What the Observatory cannot do

Eight methodological constraints on the activities of the planned International Observatory on Information and Democracy.

Luke Thorburn
King's College London

June 2022

The planned International Observatory on Information and Democracy aspires to be the "IPCC of information and communication" [12]. To achieve a level of success comparable to that of the IPCC, the Observatory will need to synthesise data from social media platforms 'at scale' and will need to build a reputation of legitimacy across politically diverse jurisdictions. As a result of these twin necessities, I believe that the remit and methodologies of the Observatory are subject to at least eight constraints.

The constraints are not strictly binding—it may be possible to circumvent them in small studies, or to ignore them at the cost of a loss in the perceived legitimacy of the Observatory and an increased risk that its work becomes politicised. Nonetheless, in the interests of clarity and to provoke discussion I phrase them below using language that is absolute.

The first six constraints are positive (rather than normative). Their purpose is primarily to promote accuracy.

1. We cannot assume things are getting worse.

In its remit and research plan, the Observatory cannot presuppose that "democracy" or the "information environment" are deteriorating. While there are certainly some measures that suggest this is the case in some places [38, 27], it is not clear that these trends exist when considering the international community in aggregate. The situation is certainly not uniformly bad across jurisdictions [3]. If the research directions pursued by the Observatory are predicated on the assumption that things are getting worse—if they ask leading questions—then the Observatory may produce findings that are inaccurate. Moreover, it may exacerbate the very risks it is intended to address, becoming a self-fulfilling prophecy. Research has shown that exposure to concerns about the information environment can in some cases exacerbate polarisation, lead to reality apathy and reduce satisfaction with democracy [28, 34, 29, 24, 1, 33].

2. We cannot assess truth.

The Observatory cannot assess the truth of information circulating in the public domain. This is for both political and practical reasons. Politically, becoming the "arbiter of truth" would alienate those whose preferred information sources are labelled false and likely lead to distrust and politicisation of the Observatory's work. Practically, it is simply not possible to properly assess the truth of claims at scale, because it requires far more work to evaluate false claims than it does to generate them. A strong indication of the difficulties in assessing truth is that the main organ of the US intelligence community tasked with monitoring the quality of intelligence work—the Analytic Standards and Integrity (AIS) division of the Office of the Director of National Intelligence—does not include accuracy among the criteria it evaluates. Dr Barry Zulauf, former Chief of AIS, states:

> We in AIS have not evaluated products for accuracy for 6 or 7 years. In order to put judgements aside to test for later accuracy, they had to be clearly stated, falsifiable, and include a time-frame. THEN we had to devote personnel to doing the research through other reporting to assess accuracy, taking personnel away from the main line of work. AIS has steadily declined in personnel resources for the past 6 years, and has done NO such evaluation. Before, only a small proportion of the products we sample, which, in turn was a representative cross-section, not a statistically significant sample of all production, was evaluated for accuracy. [40]

If the US intelligence community finds it difficult to evaluate the truth of their products, when they have strong incentives to be accurate, then it is implausible that the Observatory could evaluate accuracy at scale. In particular, this means the Observatory cannot quantify the amount of false information in circulation.

3. We cannot observe intent.

The distinction between misinformation and disinformation hinges on the intent of the person communicating it [37]. However, intent is not directly observable and, as evidenced by various complex and lengthy legal cases [7], can require considerable amounts of time to adjudicate. Even if it were possible to identify false information, it is likely not feasible to partition that into the mutually exclusive categories of misinformation and disinformation.

4. We cannot verify identities.

Verifying identities online—either to ensure that online accounts are linked to real people, or to ensure that each real person has at most one account on a given platform—is difficult to do at scale. There are significant open research questions relating to the prevention of duplicate accounts, often framed in the language of "Sybil attacks" [5, 32]. While there are some proposed solutions [4, 30, 17, 16, 26, 6, 14], none are widely implemented or adopted.

The difficulties of verifying identities at scale mean that it is likely not feasible to accurately quantify the prevalence of bots or inauthentic behaviour on online platforms.

5. We cannot measure persuasion.

Many of the concerns about information and democracy are premised on the belief that communicated messages can change people's beliefs and actions in the real world. For example, there are concerns that medical misinformation disseminated on social media may cause people to take action that harms themselves or others [23], and concerns that foreign actors may influence the way people vote in domestic elections [31]. However, outside of narrow experiments, it is difficult both to measure actions and beliefs, and to attribute those to information consumed online. Perhaps the most salient example is digital advertising: many studies have failed to find evidence that digital advertising has any effect, on average, despite the vast sums companies spend on it [15, 13]. Similarly, money spent on political campaigns does not appear to be a deciding factor [25, 9].

