US.

Institutional Power:

- **H5a:** Agreements granting more enforcement capacity will command more support than those granting less.
- **H5b:** Agreements granting less enforcement capacity will command more support than those granting more.

Issue Areas Covered:

- **H6a:** Agreements on higher concern issues like weapons, jobs, and privacy will command more support than lower concern issues like misinformation or discrimination.
- **H6b:** This positive effect will be stronger among citizens more concerned about AI's impact in that area.

Research Question 2: Does public sensitivity to the institutional features of international AI governance agreements vary across subgroups?

National Interests:

- **H7a:** Citizens are more likely to support agreements proposed or led by their own country or region of origin.
- **H7b:** The effect of a national lead will be even stronger among citizens with more nationalistic attitudes.

By Personal Concern:

- **H8a:** Individuals ranking worker substitution as the highest concern will be more supportive of AI agreements dealing with that concern.
- **H8b:** Individuals ranking autonomous weapons as the highest concern will be more supportive of AI agreements dealing with that concern.
- **H8c:** Individuals ranking data privacy as the highest concern will be more supportive of AI agreements dealing with that concern.
- **H8d:** Individuals ranking privacy and data collection as the highest concern will be more supportive of AI agreements dealing with that concern.
- **H8e:** Individuals ranking misinformation as the highest concern will be more supportive of AI agreements dealing with that concern.

  By Predispositions:
- **H9a:** Individuals with more nativist attitudes will be more opposed to AI governance agreements involving multilateral institutions or international courts/enforcement.
- **H9b:** Individuals with more anti-globalization attitudes will be more opposed to agreements involving multilateral institutions or international courts/enforcement.

  By Gender:
- **H10a:** Women are more likely to value a larger number of countries signing the agreement than men.
- **H10b:** Men are more likely to prioritize agreements mitigating risks related to the economy and security compared to women.

  Research Question 3: Does the relative importance of different institutional features vary across countries? Historical experiences, current geopolitical relationships, and domestic concerns about economic impacts or social values create strong reasons to expect cross-national variation in the weight of considerations shaping support for international agreements on AI governance. For instance, citizens in countries with heightened security concerns may prioritize agreements addressing the regulation of AI-based autonomous weapons. Conversely, in democratic nations where free speech and personal rights are paramount, concerns about misinformation, privacy, and ethical AI use might be more prominent. The geopolitical rivalry and mistrust between the US and China offer another example, where citizens of these countries might prioritize considerations related to the leading actor in the agreement, specifically showing aversion to proposals led by the rival country. Americans, on the other hand, might be more open to agreements led by the EU or UN due to shared values and historical alliances. Therefore, our research remains agnostic about the specific directions of these sensitivities, aiming instead to explore them as an open empirical question.

C.2 Design Plan

Study type
Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

Blinding
For studies that involve human subjects, they will not know the treatment group to which they have been assigned.

Is there any additional blinding in this study? No data

Study design
To explore these questions and hypotheses, we will employ a conjoint experiment embedded in a large-scale, cross-national survey to assess how key institutional design features of global AI governance agreements influence public willingness to support such frameworks. Specifically, we will use a standard conjoint design, in which respondents will evaluate 3 pairs of hypothetical proposals for international AI governance cooperation. Respondents will be presented with pairs of hypothetical proposals for international AI governance cooperation. We will use a standard conjoint design, where respondents will evaluate three pairs of hypothetical proposals for international AI governance cooperation. Each proposal will vary randomly across four relevant dimensions: (1) Number of participating countries (40, 80, 120, or 160 out of 195) (2) Key actors drafting the proposal (US, China, EU, India & Brazil, or UN) (3) Type of enforcement (no enforcement, domestic enforcement, multilateral with fines, or global court with sanctions) (4) Topic focus (autonomous weapons, misinformation, discrimination, job displacement, data collection, or all aspects of AI). For each pair, respondents will indicate their preferred proposal for their country to adopt and rate their support or opposition to each proposal. This within-subjects design allows us to measure preferences and support levels directly.

Exact wording for respondents
"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.
Which option would you prefer your country to adopt? [Option 1, Option 2]
Please rate whether you support or oppose option 1. [Strongly support, Somewhat support, Neither support nor oppose, Somewhat oppose, Strongly oppose]
Please rate whether you support or oppose option 2."

C.3 Sampling Plan

Existing Data
Registration prior to any human observation of the data

Explanation of existing data
This report pre-registers an experiment embedded within a survey that will be conducted in five countries. Data collection for the UK and US is underway, with Germany, Japan, and China set to begin shortly. The survey firm Respondi is collecting and currently storing the data. We will gain access to the data only after this pre-registration report is submitted.

Data collection procedures
The survey will only include participants over the age of 18 who give their consent to participate and correctly answer an attention (screener) check at the beginning of the survey. This screener includes a definition of artificial intelligence to ensure that respondents have a similar understanding of the concept of technology before proceeding with the survey.

Sample size
Our target sample size is ~2,000 respondents for each country.

Sample size rationale
The sample size is based on a power calculation, assuming a power of 0.8 and focusing on the attribute with four levels to estimate an Average Marginal Component Effect (AMCE) of 0.04.

C.4 Variables

Manipulated variables
We will experimentally manipulate: (1) Number of participating countries (40, 80, 120, or 160 out of 195) (2) Key actors drafting the proposal (US, China, EU, India & Brazil, or UN) (3) Type of enforcement (no enforcement, domestic enforcement, multilateral with fines, or global court with sanctions) (4) Topic focus (autonomous weapons, misinformation, discrimination, job displacement, data collection, or all aspects of AI).

Measured variables
Our primary outcome of interest is binary: an indicator for whether the respondent chose proposal 1# rather than 2#. We will use rating question as a secondary outcome.

C.5 Analysis Plan

Statistical models

Upon receiving the data, we will first make sure that all the attributes are labeled in an appropriate format (for instance topic1-topic6). Then, we will generate a new variable for each agreement (agreement1agreement6) and generate and id (id_conjoint) for each agreement. We will reshape the dataset from wide to long using id_conjoint (Conjoint1_Version) and the respondents' id (record) as identifier so that there are six observations for each respondent, one for each agreement they have encountered. Our main interest is in evaluating how support for an international proposal to regulate AI shifts as a function of changes in the attributes the regulation proposal takes along the following main dimensions of interest: Type of agreement, and Focus of global agreement, Number of participating countries, Key actors writing the proposal. To assess the impact of each feature along those dimensions, we compute the average marginal component-specific effects (AMCEs), which measure the average impact of a change in an international agreement feature on the probability of supporting the country to adopt this regulation. We will regress the dependent variable on a set of factor variables that capture the specific values that the international proposal takes on each of the policy attributes. For each dimension, we omit one of the attribute values and use it as the baseline category. We will store the estimates of the marginal effects associated with each attribute and their 95 percent and plot them.