Due: March 5th

Tie it into – EU register of data intermediation services

What the Grant is Looking For

- Initiatives that involve a novel legal, technological or participatory approach, and therefore can be learnt from. For example, a creative new approach to running a public dialogue or other deliberative process.
- Initiatives that have significant public policy relevance. For example, data intermediation services

Summary of 1st Meeting

Problem – Trust and Safety (T&S) professionals and researchers have problems tracking malicious content spreading across different platforms fast enough, resulting in serious outcomes, including genocide, rioting, and election interference. Policymakers globally are developing guidelines and regulations to ensure that companies are required to share data to measure online risks and harms. However, ample technical expertise is needed to track the spread of harmful content across various social platforms, and T&S resources are limited. There are tradeoffs to prioritize which online escalations platforms must address based on potential virality, societal impact, brand and reputational risk, and other consequences. The nature of how flagged content is spread across different platforms provides a valuable signal on prioritizing escalations. For example, if there are multiple flagged misleading political deepfakes, the deepfake that has been populated across many platforms, gaining ample views, should be considered more harmful than a comparable deepfake that has limited engagement and has only been shared on one platform. This is challenging to do internally as teams at platforms are siloed and platforms have historically not collaborated on such issues beyond child safety and terrorism content.

Solution – To address these problems, we've developed an easy-to-use platform for the community, called Arbiter, a tool that allows users to track how individual pieces of content are populated across various social platforms. A T&S researcher can use Arbiter to monitor how videos that contain political advertisements or disinformation on YouTube are spread across other platforms such as X/Twitter, Truth Social, Bluesky, and Meta-owned platforms, Facebook, WhatsApp, and Instagram. Notably, the researcher tracking this populated information does not need to have a robust technical background because Arbiter is designed to be user-friendly with natural language queries, and can empower vetted third-party researchers from academia, journalism, and newsrooms to scale their online safety research. Tracking how information spreads across multiple platforms is crucial because people are more likely to believe in falsehoods following multiple exposures.

The grant allows us to develop a harm-monitoring system, through participatory policy research using Arbiter. We have received commitment from our partners across journalism, civil society,

newsrooms, , and fact-checking who would like to build community focused features into Arbiter. . We have written up two prior reports using Arbiter independently, and seek to make this generally available for tech and policy researchers in industry and academia, especially students. We have run a Google and Mozilla backed fellowships program to advance responsible computing education in India with Arbiter and hope to use the grant to make this tool community-friendly and open-access. We aim to produce multiple reports for academic journals made in collaboration with the Data Empowerment Fund.

Questions for Application

1. **Name of Initiative**
    a. Arbiter

2. **In what country will your initiative have impact?**
    a. We aim to launch this initiative across India, Mongolia, and Bangladesh. Although Arbiter is agnostic to regional influences, we are focusing on the global south with our work, given decades of lived experience and significant experience building public health and AI tools for nonprofits operating in over a hundred villages, benefitting tens of thousands of people in India.

3. **Who is your primary contact?** ** *This is the person and organisation that would lead delivery of the initiative, and would be the recipient of funding from the Data Empowerment Fund.*
    a. One Fact Foundation (US 501 (c)(3) nonprofit, fiscal sponsor of SimPPL)

4. **Describe the problem you are working to solve, or your diagnosis of the status quo and why it must change.**

Trust and Safety (T&S) professionals and researchers have problems tracking malicious content spreading across different platforms fast enough, resulting in serious outcomes, including genocide, rioting, and election interference. Policymakers globally are developing guidelines and regulations to ensure that companies are required to share data to measure online risks and harms. The EU's Digital Service Act, the UK's Online Safety Act, and the US Surgeon General's Advisory Board have all acknowledged that professionals across academia, fact-checking organizations, and other civic society organizations must independently measure online risks and ensure platforms mitigate harm.

However, ample technical expertise is needed to track the spread of harmful content across various social platforms, and T&S resources are severely limited outside of large tech companies or coalitions. Platforms generally prioritize which online escalations to address by weighing the tradeoffs between their scope of harm, potential virality, societal impact, brand and reputational risk, and business impact.