These studies do not suggest that communicated information has no effect. Indeed all human beliefs and behaviour must be based on some form of information transfer. But they do indicate that the relationship between information consumption and subsequent behaviour is complex and difficult to observe. It is unlikely that the Observatory would be able to attribute beliefs or behaviour to the dissemination of online content, especially in real-world cases where we cannot observe the counterfactual world in which the content was not disseminated.

6. We cannot pretend social science is physics.

Unlike the IPCC, the Observatory's remit will predominantly fall within social science. This constrains the types of activities the Observatory can undertake. Unlike in climate science, where there is one overriding driver of harm (rising greenhouse gas emissions), the set of factors influencing the quality of the "information space" is much more diverse, and it is not yet clear which are the most important. In addition, it is not plausible that the Observatory could define forward-looking scenarios in the same way that the IPCC does because the trajectory of the information space is highly contingent on human behaviour in ways that are not able to be simply described.

The final two constraints are normative. Their purpose is primarily to promote perceived legitimacy in politically diverse jurisdictions.

7. We cannot stipulate where the free speech line should be.

Most jurisdictions have some notion of a right to freedom of expression [18], whilst also criminalising or allowing legal action against those who commit certain speech acts [39]. In online platforms, legal speech is subject to additional platform-specific content moderation policies that specify what types of speech are allowed on the platform [22, 10]. These policies are constantly evolving, are politically contested, and can differ by cultural context [20]. The Observatory cannot take a granular, detailed stance on where the line between permissible and impermissible speech falls without risking the politicisation of its work.

8. We cannot define democracy.

Currently there are 195 member countries of the IPCC [19]. In contrast, only 40 states have so far become signatories to the International Partnership on Information and Democracy (hereafter, simply the Partnership), with notable absences including China and the United States [11]. Together, the US and China have more than 20% of the world population, so it is important that an inclusive international institution such as the proposed Observatory can accommodate them both. Both China and the US have very different views of what it means to be a democracy, including how human rights should be prioritised [35, 8]. However, the Partnership requires signatories to declare *inter alia* that

> The global information and communication space, which is a shared public good of significant democratic value, must support the exercise of human rights, most notably the right to freedom of opinion and expression, including the freedom to seek, receive and impart information and ideas of all kinds, through any media of one's choice regardless of frontiers, in accordance with the International Covenant on Civil and Political Rights (Article 19). [11]

It seems unlikely that China would be willing to make such a declaration, given that it has not ratified the International Covenant on Civil and Political Rights [36]. However, both the US and China publicly claim that democracy and human rights are important [2, 21]. Thus, to the greatest extent possible without compromising its usefulness as an institution, the Observatory should be kept separate from the Partnership, and avoid stipulating or requiring particular interpretations of key terms such as "democracy" and "human rights" in its work. This will increase the likelihood that jurisdictions as diverse as China and the US will be willing to participate as member states, contribute researchers, and work towards common ground.

The eight constraints I describe above are not insignificant. That said, I highlight them in good faith, with the goal of steering the Observatory in directions I think are going to be most positively impactful and most likely to succeed. So what *can* the Observatory do?

Establishing a long-term, global database of descriptive data describing the information space and how it changes over time would be extremely valuable. In each country, what is the degree of affective polarisation? How do people spend their attention—on what information sources, on what platforms, and in the presence of what incentives? (This could be called *attention accounting*.) What are the levels of trust in the available information sources? To what extent do people's reflexive opinions agree with their more considered, deliberative judgements? What is the degree of reality apathy? How are all these variables trending, both within countries and globally?

A second line of work is both philosophical and experimental. How do we define and measure quality in the information space, in ways that remain impartial to the semantics of the information and particular political issues? Having defined quality, what types of mechanisms and incentive structures promote it? Synthesising such research could be done without going beyond the above constraints and would provide a roadmap for improvement.

About the Author

Luke Thorburn is a PhD candidate in the UKRI Centre for Doctoral Training in Safe and Trusted AI at King's College London. His research focuses on the use of social media recommender systems to mitigate political polarisation, and the design of bottom-up governance mechanisms for online information environments. He is currently collaborating with Aviv Ovadya on the design of "bridging-based" recommendation algorithms, and with Jonathan Stray (Center for Human-Compatible AI, UC Berkeley) and Pri Benghani (Tow Centre for Digital Journalism) on a series of articles about recommender systems and their societal impacts. He has also spent an unusual amount of time thinking about what language and metaphors are best used to describe the "information environment".

