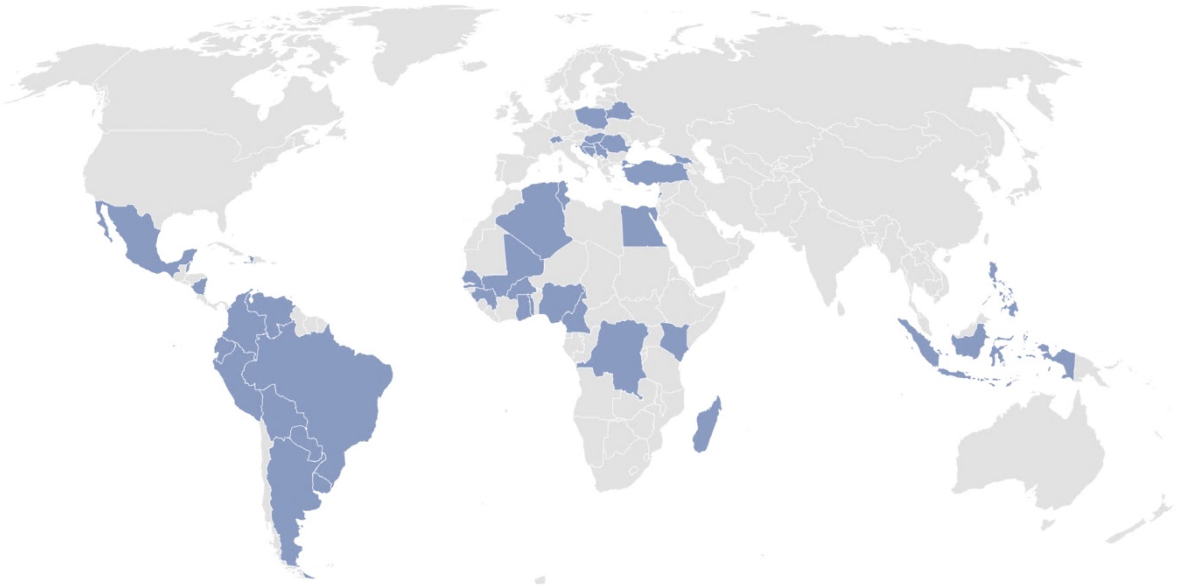
Project Final Report

Platform Governance Survey 2022 – A Global Study of Public Attitudes Towards Content Moderation



- Funded by the Swiss Federal Office of Communications (OFCOM)
- Principal Investigator: Dr. Dennis Redeker, ZeMKI, Centre for Information, Communication and Information Research, University of Bremen, Germany
- Research assistants: Fee-Sofie Cohausz and Bestian van der Neut
- Project Duration: 1 August 2022 to 31 March 2023
- Project Volume: 22,500 CHF

1. Introduction

In the course of the project "Platform Governance Survey 2022 – A Global Study of Public Attitudes Towards Content Moderation", I conducted a large-scale public opinion survey of users of Facebook and Instagram in 41 countries in order to understand their attitudes toward platform content moderation. A particular focus was on the acceptance (and legitimacy) of different actor groups (stakeholder groups within the language of the dominant multistakeholder discourse) in the process of content moderation.

The survey project has also received funds from the *Exploring Digital Transformation: Special Funding Program* provided by the State of Bremen. In that context, the EU Horizon Project REMIT (2023-2027) has been prepared using the questions included using that part of the budget. Further funds for staff costs have been contributed by ZeMKI, Centre for Information, Communication and Information Research, University of Bremen.

This final report commences by describing the theoretical background of the project and the research questions addressed. In the third section, the report elaborates on the methods used to collect data and analyze attained data. The fourth section entails a detailed description of the sample, including a comparison of the study sample to both populations on the two social media services and the target countries' populations. The final section offers preliminary findings from this study.

The results are being prepared for submission as part of a special issue of *Internet Policy Review*. Very early results were presented in the course of three conferences: the *International Studies Association Annual Convention* in Montréal (March 2023), the *Platform Governance Research Network Online Conference* (April 2023) and the *European Multidisciplinary Conference on Global Internet Governance Actors, Regulations, Transactions and Strategies* (GIG-ARTS, Padova, May 2023). The conference presentation at the ISA Annual Convention led to a further after-sampling effort in order to increase and further improve the sample until the project end on 31 March 2023.

2. Theoretical Background and Research Questions

Social media platforms like Facebook, Twitter and TikTok are the "new governors" or "custodians" of the Internet (Klonick 2018; Gillespie 2018). How they moderate global speech online affects the communication practices of billions of people and it can make or break social movements and political resistance, and generally be a critical risk factor for human rights violations. These platforms are increasingly joined by states, international organizations, civil society, journalists and others in defining and interpreting the limitations of speech online, be it through legislation, guidelines or by helping platforms to distinguish misinformation from legitimate content. Increasingly, researchers ask serious questions concerning the legitimacy of various approaches of content moderation (Haggart & Keller 2021; Suzor 2019), which must extend to the question of which actors ought to fulfill which function in content moderation. A legitimate content moderation constellation (and potentially division of labor) is arguably one that is perceived to be legitimate by the "governed" themselves (for whatever qualities are appraised by them). As of today, however, we have little empirical knowledge about what users think about content moderation in general. Even less so, we know what users think about different roles for states, international organizations or NGOs in platform content moderation.

When platforms develop their content moderation setups, they increasingly think about the involvement of different kinds of stakeholders to full different roles such as rule-deliberation, fact-checking, or evaluation of platform practices. Likewise, policy innovations – such as social media councils are increasingly discussed both in national and international digital governance fora. In February 2023, more than 1,500 representatives of different sectors assembled at UNESCO's headquarters to deliberate on a proposal of a global guideline for content moderation regulation involving such multistakeholder councils (UNESCO 2023). In the funded project I investigate who prefers which actors' involvement in content moderation and in which specific role (rule-making, rule-applying and rule-adjudicating). The *who* question

here relates to country-differences as well as differences between genders. It is from these different aspects that the research project poses three different empirical research questions. The first question may be the most comprehensive:

> *RQ1: Which actor type do Meta platform users prefer for which functional role in content moderation? How do preferences differ based on demographic characteristics such as age, gender and education?*

For this question, a number of variables require definition. Functional roles include the *making of rules* in content moderation (legislative function), which may include making laws about what content can be posted by users online or the creation of platform-specific community standards that outline what can and cannot be posted on a given platform. Secondly, the *application of rules* (executive function) entails the actual act of content moderation on platforms including though human moderators and algorithmic systems but potentially broader conceived including roles for police, civil society or users (as in the case of more decentralized platforms). Lastly, the *adjudication of rules* (judicial function) is an important role to take on with regard to platform content moderation. For instance, platforms have set up highest-level appeals bodies for their content moderation decisions, such as Meta's Oversight Board. In addition, increasingly, external social media councils may take on such rules for specific jurisdictions. In general, nation states have shown to be able to take on appeals functions if need be. What actor types are relevant for this question. Based on a review of the literature and news sources of what actually happens already (or is considered) for at least one of the three functions, and based on feedback from colleagues, I opted to include seven different actor groups in the questionnaire: Meta Inc. (the company itself inter alia through the Oversight Board), public authorities (parliamentary commissions, policy and prosecutors or courts respectively), organized civil society (NGOs, etc.), third party commercial vendors, academics, journalists or users themselves. The second question is based on the differentiation entailed in RQ1 but adds a comparative layer:

> *RQ2: How do the preferences vary across countries and country groups (e.g., Global South vs. Global North) and based on demographic characteristics such as age, gender and education?*

The third research question asks causal questions about the drivers of the trends unearthed in RQ1 and RQ2.

> *RQ3: What are the drivers of variation across countries? Specifically, how is a higher propensity to wanting public authorities/Meta/civil society, etc. involved in the three functions of content moderation explained by levels of levels of trust in these institutions in general?*

Since a number of factors for a full-fledged correlational analysis are not in place, the report shows trends through sub-group analysis. All in all, the project was successful in creating new empirical evidence on how users perceive platform content moderation and how they perceive

content moderation roles of different *alternative governors* of speech.[1] Knowing user preferences better allows us to move forward toward evidence-based reform proposals. Specifically, knowing what users in different countries or of different genders think about who they want to be in charge of content moderation can help to develop new-generation institutions and agreements.

3. Methods for Data Collection and Analysis

One assumption of the project – initially instilled by previous research presented at the *ICA Pre-conference on Alternative Content Regulation on Social Media* in Paris in 2022 – is that the approval for certain actors' involvement in content moderation differs across national boundaries. This motivated a sampling across as many countries as possible and including OECD countries and those of the "Majority World"[2]. The aim was to include more countries from the so-called "Global South" than would normally be included in comparative ross-sectional survey research projects. Another core aim was to include both liberal democratic and less democratic countries among countries selected for recruitment. Variation in political regime type would likely reflect on trust in certain (state) institutions, which again is seen as a potential factor for driving trust in these actors in functional roles in content moderation on platforms.