In 3 years of working with them, we have seen platform professionals and civic society organizations have trouble efficiently tracking flagged content spread across different social platforms, even though understanding the spread of harmful content can inform how to prioritize escalations. For example, if there are multiple flagged misleading political deepfakes, the deepfake that has been populated across many platforms, gaining ample views, should be considered more harmful than a comparable deepfake that has limited engagement and has only been shared on one platform. Understanding how malicious content is spread across different platforms is challenging because siloed social media teams and companies seldom collaborate on such issues (child safety and terrorism being exceptional topics). This problem is exacerbated during elections where deepfakes showing candidates engaging in illegal activities and electoral misinformation are widely shared to reinforce a narrative and change voters' perceptions. This calls for technical solutions to identify the spread of misinformation and mark content as such, given over 80 elections globally in the year 2024.

5. Describe how your initiative is designed to address this problem or bring about this change. ** Please make this short and sweet. We will ask for information about specific activities, outputs and outcomes in a moment.

To address these problems, we've developed an easy-to-use platform called Arbiter, which allows users to track how individual pieces of content are populated across various social platforms. For example, a researcher can use Arbiter to see how videos that contain political advertisements or disinformation on YouTube are spread across other platforms such as Meta and Twitter. Notably, the researcher tracking this populated information does not need to have a robust technical background because Arbiter is intuitively user-friendly and can empower vetted third-party researchers from academia, journalism, and newsrooms to scale their online safety research. Tracking how information spreads across multiple platforms is crucial because people are more likely to believe fake content is true when they see it multiple times.

With this grant, we plan to test Arbiter with our various global community partners across journalism, fact-checking organizations, and newsrooms. Arbiter can help our current and past collaborators at the Wikimedia Foundation, Mozilla, The Sunday Times, Deutsche Welle (DW), Tech Global Institute, and IFCN-verified fact-checkers to foster interdisciplinary online safety research with a user-friendly tool. This grant will lead to reports for academic journals made in collaboration with the Data Empowerment Fund to spread awareness of Arbiter's potential to scale online safety research across social platforms for industry professionals and third-party watchdogs.

6. Tell us about the composition, experience and diversity of your team. ** *Also list any other organisations who would be involved in your initiative, such as delivery partners or advisors.*

We are a student-led research collective called SimPPL, fiscally sponsored by a US-based 501(c)(3) nonprofit, the One Fact Foundation, developing open-access civic integrity tools for journalists and media development organizations. SimPPL was founded in 2021 by Dr. Swapneel Mehta, a postdoctoral associate at MIT and Boston University, researching platform governance and free speech. Insights from teams of student researchers across the US, UK, and Global South drive our organizational decision-making. Our diverse team has received grants from Google, the Wikimedia Foundation, Mozilla, Amazon, the Goethe Institute, the Belfer Fellowship, the NYC Media Lab, and others. We have successfully built multiple student organizations that are training hundreds of students each year called DJ Unicode (https://djunicode.in) and the NYU AI School (https://nyu-mll.github.io/nyu-ai-school-2023/).

Our core team comprises members who are postdoctoral associates, full-time data scientists working with cloud infrastructure, ML engineers, production engineers, political scientists, professors, and policy researchers. We've fostered a global community that champions community-based participatory research.

SimPPL offers technical and personal mentorship to students from non-premier institutes and underserved communities while paying them to develop scalable Trust and Safety tools to protect the public from online harm. Many of our student contributors have faced economic constraints and gender-based limitations, and their lived experiences inform our mission to build civic integrity tools to protect global users from overlooked communities. We recently launched a fellowship program to invite 60+ new students from Tier-II and Tier-III institutes in India and abroad to contribute to our goal.

In three years, we have built technology products like Parrot, analyzing hundreds of millions of online posts for The Sunday Times in the UK, Tech Global Institute, and Sakhi, a Bengali-speaking WhatsApp chatbot, securing research pilot commitments with communities in Bangladesh with additional interest for a Marathi and Hindi version with nonprofits in India. In 2022, we identified 400k malicious accounts and reported them to integrity professionals at X/Twitter. Our work and commentary is featured by Google, Deutsche Welle, Fast Company, the Wikimedia Foundation (WikiCred).

7. Describe the maturity of your initiative. ** *Tell us how long it's been running, if you have participants involved already and any impact it has achieved to date (if any).*

We have successfully built Parrot, a precursor to Arbiter, a system to support the transparent investigation of coordinated networks of accounts on Twitter spreading manipulated information from state-backed Russian media providers by journalists from The Sunday Times. Our agenda was to step beyond reliance on conventional “third-party ratings” alone as a measure of malintent (e.g. the US saying Russian state-backed media spreads misinformation). We scaled Parrot to analyze hundreds of millions of tweets, filter tens of millions of accounts and identify at a set of about 400,000 coordinated accounts that were successfully reported through our conversations with the Integrity Leads for X, formerly Twitter, in 2022. We wrote a report about actor analysis, a draft (https://jhagrutlalwani.vercel.app/blog/network_analysis_simppl) of which has been viewed by thousands of visitors, in anticipation of its final publication.