Before starting his PhD, Luke completed undergraduate and master's degrees in probability and stochastic processes at the University of Melbourne, and spent two years working to improve the information environment in Five Eyes intelligence communities at the Hunt Lab for Intelligence Research.

For a full CV, please see `lukethorburn.com/cv/`.

Conflicts of Interest

The author has no conflicts of interest.

Acknowledgements

Thank you to Zheng Hong See, Barry Zulauf, Tim van Gelder and Ashley Barnett for their assistance in preparing this submission, and to Aviv Ovadya for useful feedback. Any remaining errors are my own.

References

[1] Douglas J. Ahler. "Self-Fulfilling Misperceptions of Public Polarization". In: *The Journal of Politics* 76.3 (July 2014), pp. 607–620. ISSN: 0022-3816. DOI: 10.1017/S0022381614000085.

[2] Joe Biden. *Remarks By President Biden At The Summit For Democracy Opening Session*. https://www.whitehouse.gov/briefing-room/speeches-remarks/2021/12/09/remarks-by-president-biden-at-the-summit-for-democracy-opening-session/. Dec. 2021.

[3] Levi Boxell, Matthew Gentzkow, and Jesse M. Shapiro. *Cross-Country Trends in Affective Polarization*. Working Paper 26669. National Bureau of Economic Research, Jan. 2020. DOI: 10.3386/w26669.

[4] brightID. *Universal Proof of Uniqueness*. Tech. rep. brightID, Jan. 2022.

[5] Vitalik Buterin. *Hard Problems in Cryptocurrency: Five Years Later*. https://vitalik.ca/general/2019/11/22/progress.html. Nov. 2019.

[6] Qiang Cao et al. "Aiding the Detection of Fake Accounts in Large Scale Social Online Services". In: *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. 2012, pp. 197–210.

[7] Winnie Chan and A.P. Simester. "Four Functions of Mens Rea". In: *The Cambridge Law Journal* 70.2 (July 2011), pp. 381–396. ISSN: 0008-1973, 1469-2139. DOI: 10.1017/S0008197311000547.

[8] China. *The Right to Subsistence–The Foremost Human Right The Chinese People Long Fight For*. http://www.china.org.cn/e-white/7/7-I.htm. 1991.

[9] Alexander Coppock, Donald P. Green, and Ethan Porter. "Does Digital Advertising Affect Vote Choice? Evidence from a Randomized Field Experiment". In: *Research & Politics* 9.1 (Jan. 2022), p. 20531680221076901. ISSN: 2053-1680. DOI: 10.1177/20531680221076901.

[10] Evelyn Douek. "Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability". In: *Columbia Law Review* 121.3 (2021), pp. 759–834. ISSN: 0010-1958.

[11] Forum on Information & Democracy. *International Partnership for Information and Democracy*. https://informationdemocracy.org/principles/. 2020.

[12] Forum on Information and Democracy.
Prefiguration Group of the International Observatory on Information and Democracy (IOID).
https://informationdemocracy.org/working-groups/ioid/. 2021.

[13] Jesse Frederik and Maurits Martijn.
"The New Dot Com Bubble Is Here: It's Called Online Advertising".
In: *The Correspondent* (Nov. 2019).

[14] Mike Goldin. *Token-Curated Registries 1.0.* https://docs.google.com/document/d/1BWWC___-Kmso9b7yCI_R7ysoGFIT9D_sfjH3axQsmB6E/edit?usp=embed_facebook. 2017.

[15] Brett R. Gordon et al. "Inefficiencies in Digital Advertising Markets".
In: *Journal of Marketing* 85.1 (Jan. 2021), pp. 7–25. ISSN: 0022-2429.
DOI: 10.1177/0022242920913236.

[16] ID2020. *Digital Identity.* https://id2020.org/digital-identity. 2020.

[17] IDENA. *IDENA: Proof-of-Person Blockchain.* https://idena.io. 2022.

[18] Institute for Democracy and Electoral Assistance. *Freedom of Expression.*
https://www.idea.int/gsod-indices/democracy-indices?attr=%5B%22SC_02_02_a%22%5D.
2017.