In order to attain a sufficiently large sample in a set of countries in both the Global South and Western and Eastern Europe, together with two research assistants – Fee-Sofie Cohausz and Bastian van der Neut, I conducted an online opinion survey in 41 countries. We used online advertisement, specifically on Facebook and Instagram, to recruit social media users in the target countries. This recruitment method, albeit relatively novel, has already been demonstrated to be a viable alternative to classical recruitment for representative country-level samples (Pötzschke & Braun, 2017; Redeker & Sturm, 2019; Rosenzweig et al., 2020; Zhang et al., 2020). We used a quota-sampling approach, based on distribution of age and gender in the population of the sampling countries. The underlying quotas stem from the most recent census data available to us. The ad micro-targeting options afforded by the *Meta Ads Manager* – albeit all its shortcomings – allows to engage in effective quota-targeting.

An advertisement in the (mostly) country-appropriate language led those who clicked to a questionnaire, administered through LimeSurvey hosted on servers of the University of Bremen. Participants could win an equivalent of 100 US dollars through a raffle, they were informed about the purpose of the study and all rights they had regarding the retainment of the data and the possibility to withdraw from the study (informed consent). Respondents were also asked whether they would want to be recontacted for future surveys, which around half confirmed. On average, respondents included in the final dataset spent around 29 minutes on the survey, which is long for an online survey. The Swiss sample is a special case. Here, only

[1] Speaking of „alternative governors", I have in mind Kate Klonick's (2018) paper on "the new governors". Consequently, the alternatives are those who are not the platform companies themselves.
[2] In this report, I use non-OECD, Global South and Majority World interchangeably, aware of the lack of precision connected to such use.

a smaller questionnaire focused on social media content moderation was adopted – on average respondents from Switzerland took around 21 minutes to complete the questionnaire. This is due to the high costs of recruiting a well-balanced sample in Switzerland. Costs for a valid response were easily fifty-times as high as for those countries for which costs were lowest, even adopting these measures.[3]

In general, next to higher-level theoretical reasons for inclusion of countries into the study – based on assumption expressed in the research questions, a number of other reasons affected the viability of doing so. Two main qualities were required to conduct survey research effectively in countries: First, a relatively high spread of Facebook or Instagram use in the population so that opinions are not merely those of a very small elite or dissident group (as in Cuba, China or a number of countries in Sub-Saharan Africa). This means that during the selection of countries at the proposal stage and also later on when replacement countries had to be identified, a focus was on identifying the countries with higher levels of use of Meta's platform services. The second, no less important criterion for inclusion was the cost of running an effective ad-based recruitment campaign. Here, at the proposal stage, I relied on data from previous smaller studies (covering only a dozen countries) and on the estimates provided by Meta's Ads Manager. Importantly, and in addition to other reasons for exclusion, some countries could not be included due to international or US sanctions, including Iran, Russa, North Korea. In addition, access to translators and research assistance in different languages was important for the choice of countries eventually included. The survey (and ads) could have worked in Thailand or Vietnam but not without a translation into local languages. The reality of sampling for the survey showed results as expected in some countries (such as Argentina, Venezuela or Kenya) and results that were surprising and underwhelming. Countries with higher incomes turned out to be even more difficult to recruit in than previously experienced or even as company data would suggest. This meant that many of the European countries and general OECD countries were not feasibly generating a 300-500 range of responses without spending at least 3,000-5,000 for each of these countries. Surprisingly, some countries performed particularly poorly even though the ad markets are not particularly developed/expensive: for instance, India and Pakistan fall into that category.

It should be noted that, in addition to the 41 countries included, we tested around 20 *other* countries for viability for this study. We largely excluded small island states from the sample, with the exception of Haiti and the Comoros, as we previously studied Pacific small islands states (Redeker et al., unpublished). The survey was conducted in three waves as depicted in Table 1. The survey was available in seven languages in total: English, French, Spanish, German, Italian and Portuguese. These languages as spoken by at more than 1.3 billion people as a first language and likely several billion more as an additional fluent language.

[3] Partly, but not completely, the timing of the sampling in Switzerland may have driven ad prices (including December 2022) but this has not been analyzed in detail. In general, the less developed an online advertisement market (and presumable e-commerce in general), the more affordable advertisement to recruit for online surveys. Low costs have been demonstrated in Kenya, Venezuela and a number of other countries with relatively low incomes.

	Languages	*Number of countries*	*Dates of data collection*
Wave 1	English	16	9 Nov 2022 – 19 Jan 2023
Wave 2	English, French, Italian, German	1	12 Dec 2022 – 16 Jan 2023
Wave 3	English, French, Spanish, Portuguese	24	11 Feb 2023 – 31 Mar 2023

Table 1: Waves of data collection

Conducting the research across a certain period (here: more than four months) is not unusual for survey research at all. For instance, each Wave 7 of the *World Value Survey* started in January 2017 and ended in December 2021. In principle, some research questions, for instance about electoral preferences, require a narrower period of data collection. An ad-driven methodology does generally allow for scaling and thus also for data collection in a shorter period of time that an in-person interviewing of thousands of research participants does not allow for. However, due to the technical limitations of Meta's systems the ad buys were limited time and time again. New ad accounts only have a daily maximum spending of approximately US$ 50 (even this amount was held a secret by the call center agents frequently contacted; an equivalent in EUR was usually the limit at which daily ad spending stopped). These limitations were put in place to limit ad-based misinformation (but also hindered research). Only after running ads successfully to the limit over a "a few days" or "a few weeks" (call center was either not informed about the exact rules or not allowed to share the information) would the amount slowly and gradually increase. Even the previously existing accounts were set back to US$ 50 a number of times, after a non-collection on the credit card (the credit card issuer flagged the many small transactions).

The 41 countries eventually included in the study can be seen in Table 2. Largely, the countries can be divided into a subset in the Americas, in Africa and in Europe – mostly Eastern Europe. Only three countries were included from Asia, none from Oceania.

Region	***Country***	***Region***	***Country***
Africa	Algeria	Asia	Lebanon
Americas	Argentina	Africa	Madagascar
Europe	Belarus	Africa	Mali
Americas	Belize	Americas	Mexico
Americas	Bolivia	Americas	Nicaragua
Europe	Bosnia & Herzegovina	Africa	Nigeria
Americas	Brazil	Americas	Paraguay
Africa	Burkina Faso	Americas	Peru
Africa	Cameroon	Europe	Poland
Americas	Colombia	Europe	Romania
Europe	Croatia	Africa	Senegal
Africa	DR Congo	Europe	Serbia
Americas	Ecuador	Europe	Switzerland
Africa	Egypt	Africa	The Comoros
Europe	Georgia	Asia	The Philippines

Africa	Ghana	Africa	Togo
Africa	Guinea	Africa	Tunisia
Americas	Haiti	Europe	Turkiye
Europe	Hungary	Americas	Uruguay
Asia	Indonesia	Americas	Venezuela
Africa	Kenya		

Table 2: Countries included in study

Entailed in the project proposal but after attempts rejected as a viable source for survey participants are the following countries: South Africa, Gabon, Côte D'Ivoire, Congo, Angola, India, Thailand, Pakistan, Vietnam, Malaysia, El Salvador, Portugal, Spain and the United States. In total, 41 instead of 40 countries are represented in the final sample.

The questionnaire entailed both demographic and substantive questions. It also entailed a survey experiment that has so far not been analyzed and questions put on through REMIT-related funding (about geopolitics and technology). The demographic questions were the following:

- Country of residence
- Canton/region/state/etc. of residence
- Age
- Gender
- Location (urban/rural)
- Political orientation (how conservative?)
- Education
- Income

The substantive questions related to platform governance, and institutional trust.

- Trust in UN, African Union, EU, FIFA, World Bank, WHO, Catholic Church, own government, own parliament, own constitution, own courts, NGOs, Meta, Volkswagen AG, BBC, journalists in own country, academics in own country
- What topics should be included in the Global Digital Compact: protecting children online, fighting hate speech online, protecting privacy online, no censorship online, right to encryption of data, cultural diversity online, more innovation – less regulation, protection for intellectual property, network neutrality, open-source software, none of the above
- Who should the UN listen to when drawing up the Global Digital Compact? Governments, academics, NGOs, businesses, technical experts, citizens, none of the above
- In reality, who does the UN listen to when drawing up the Global Digital Compact? Governments, academics, NGOs, businesses, technical experts, citizens, none of the above
- Do you think that the protection of the following human rights has either increased or decreased because of the Internet? Right to privacy, freedom of expression, right to information, Security of person, equality before the law
- How often do you use Facebook/Instagram?