Parrot led us to create tools that have collected 100M+ posts across Bluesky and Truth Social, 100M+ comments on YouTube, and nearly half a million instances of flagged content, resulting in 98 million interactions on Facebook. These efforts have informed teams at YouTube and Meta, which led to expanding our data access, given our previous work in reporting coordinated pages and groups spreading misleading information. Parrot helped us build the technical, multilingual, and multimodal understanding needed to create our new tool, Arbiter. We uniquely understand how to identify manipulated information, and the next step is understanding how that information and other types of malicious content are spread across different platforms.

SimPPL will be featured on the Google Cloud blog (past post) in addition to existing media mentions with Deutsche Welle, the NYC Media Lab, Wikimedia, and others. We have been invited to present our work at MIT, Oxford, X (Twitter), Meta (Facebook), and tens of conferences, including the International Conference on Machine Learning, NeurIPS, Mandiant mWise, Truth and Trust Online. We have presented our work at MIT, Oxford, Stanford, X (Twitter), and Meta, with awards totaling USD 150,000 from entities like Google, Mozilla, Amazon, Wikimedia Foundation, and Goethe Institute.

To inform the development of Arbiter, we need support to help us present at prestigious conferences, publish in academic journals, and adapt Arbiter to support bespoke use cases grassroots organizations require to be developed, e.g., in India, they first needed us to build an Android app, to automate data entry. Through this process, we’ve earned the trust of various communities worldwide, and hope to cater a meaningful technical system to support their work.

8. What value of grant are you seeking from the Data Empowerment Fund?
    a. **Answer**: $100,000

9. Describe what activities you would use the grant to undertake.
    a. **Answer**: Offset computation costs, support staff, and feature development to cater to education and partner requirements in the journalism and civil society spaces.

We will use 30,000 USD from the grant towards computational and infrastructure costs, including cloud infrastructure and technical equipment required to develop the Arbiter system. 50,000 USD will be used to support our staff (including salary, fringe benefits, tuition, or other direct compensation) to feature development to support professionals across journalism and civic society spaces. 10,000 USD will cover registration fees and travel expenses for conferences, seminars, and networking events that we attend to present our work while an additional 10,000 USD will cover miscellaneous expenses and indirect costs. These cost estimations are informed by actual numbers from the previous grants we have received via Google, Mozilla, Amazon, Wikimedia Foundation, the Center for Tech and Society, and the Goethe Institute.

!!! EXTRA QUESTION: Describe what outcomes you would use the grant to achieve.

We will use the funds to grow Arbiter into a community-based research tool to investigate coordinated networks, news, narratives, threat actors, and communities exposed to mis and disinformation on the social web. We aim to have three large community partners in journalism, civil society, and policy (e.g. govt. agencies that we have liaised with) that inform our development of this and whom we pilot this system with, ensuring we are guided by user requirements. We aim to fill the existing gap in the availability of intuitive and user-friendly trust and safety tools for professionals to investigate, monitor, and report on harmful online content with offline consequences. The grant will support compute costs, a project lead, research engineers, and staff members engaged in educational and research activities to build a sustainable community around Arbiter, including in the publication of two research papers investigating its utility and applications.

SAMPLE : Arbiter is a centralized platform that quickly queries multiple social media datasets across several platforms to help researchers, fact-checkers, journalists, and other civic society organizations form shared narratives of the digital ecosystem. This grant will help us strengthen existing community relationships and build new ones by supporting us in attending conferences, publishing in journals, and testing our new model with various actors. Online safety researchers across industries have to exhaust technical talent and time to analyze how content spreads on one platform, and this grant will help us give them an efficient tool to analyze viral content spreading across multiple platforms. This grant will help us share our knowledge with a multidisciplinary audience that aims to build a safer online world.

10. Describe what outputs you would use the grant to produce.
    a. **Answer**: Arbiter product, reports, testing with external partners

We aim to use the grant money to support the development of the Arbiter to facilitate the analysis of the spread and interaction of a piece of content across multiple social media platforms. We aim to publish reports independently and in collaboration with existing and potential new partners about how different social media is used at the time of elections and to reinforce a narrative in different countries. These reports aim to drive digital literacy programs to educate people about how to look for misinformation and how to scrutinize a piece of content they see on social media platforms for potential deepfakes or misinformation. We plan to test Arbiter with our external collaborators, such as IFCN-verified fact-checkers, to help them identify the coordinated spread of misleading content and facilitate the fact-checking for such information. This will reduce costs and the time it takes to analyze content populating across platforms. Another outcome of this is to use Arbiter for newsrooms to educate journalists about how coordinated networks work to spread fake news and help them understand what type of content or articles are being circulated and potentially identify bad actors and fake news.