[19] IPCC. *List of IPCC Member Countries.* Feb. 2019.

[20] Jialun Aaron Jiang et al. *A Trade-off-centered Framework of Content Moderation.* Tech. rep.
June 2022. DOI: 10.1145/3534929. arXiv: 2206.03450 [cs].

[21] Xe Jinping. *Speech by H.E. Xi Jinping President of the People's Republic of China at the Conference Marking the 50th Anniversary of the Restoration of the Lawful Seat of the People's Republic of China in the United Nations.*
https://www.fmprc.gov.cn/mfa_eng/zxxx_662805/202110/t20211025_9982254.html. Oct. 2021.

[22] Kate Klonick. *The New Governors: The People, Rules, and Processes Governing Online Speech.*
SSRN Scholarly Paper ID 2937985. Rochester, NY: Social Science Research Network, Mar. 2017.

[23] Heidi J. Larson. "The Biggest Pandemic Risk? Viral Misinformation".
In: *Nature* 562.7727 (Oct. 2018), pp. 309–309. DOI: 10.1038/d41586-018-07034-4.

[24] Matthew Levendusky and Neil Malhotra.
"Does Media Coverage of Partisan Polarization Affect Political Attitudes?"
In: *Political Communication* 33.2 (Apr. 2016), pp. 283–301. ISSN: 1058-4609.
DOI: 10.1080/10584609.2015.1038455.

[25] Steven D. Levitt. "Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U.S. House".
In: *Journal of Political Economy* 102.4 (Aug. 1994), pp. 777–798. ISSN: 0022-3808, 1537-534X.
DOI: 10.1086/261954.

[26] Litentry Technologies. *About.* https://www.litentry.com. 2022.

[27] Nahema Marchal and David Watson.
"The Rise of Partisan Affective Polarization in the British Public".
In: *SSRN Electronic Journal* (2019). ISSN: 1556-5068. DOI: 10.2139/ssrn.3477404.

[28] Erik C. Nisbet, Chloe Mortenson, and Qin Li. "The Presumed Influence of Election Misinformation on Others Reduces Our Own Satisfaction with Democracy".
In: *Harvard Kennedy School Misinformation Review* (Mar. 2021). DOI: 10.37016/mr-2020-59.

[29] Uwe Peters.
"How (Many) Descriptive Claims About Political Polarization Exacerbate Polarization".
In: *Journal of Social and Political Psychology* 9.1 (Feb. 2021), pp. 24–36. ISSN: 2195-3325.
DOI: 10.5964/jspp.5543.

[30] Proof of Humanity. *Proof Of Humanity.* https://www.proofofhumanity.id/. 2022.

[31] Scott Shane. "The Fake Americans Russia Created to Influence the Election".
In: *The New York Times* 7.09 (Sept. 2017).

[32] Divya Siddarth et al. "Who Watches the Watchmen? A Review of Subjective Approaches for Sybil-Resistance in Proof of Personhood Protocols". In: *Frontiers in Blockchain* 3 (2020).
ISSN: 2624-7852.

[33] Olga Stavrova, Daniel Ehlebracht, and Kathleen Vohs. *Victims, Perpetrators, or Both? The Vicious Cycle of Disrespect and Cynical Beliefs about Human Nature*. Jan. 2020. DOI: 10.31234/osf.io/thuq8.

[34] John Ternovski, Joshua Kalla, and Peter Aronow. "The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments". In: *Journal of Online Trust and Safety* 1.2 (Feb. 2022). DOI: 10.54501/jots.v1i2.28.

[35] The Economist. "China Says It Is More Democratic than America". In: *The Economist* (Dec. 2021). ISSN: 0013-0613.

[36] United Nations. *Status of Treaties: International Covenant on Civil and Political Rights*. https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-4&chapter=4&clang=_en. June 2022.

[37] Claire Wardle and Hossein Derakhshan. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Tech. rep. Council of Europe, Aug. 2018, p. 110.

[38] Steven W. Webster and Alan I. Abramowitz. "The Ideological Foundations of Affective Polarization in the U.S. Electorate". In: *American Politics Research* 45.4 (July 2017), pp. 621–647. ISSN: 1532-673X. DOI: 10.1177/1532673X17703132.

[39] Wikipedia Editors. "Freedom of Speech by Country". In: *Wikipedia* (May 2022).

[40] Barry Zulauf. *Evaluating Accuracy (Private Communication)*. June 2022.