- How often do you do the following things on Facebook/Instagram? Browse content, comment on other users' content, read news, exchange private messages, create content/posts, I do not use the platform
- How concerned are you about the following topics concerning social media? Spam, bots, misinformation, bullying, surveillance, advertisement, hate speech, censorship
- Experiment: Control or one of nine treatment conditions related to content moderation
- In general, how much do you trust Meta with content moderation?
- Have you ever reported inappropriate content on Meta's platforms? If so, how would you rate the reporting process? Do you have any specific comments about the process?
- Have you ever been reported for inappropriate content on Meta's platforms? If so, how would you rate the moderation process? Do you have any specific comments about the process?
- How much do you agree with the following statement? I feel comfortable posting on Facebook/Instagram.
- How much do you agree with the following statement? I trust that other users are fairly reporting my content.
- How involved should the following actors be in making [enforcing/adjudicating appeals related to] the rules for what content can be posted/should be removed on Instagram and Facebook? Meta, civil society/NGOs, state institutions (parliament, government, courts), academics, international organizations, journalists, users
- If you had the choice, in which country should a body that makes final content moderation decisions for your country be located? Your country, other country, no preference
- Should human moderators or AI make rules in these cases? Hate speech, spam, pornographic content, child sexual abuse material, copyrighted material, content that hurts some people's sensibilities, violent content
- How concerned are you that the following actors know things about you, based on what you post and what you do on social media platforms? Private companies (the platforms themselves and advertisers), your government, your friends and family, your employer or school

A number of questions were randomly assigned and not shown to the entire sample (but rather to a quarter, a third or half of the full sample).

4. Sample Characteristics

This section describes the final sample used for analysis. After the data collection was concluded, a full dataset was compiled. Only complete responses were exported, then those removed that indicated (a) non-agreement with the terms of participation, (b) age under 18, or (c) residence in another country but those included in the respective wave of data collection. After these adjustments, 16,865 valid responses from individuals residing in 41 countries were retained (see Figure 1). The country-level samples range from 164 respondents on the Comoros to 616 in Kenya.

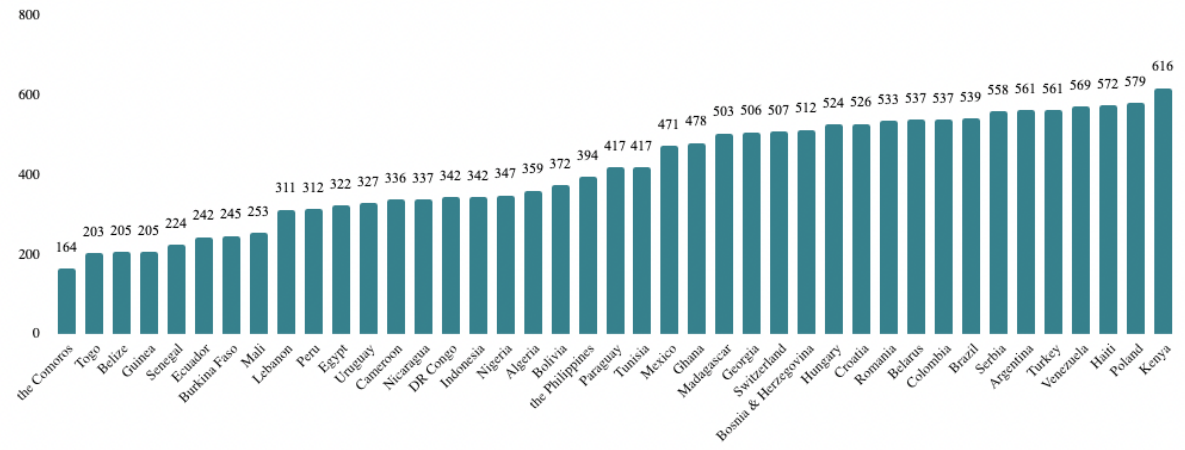


Figure 1: Sample size by country

In terms of gender and age, quota sampling did work very well in some countries and less well in others. For the example of gender, see Figure 2 for an overview. In countries such as Colombia, Venezuela and Switzerland, a balanced gender quote (between those identifying as female and those identifying as male) has been achieved. In other countries, especially many of those in which getting a sample for a reasonable cost has been a challenge in the first place, a balanced sample has not fully been achieved (neither for age, nor for gender). Most of those countries are located in Africa, with countries in the Americas and Eastern Europe generally being easier to sample a quota-adhering sample. In later analyses, different forms of weighting can make up or the lower number, e.g., of female respondents. Such weighting will be applied later on, following Lumley (2020).

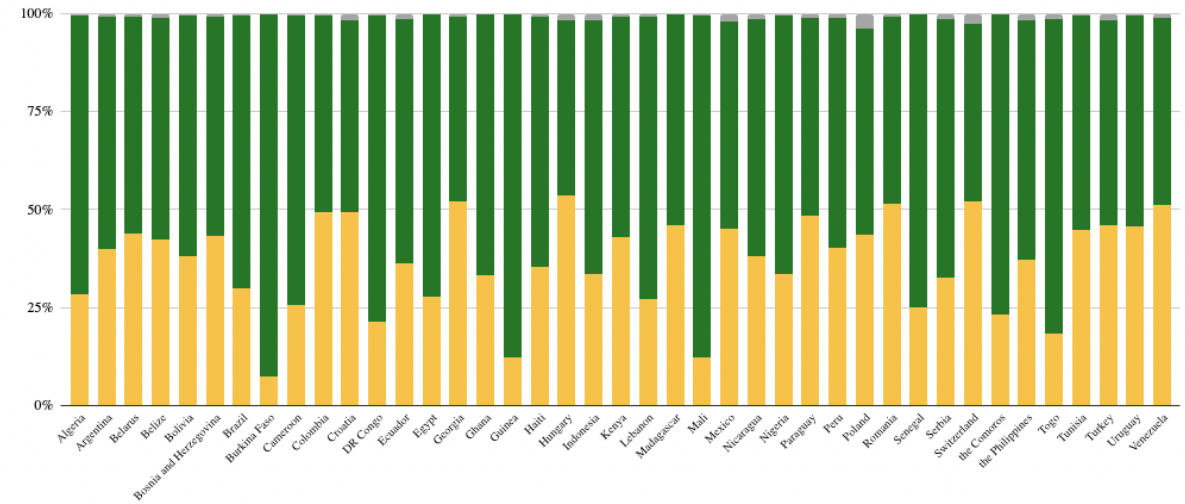


Figure 2: Sample by gender self-identification (green: male, yellow: female, grey: other)

The original plan for the study was to include 40 countries with 22,000 or more valid responses from these (either sampling 500 or 1,000 respondents per country). The 16,865 valid responses entailed in the final dataset fall short of this. The reasons are manifold and exclusively relate to the costs of advertisement that have been largely underestimated based on previous studies. While cost per click increases cannot be ruled out (i.e., inflation of the costs of ads), there are

mainly three reasons for the higher-than-expected cost per valid response. First, more countries than expected had to be replaced from the original planned lineup and the replacement candidates had to be extensively tested with many underperforming in these tests (but funds had to be allocated to these, too). Second, it turned out to be more difficult to collect questionnaires in a way that adheres to the set quotas in some countries. In a number of countries in Africa, recruiting female participants for the survey turned out to be extremely expensive (a female respondent submitting the form from the Democratic Republic of Congo cost nearly as much as a respondent in Switzerland). The same holds true for age of the respondents. I decided to rather strive for a more balanced sample (in most of the countries this worked), than to have more of the same (or similar) respondents, i.e., young men. Third, a few anchor countries that I did not want to replace in the data collection turned out to be more expensive to gather data. This category of cases includes Switzerland, Lebanon, the Philippines and Indonesia (the first being the only Western European case and the latter being the only three remaining countries in Asia).

Next, what remains to be addressed with respect to the sample composition, is how the sample compare to who uses Instagram and Facebook and can thus be reached by the advertisements, and to the countries' residential population. While the latter information can be attained through census data, in this case – for the sake of simplicity – based on data from one source (World Bank, 2022A), the former is drawn from the Meta Ads Manager[4] – retrieved at the time of data collection, in February 2023. Table 3 shows for all 41 countries how the share of female individuals in the sample differs from the share of female users among the total amount of users that can be reached through ads (virtually all users of Facebook and Instagram in one country combined) and again from the share of females within each of the countries sampled. While the sample contains three categories of gender self-identifications (male/female/other), no "other" identification exists in census data (at least not in all countries). Data from the Meta Ads Manager would allow to deduct the share of the "other" category, by subtracting from the high-end estimate of the total usership the high-end estimates of the female *and* the male user estimates.