11. Describe the ultimate impact you would use the grant to achieve.
    a. This is the longer term result of your work, likely to occur beyond the grant period.

Our vision is for Arbiter to be a user-friendly tool that will reduce the time and cost of analyzing content populating across platforms for a global audience of researchers, journalists, and other civic integrity professionals. Fraudulent news costs the global economy $78 billion a year, so the European Commission and U.S. Department of Defense Advanced Research Projects Agency are putting millions of dollars towards technical solutions to spot malicious and manipulated videos. As fraudulent content gets even more dangerous with new technology, we aim to position Arbiter as a leader in tracking generated content across social platforms. There is a clear need for identifying and pruning harmful content to mitigate economic and societal harms.

Arbiter aims to bridge the knowledge gap across platforms to help professionals stop deceitful content from spreading virally.

Due to Arbiter's intuitive interface, we plan to continuously refine it so professionals and researchers from across industries can participate in shaping the online landscape to reflect real-world values. To achieve this, we would appreciate the support and collaboration of the Data Empowerment Fund to fine-tune Arbiter for diverse use cases while helping spread awareness of the tool through academic journals, conferences, and civic society organizations

12. What will others be able to learn from your initiative? ** *We're particularly interested in supporting initiatives that: have significant public policy relevance; enable people to control how data is used to train AI models; and/or involve a novel legal, technological or participatory approach.*

Our centralized platform queries multiple social media datasets across platforms such as YouTube, Twitter, Bluesky, and Truth Social, jointly to obtain shared narratives. We have supported Policy Leads at Meta in South Asia to understand the complex political ecosystem in the Global South, and we are currently working on using our unique insights to inform similar teams across private platforms, academia, and public policy. This grant would help position Arbiter as a neutral tool that can enforce the DSA and Online Safety Act's regulations and transparency efforts to build cross-discipline efforts to track online risks and threats. Supporting us will help support an open-source, costless platform that can empower the public to audit the influence of bad actors, advertisers, and public figures across social channels. Arbiter's use to produce https://infolab.techglobalinstitute.com/how-facebook-has-become-a-political-battleground-in-bangladesh/ has informed discussions in Bengali Parliament, EU, and the US. It was featured in the national newspaper on the front page with a half-page feature.

13. Describe how the Data Empowerment Fund could support your work beyond providing access to funding. ** *E.g. by providing guidance on a particular challenge you're facing, or by connecting you to particular types of expertise.*

SimPPL has been able to develop culturally competent platforms because of our strong relationships with grassroots organizations, and we would appreciate introductions to other civil society partners to deploy hyper-local technological solutions.

We've partnered with local organizations such as Mongolia's only IFCN-verified fact-checker, the Queen of the Marma tribe representing the last speakers of their language in Bangladesh, and Aadhar Bahuddeshiya Sanstha who serves women and children in over 100 villages in rural Maharashtra, India, national newsrooms in Germany and the United States. With the trust of various community advocates, we've been able to build

multilingual technical solutions to help detect disinformation about women's health and manipulated narratives from state-run media.

Frankly, it has been a challenge to build Arbiter to query over half a billion social media posts in real time without funding, but we've still managed to make things work. As we work to scale Arbiter through community-based usage for independent research, we would immensely appreciate the Data Empowerment Fund's support to identify grassroots organizations that could use our technical solutions.

14. **Are you happy for us to share your application with our partners, the Patrick J. McGovern Foundation and Omidyar Network?**

Yes.

Caitlyn & Dhara's Notes

How can we tie Arbiter into this?

- "Coordination sharing" data insights across platforms.
- **Use Cases** – 1. How to prioritize escalations using Arbiter (which flagged videos have been populated across platforms the most),
- **Potential Clients** – Fact Checkers (Can use Arbiter to understand which videos), Journalists, News Rooms, Advertisers
- **Theme** – A tool to scale T&S research. This is not suggesting any sort of intervention but rather understanding how to prioritize which escalations to address.
    - We provide neutral third party that aids many T&S researchers and adjacent actors.
    - We aim to expand on existing policies by providing T&S researchers and 3rd parties by giving them a simple tool.