As can be seen in Table 3, the quota aims, which were based on census distribution not Meta user distribution, have been reasonably well met in a number of cases. The deviations of the sample characteristics from the Meta usership and the census data respectively can be seen in the last two columns of the table (*Dev_meta* and *Dev_country*). For instance, the samples from Colombia, Croatia, Georgia, Hungary, Paraguay, Romania, Switzerland and Venezuela each show a deviation of only two percent or less. They may approximate the aimed for quotas as share of the overall population well but that does not mean that they necessarily fit the usership of Meta's platforms well. They do, however, track these, too – logically – where a high share of the population is also a user of Meta's platforms. An overview of the reach of Facebook and

[4] The Meta Ads Manager produces a range of the population that can be targeted with ads (high-end vs. low-end estimates). It should be noted that for the share of female users, the high end of female users has been subtracted from the high end of total users. In addition, the Meta provided data here applies to the population above the age of 18 (as did the sampling for the survey). The census data is for the female share of the total population (all ages).

Instagram (combined) can be seen in Table A1. Country samples with a bad gender balance tend to be countries with a lower degree of uptake of Meta's two platform services, including Burkina Faso, the Democratic Republic of Congo, Guinea and Togo.[5]

Country	*Sample_female*	*Meta_female*	*Country_female*	*Dev_meta*	*Dev_country*
Algeria	28.41%	36.88%	49.10%	-8.47%	-20.69%
Argentina	40.11%	54.74%	50.50%	-14.63%	-10.39%
Belarus	43.76%	60.00%	54%	-16.24%	-10.24%
Belize	42.44%	52.42%	49.70%	-9.98%	-7.26%
Bolivia	38.17%	47.37%	49.90%	-9.20%	-11.73%
Bosnia and Herzegovina	43.36%	48.66%	50.80%	-5.30%	-7.44%
Brazil	30.06%	54.03%	50.90%	-23.98%	-20.84%
Burkina Faso	7.35%	34.15%	50.20%	-26.80%	-42.85%
Cameroon	25.60%	40.00%	50.10%	-14.40%	-24.50%
Colombia	49.53%	51.00%	50.70%	-1.47%	-1.17%
Croatia	49.24%	46.15%	51.30%	3.09%	-2.06%
DR Congo	21.35%	31.15%	50.40%	-9.80%	-29.05%
Ecuador	36.36%	50.75%	50.10%	-14.38%	-13.74%
Egypt	27.64%	40.35%	49.40%	-12.72%	-21.76%
Georgia	51.98%	53.13%	53%	-1.15%	-1.02%
Ghana	33.26%	42.86%	50.10%	-9.59%	-16.84%
Guinea	12.20%	38.26%	50.50%	-26.06%	-38.30%
Haiti	35.31%	48.00%	50.50%	-12.69%	-15.19%
Hungary	53.63%	52.94%	52%	0.68%	1.63%
Indonesia	33.63%	45.79%	49.70%	-12.17%	-16.07%
Kenya	43.02%	48.41%	50.40%	-5.39%	-7.38%
Lebanon	27.33%	45.00%	51.50%	-17.67%	-24.17%
Madagascar	46.12%	45.45%	49.90%	0.67%	-3.78%
Mali	12.25%	23.49%	49.50%	-11.24%	-37.25%
Mexico	45.01%	51.72%	51.20%	-6.71%	-6.19%
Nicaragua	37.98%	50.00%	50.70%	-12.02%	-12.72%
Nigeria	33.43%	38.12%	49.50%	-4.69%	-16.07%
Paraguay	48.44%	50.00%	49.80%	-1.56%	-1.36%
Peru	40.38%	49.62%	50.50%	-9.23%	-10.12%
Poland	43.70%	52.03%	51.60%	-8.34%	-7.90%
Romania	51.59%	52.46%	51.60%	-0.86%	-0.01%
Senegal	25.00%	32.35%	50.80%	-7.35%	-25.80%
Serbia	32.62%	47.83%	52.10%	-15.21%	-19.48%
Switzerland	52.27%	48.28%	50.30%	3.99%	1.97%
the Comoros	23.17%	39.49%	49.80%	-16.32%	-26.63%
the Philippines	37.31%	53.27%	49.20%	-15.96%	-11.89%
Togo	18.23%	28.92%	49.70%	-10.69%	-31.47%
Tunisia	44.84%	48.65%	50.60%	-3.80%	-5.76%

[5] Compare Table 3 and Table A1.

Turkiye	45.99%	42.51%	49.90%	3.48%	-3.91%
Uruguay	45.87%	53.13%	51.50%	-7.25%	-5.63%
Venezuela	51.14%	53.37%	50.50%	-2.23%	0.64%

Table 3: Share of female individuals in sample, Meta population and country population

A specific sub-national unit quota sampling has not generally been conducted for this study. This means that in many cases, the economic and cultural centers, the capital cities and other well-connected regions are over-represented. This could also later be corrected for using weights but it might not have to be done, depending on the analysis in question. In two cases, the sampling was corrected with strong quota enforcement. Next to the case of Belarus, this was done in the case of Switzerland. Figure 3 shows two distributions of populations of Switzerland based on a canton-level analysis: On the left, the distribution of the residential population of Switzerland according to the most recent census and on the right the share of the Swiss sample reporting to be residing in different cantons. This quota sampling proved quite effective overall, with German-speaking big-city cantons (Zürich, Basel) somewhat over-represented and French-speaking cantons somewhat under-represented.

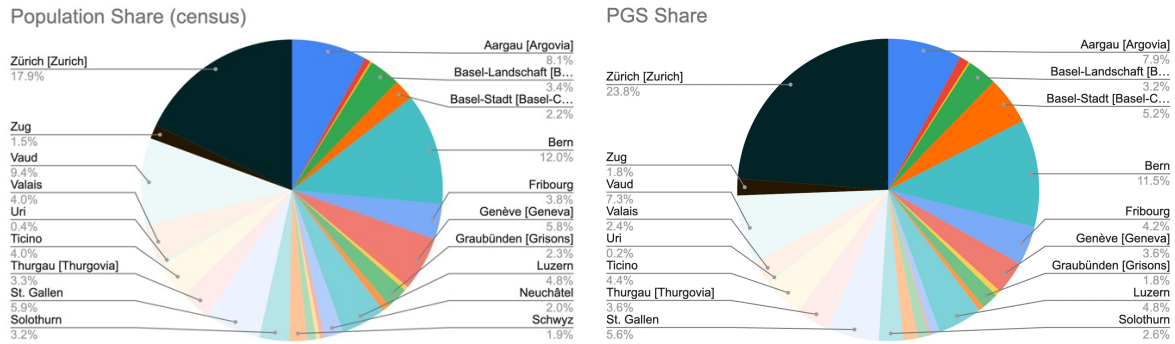


Figure 3: Share of Swiss population and sample population by residence in cantons

Overall, the sample did not quite meet the high aspirations in terms of size and quota-sampling I aimed at. Nonetheless, apart from the substantive findings discussed in section 5 of this report, there are a number of methods-related conclusions that can be drawn. The main insight is that a well-prepared survey recruitment strategy using social media advertisement may be a useful alternative to other established ways of quota sampling. While country-level and even comparative sampling studies have been conducted (as indicated), there has – to my knowledge – not been such a comprehensive study so far. The country-level samples show that quota sampling can work well for gender, age and even sub-national subdivisions such as oblasts and cantons. The difficulty is to make a trade-off between high sample size and better quota sampling against the background of scarce resources.

5. Study Findings

This section follows the research questions outlined in section 2. The specific question on the questionnaire relating to the research question of the project pertaining to *alternative governors* was phrased in the following:

How involved should the following actors be in making [enforcing/adjudicating appeals related to] the rules for what content can be posted/should be removed on Instagram and Facebook?

Respondents rated all seven actors included in the study by each providing a score on a 10-point scale (1-10) from "not involved at all" to "highly involved". Here, I first report full sample average confidence in seven different actors, including a subdivision by gender, for each of the functional roles defined.

5.1. Trust in Seven Alternative Governors in Different Roles (RQ1)

The data shows that respondents express a different level of trust in different functional roles in the context of content moderation. For the "legislative" making of rules for content moderation, trust is highest in users (6.7 out of 10) and Meta itself (6.4 out of 10). The lowest trusted for making rules for content moderation are the respondents' respective national governments (4.0 out of 10). Figure 4 also shows that male and female respondents differ slightly regarding their views on certain actors. Male respondents tend to trust parliaments more; whereas female respondents have more confidence in Meta and NGOs when it comes to the making of rules. I will need to establish later whether these differences are significant.

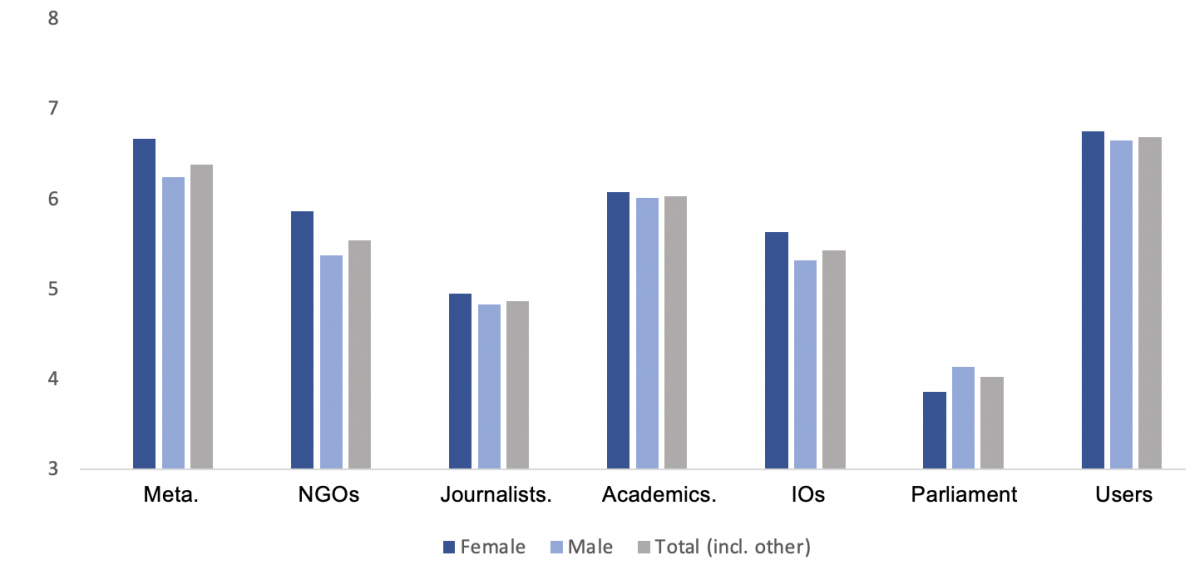


Figure 4: Preference for different actors with regard to "making rules" for content moderation

A similar tendency can also be observed for levels of trust in these seven actors with regard to the *enforcement* of rules in content moderation and the *adjudication* of rules (see Figure 5 and Figure 6). The differences between male, female and other respondents are similarly little pronounced but generally pointing in the same direction when it comes to the other two functional roles these actors can take.

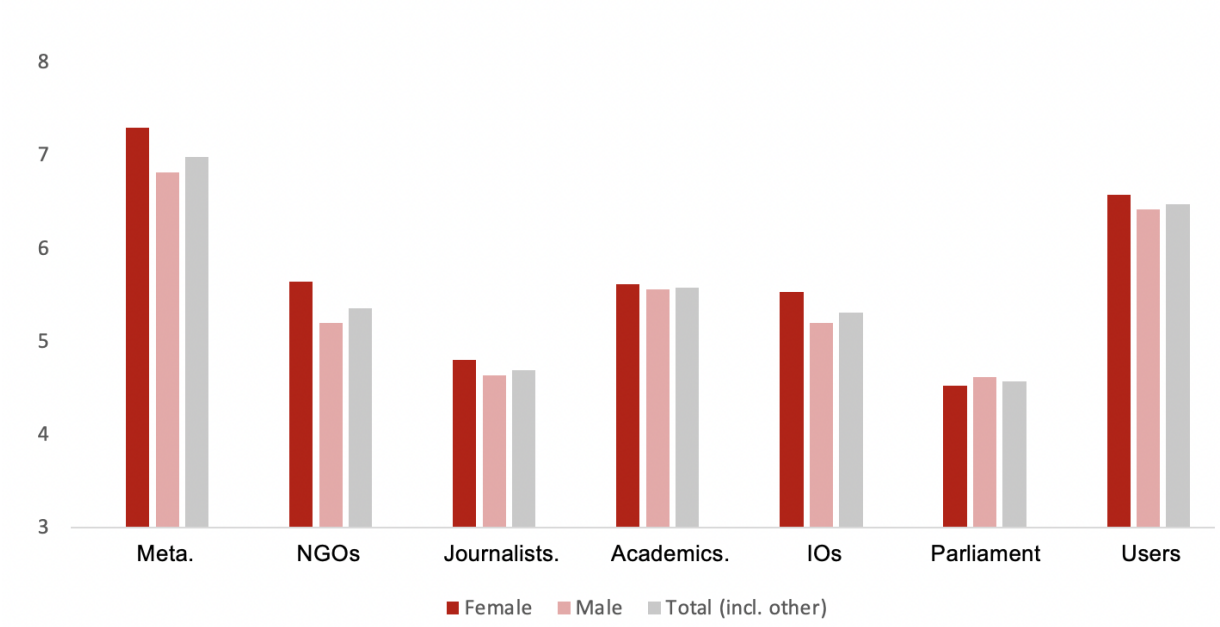


Figure 5: Preference for different actors with regard to enforcement of content moderation rules

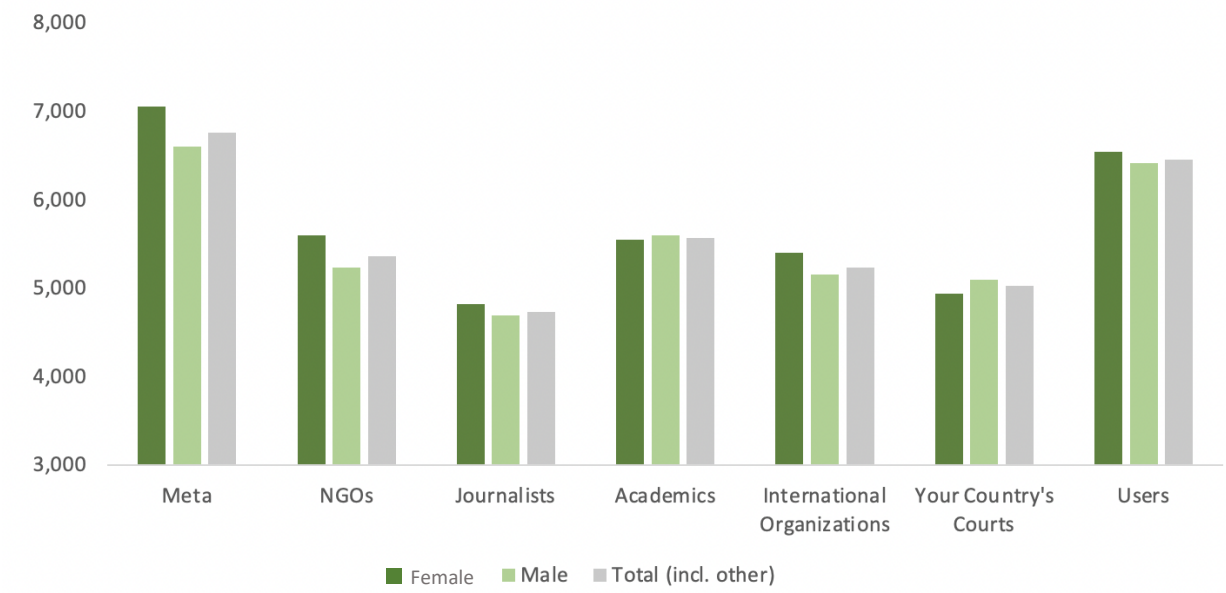


Figure 6: Preference for different actors with regard to the adjudication of appeals

The strong overall trust in Meta for all three functions in content moderation comes as a surprise given the strong criticism in international media and by civil society groups concerning the company's handling of content moderation. Users stick out as a category of actors often overlooked when it comes to content moderation on platforms. While they do play a certain role in supporting content moderation in the case of Meta, usually by flagging content they deem inappropriate, they do not have the same role that community moderators have in pre-platform or smaller forums or recently attained as part of the growth of Mastodon. Academics and NGOs are also generally relatively trusted when it comes to the three functions. Overall,

respondents have the lowest level of confidence in their own country of residence's state institutions. International organizations and journalists both land in the lower mid-field of potential alternative governors.

Having reported overall trends of confidence in different actors with respect to content moderation roles, next, I will illustrate the relative differences between the trust they enjoy with respect to different functional roles they engage in. Figure 7 illustrates that there appear to be relatively strong differences depending on the *functional role* in which actors act. For instance, respondents indicate relatively lower preference for Meta when it comes to the context of making the rules, whereas preferences for Meta are higher for enforcing and also adjudicating the rules. The most striking example of how respondents preferences differ based on different contexts is the case of state institutions. Respondents do not prefer the making of content moderation rules by parliament (4.0 out of 10), whereas they trust more in the state executive to enforce rules (4.6 out of 10) and even more in courts to adjudicate the rules (5.0 out of 10). Academics, too, are seen as more trusted to develop the rules for content moderation than to actually engage in enforcement or adjudication, albeit all on a relatively high level. What should further be established with these differences is if – after weighting – they are indeed significant.

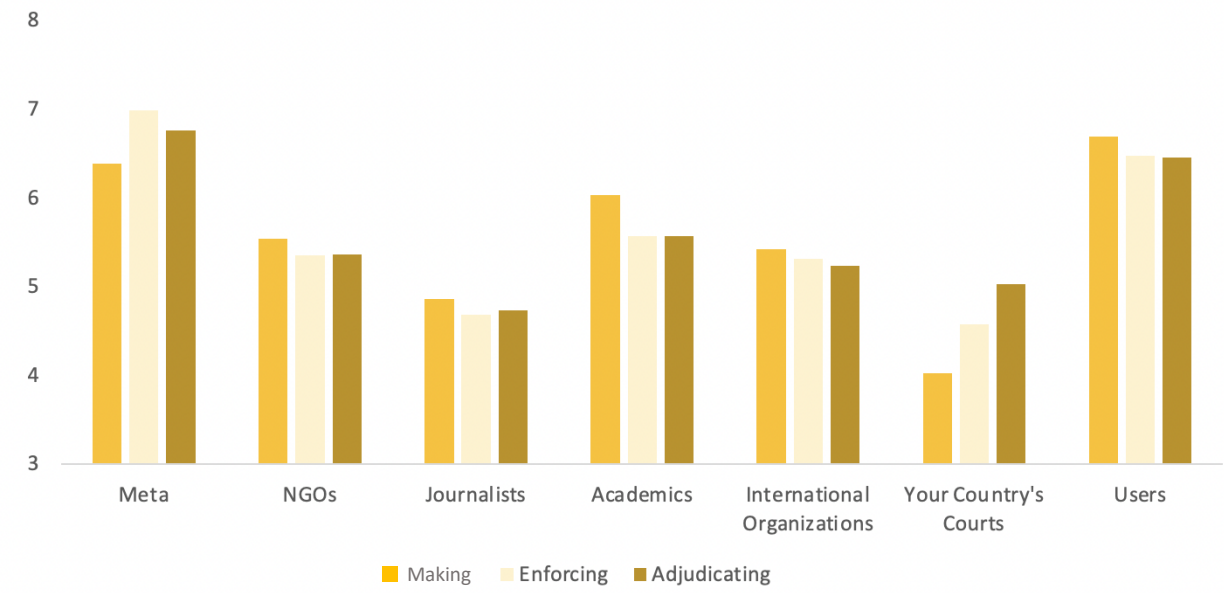


Figure 7: Preference for different actors with regard to different functional roles

Another aspect that requires analysis is that Figures 4-7 merely show average values for the entire dataset – with only a breakdown into gender groups. A meaningful comparison between countries can be conducted.

5.2. Variation of Preferences across (RQ2)

This subsection discusses how respondents in different countries prefer content moderation to occur based on the choice of different actors responsible. Based on the large number of variables, not all data can be explored (41 countries x seven actors x three roles).

First, I report data from the three countries with each the highest and lowest average scores for confidence in Meta, the traditional arbiter of such rules. In contrast, I report data on the confidence in state institutions. Data displayed in Figure 8 is for confidence in Meta and the respective country's parliament in making the rules for platform content moderation.

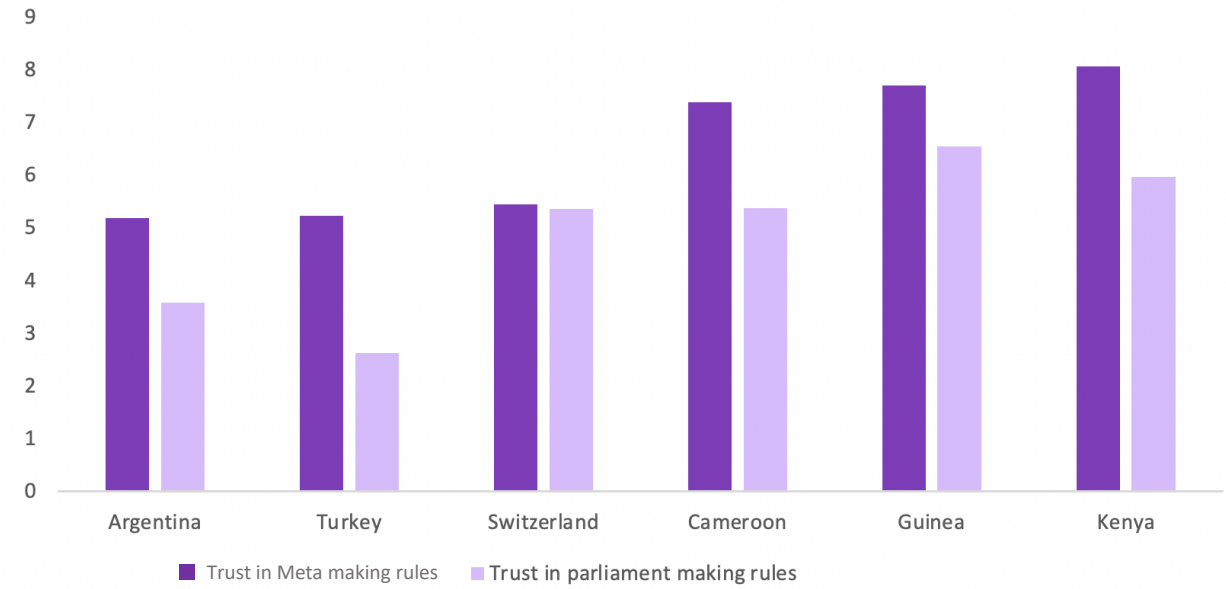


Figure 8: Preferences for Meta and the country's parliament actors with regard to making of content moderation rules

In Figure 8, country-level samples are shown that do not only differ much on how much respondents prefer Meta to make content moderation rules. Preference for Meta for this functional role is relatively high in Cameroon, Guinea and Kenya, while relatively low in Turkiye and Argentina and Switzerland. However, in comparison to Meta, parliaments are generally less preferred, apart from Switzerland where both are similarly preferred for legislative roles when it comes to content moderation on platforms. I conducted another comparison by looking at preference for state institutions in content moderation on the country-level, keeping the above-mentioned country selection. Figure 9 shows data for these six countries for the making, the enforcement and the adjudication of content moderation rules by state institutions (parliament/government (e.g., police)/courts).

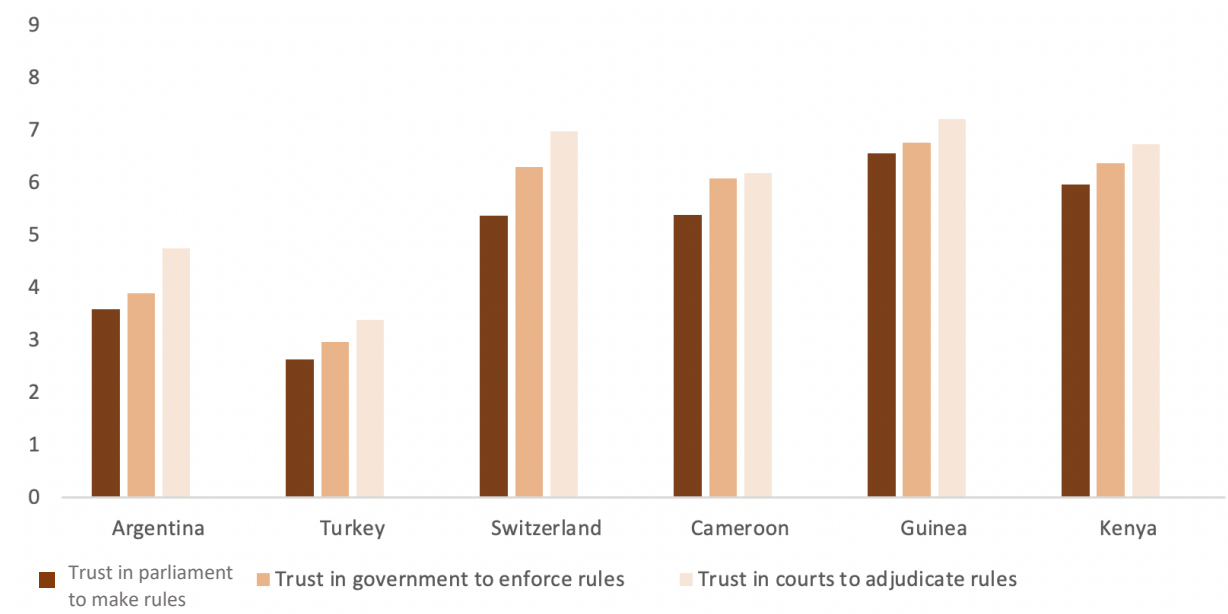


Figure 9: Preference for state institutions concerning three functional roles

The data (before data weighting and significance tests) shows that role-specific preferences are similarly pronounced on the country level as compared to the full sample averages. Swiss respondents display a particularly high contextualization of preferences for state institutions in content moderation: Their confidence in their nation state to adjudicate the decisions made concerning content moderation is much higher than when it comes to making of rules, application of rules by the government almost precisely in between. I did not calibrate the data taking into account the different national propensities to prefer institutions in general, which could alter results between countries but not within countries.[6]

Figure 10 shows the 41 countries sampled from on a map colored by the degree to which respondents want involvement of their parliaments in the process of content moderation. The mean value for this is 4.02, Belarus scores lowest with a value of 2.10 and Burkina Faso scores highest with a value of 6.67 on the question of whether respondents want their national parliaments be more or less involved in making the rules for content moderation. The mean of the scale is a 5.5 (on a 10-point Likert scale). What can be seen is that respondents in countries in Sub-Saharan Africa (but not Northern Africa), South East Asia (the Philippines and Indonesia), Haiti and Switzerland have more than average preferences for involvement of their national parliament in content moderation. In the remainder of countries, especially in Eastern Europe and around the Mediterranean, as well as in the Americas, respondents show a less-than-average inclination to have their parliaments possess a greater role in content moderation.

[6] Such calibration could take into account the fact that people in some countries are more likely to want many actors involved in content moderation whereas people in other countries would rather have the functions be conducted by fewer countries. This may well be due to trust (see 5.3). We know that general trust in institutions differs across countries (e.g., from the World Value Survey) and there are sometimes good reasons for different levels for this (e.g., due to civil war, authoritarian government, "well-governed" country according good governance criteria). Hence, calibration can control for different macro-situations that affect confidence in a number of institutions that could engage in content moderation but that may not be particularly helpful.

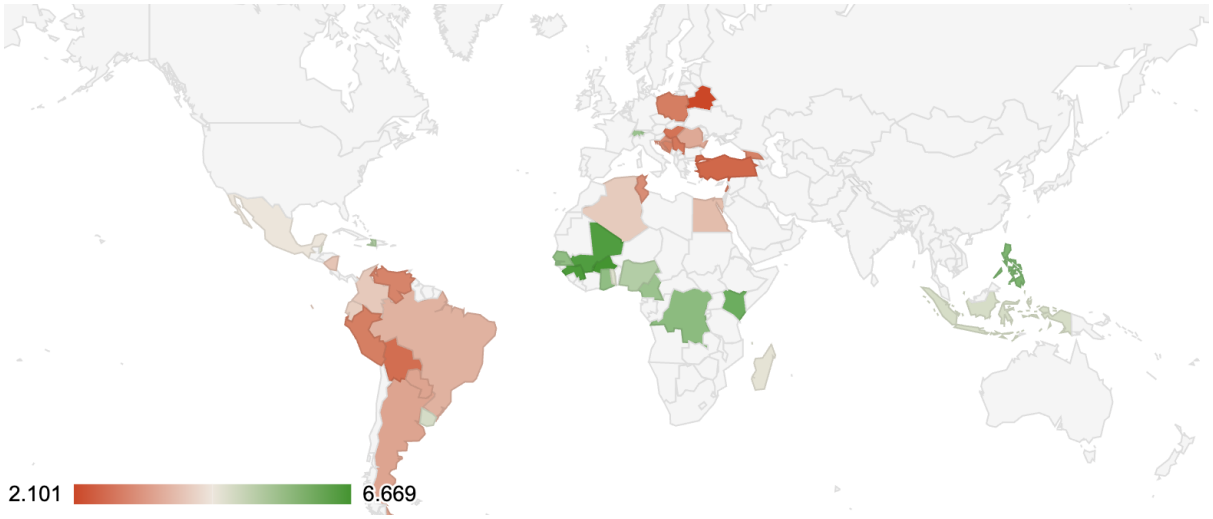


Figure 10: Mapping preferences for parliament making rules (green: high, red: low)

These findings about national level differences require explanations. While this report does not include a full-fledged quantitative analysis taking into account a variety of other outside factors, including potentially country-level variables and respondent-level variables in one model, it does next discuss the correlation between trust in institutions and preferences for these institutions (or lack thereof) in the context of platform content moderation.

5.3. Relationship Between Content Moderation Preferences and Trust in Institutions (RQ3)

The third research question asks about the drivers of the trends unearthed in RQ1 and RQ2, especially with respect to confidence or trust in institutions. For this, in Figure 11, I clustered, on a respondent-level, respondents with low to high levels of trust in their national parliament (on a scale 1-10). I then display the level of preference in parliament making rules on the y-Axis. In the mean of the sample (n=16,865), trust in national parliaments stands at 3.408 on a 1-10 scale with a standard deviation of 2.81. The mean preference for parliaments making rules on social media is at 4.024 with a standard deviation of 3.21 (also on a 1-10 scale).

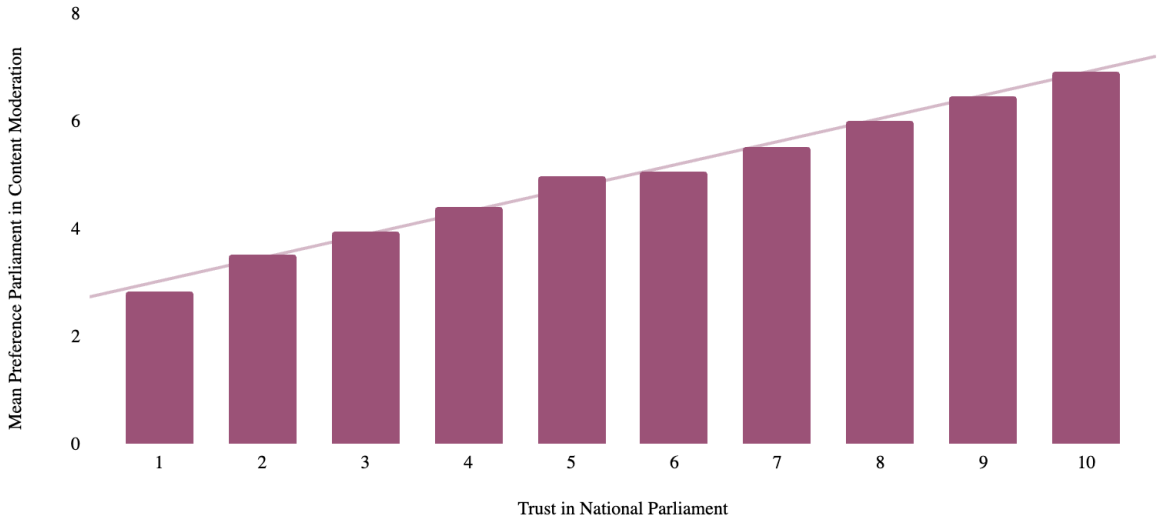


Figure 11: Relation between trust in national parliaments and preference for national parliaments in content moderation functions

The data shows a clear correlation between trust in national parliaments and preferences for national parliaments being more involved in making the rules for content moderation. Interestingly, in Figure 12, data shows that the effect of trust in Meta on preferences for Meta is generally not as high as in the case of national parliaments (and other institutions for that matter). A possible explanation may be that – unlike some of the other *alternative* actors in platform content moderation, the platform is "naturally" seen as the institution to engage in content moderation. The roles for actors such as NGOs, academics and state institutions are newer ones that may also, understandably, not be obvious to all respondents. Even here, trust in Meta does on average predict preferences for Meta to be more involved in all three functional roles in content moderation.

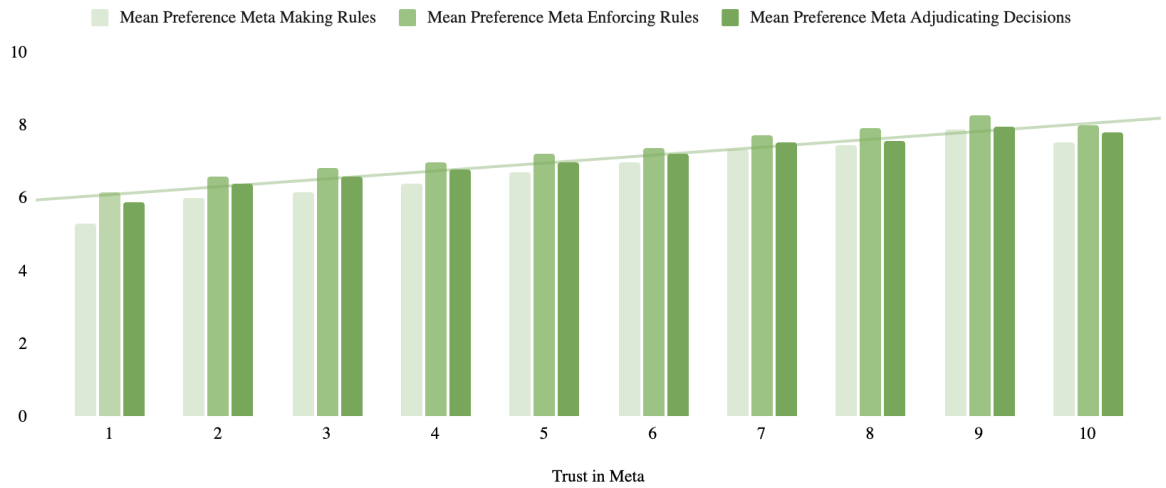


Figure 12: Relation between trust in Meta and preference for Meta in three different content moderation functions

Another analysis encapsulates two different approaches by examining the relationship between trust in institutions and trust in these institutions to engage in social media content moderation on the one hand, and by conducting a comparison between the only strictly Western European country in the sample and the global average of 41 countries on the other hand.

Figure 13 shows the general (average) levels of trust in six different institutions, including the United Nations/IOs, the respective national parliaments, civil society/NGOs, Meta Inc., journalists (in the respective country) and academics (in the respective country). These data are displayed for Switzerland (dark red) and the full sample of 41 countries (dark blue). The respondents answered on a 10-point scale from no trust to complete trust. The figure also displays to what extent the same institutions should be involved in social media content moderation (also on a 10-point scale), for both the Swiss sample (506 respondents) and the full sample (16,865 respondents).

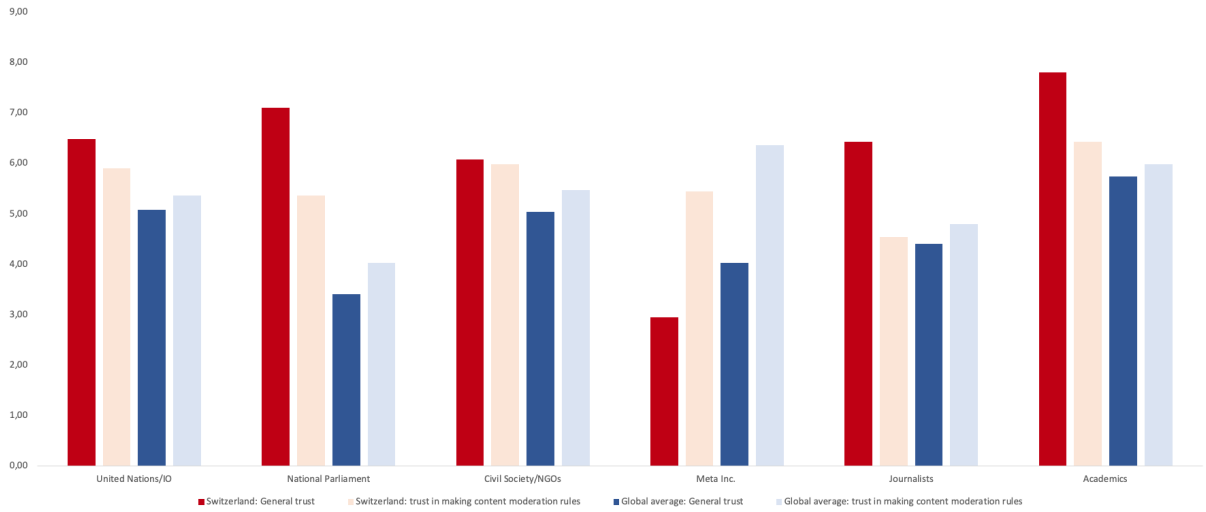


Figure 13: Comparison between general trust in institutions and trust in these institutions to make the rules for content moderation, and between Switzerland and global averages

The data shows that Swiss respondents trust institutions more than all respondents across 41 countries. The Swiss sample trusts its parliament and Swiss academics most, followed by journalists, the United Nations/IOs and civil society/NGOs. The sole exception to this is Meta Inc., which sees very little trust by the Swiss sample. When it comes to these institutions being involved in content moderation, the Swiss sample indicates that Meta Inc. should be involved in making the rules quite a bit (but only as much as other actors), suggesting at least increased levels of trust when it comes to this specific role. A similar increase can be seen in the global sample of 41 countries but here from a relatively higher base (because Meta Inc. is not globally as mistrusted as it is in Switzerland), landing Meta Inc. on top of the ranking of actors (excluding users) in terms of who is – on average – most desired to make the rules for content moderation. Further analysis with the dataset from the Platform Governance Survey 2022 and additional outside sources could explore other factors besides generalized trust in institutions behind respondent preference for content moderation actors.

Bibliography

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* New Haven: Yale University Press.

Haggart, B., & Keller, C. I. (2021). Democratic legitimacy in global platform governance. *Telecommunications Policy,* 45(6), 102152.

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131, 1598–1670.

Lumley, T. (2020). Survey: analysis of complex survey samples. R-Paket Version 4.0.

Pötzschke, S., & Braun, M. (2017). Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. *Social Science Computer Review*, 35(5), 633-653.

Redeker, D. & Sturm, I. (2019). ICT skills in Small Island Developing States: ICT capacity building, economic opportunities and brain drain. *Digital Skills Insights 2019*. Geneva: ITU.

Redeker, D., Sturm, I., Cohausz, F., and van der Neut, B. (unpublished). Pursuing Pacific Partnerships: Aid, Democracy and Public Opinion. Unpublished working paper presented at ISA 2022.

Rosenzweig, L. R., Bergquist, P., Hoffmann Pham, K., Rampazzo, F., & Mildenberger, M. (2020). Survey sampling in the Global South using Facebook advertisements. https://osf.io/preprints/socarxiv/dka8f/

Suzor, N. (2019). *Lawless. The Secret Rules That Govern Our Digital Lives.* Cambridge: Cambridge University Press.

UNESCO (2023). Guidelines for regulating digital platforms: a multistakeholder approach to safeguarding freedom of expression and access to information. CI-FEJ/FOEO/3 Rev.

World Bank (2022A). Population, female (% of total population). United Nations Population Division's World Population Prospects. https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS

World Bank (2022B). Population, total. United Nations Population Division's World Population Prospects. https://data.worldbank.org/indicator/SP.POP.TOTL

Zhang, B., Mildenberger, M., Howe, P. D., Marlon, J., Rosenthal, S. A., & Leiserowitz, A. (2020). Quota sampling using Facebook advertisements. *Political Science Research and Methods*, 8(3), 558-564.

Appendix

Country	Lower-end total	Country population	*Meta Ads Reach*
Algeria	23,900,000	44,900,000	53.23%
Argentina	32,300,000	46,230,000	69.87%
Belarus	3,400,000	9,200,000	36.96%
Belize	212,600	400,000	53.15%
Bolivia	6,400,000	12,200,000	52.46%
Bosnia and Herzegovina	1,600,000	3,230,000	49.54%
Brazil	139,100,000	215,310,000	64.60%
Burkina Faso	2,000,000	22,670,000	8.82%

Cameroon	3,800,000	27,910,000	13.62%
Colombia	34,000,000	51,870,000	65.55%
Croatia	2,200,000	3,850,000	57.14%
DR Congo	5,200,000	99,010,000	5.25%
Ecuador	11,400,000	18,000,000	63.33%
Egypt	38,300,000	110,990,000	34.51%
Georgia	2,800,000	3,710,000	75.47%
Ghana	6,000,000	33,470,000	17.93%
Guinea	2,100,000	13,850,000	15.16%
Haiti	2,100,000	11,580,000	18.13%
Hungary	5,800,000	9,580,000	60.54%
Indonesia	146,400,000	275,500,000	53.14%
Kenya	10,700,000	54,020,000	19.81%
Lebanon	3,400,000	5,480,000	62.04%
Madagascar	2,800,000	29,610,000	9.46%
Mali	1,700,000	22,590,000	7.53%
Mexico	81,600,000	127,500,000	64.00%
Nicaragua	3,100,000	6,940,000	44.67%
Nigeria	29,000,000	218,540,000	13.27%
Paraguay	3,700,000	6,780,000	54.57%
Peru	22,300,000	34,040,000	65.51%
Philippines	71,500,000	115,550,000	61.88%
Poland	20,900,000	37,560,000	55.64%
Romania	10,300,000	18,950,000	54.35%
Senegal	2,900,000	17,310,000	16.75%
Serbia	3,900,000	6,760,000	57.69%
Switzerland	4,900,000	8,760,000	55.94%
The Comoros	173,000	836,000	20.69%
Togo	736,800	8,840,000	8.33%
Tunisia	6,300,000	12,350,000	51.01%
Turkiye	56,200,000	85,340,000	65.85%
Uruguay	2,700,000	3,420,000	78.95%
Venezuela	13,900,000	28,300,000	49.12%

Table A1: Reach of Meta's ads on Facebook and Instagram (combined, low-end estimate); total number of users over 18 divided by the overall population of a country. Own calculations based on data from World Bank (2022B) and Meta Ads Manager data.