SOCIETAL DIVIDES AS A TAXABLE NEGATIVE EXTERNALITY OF DIGITAL PLATFORMS

An exploration of the rationale for regulating algorithmically mediated platforms differently

TECH & HUMANITY
Powered by ASHOKA

BUILD UP

This paper was written by Helena Puig of Build Up. It is part of a collaboration under Ashoka's Tech & Humanity initiative, a global network of leading social entrepreneurs committed to ensuring tech works for the good of people and planet. This community is concerned about the societal and environmental harms of the data economy and is building innovative frameworks and tools to mitigate these harms.

Table of Contents

Executive Summary

Online polarization is a negative externality that results from the incentives of the attention economy.

This paper argues that many societal divides today are negatively impacted by patterns of digital content consumption and interaction driven by algorithmically mediated platforms. However, these digital platforms were not set up in order to distribute divisive content or create polarizing interactions. Rather, digital platforms produce these societal divides as a by-product of a profit maximizing strategy that governs the content and interactions they promote, and results in polarization online that translates into societal divides offline. As such, online polarization should be seen as a taxable negative externality of surveillance capitalism.

There is widespread agreement that algorithmically mediated platforms are degrading the digital space and the social fabric of societies. Current policy approaches to address the negative societal impact of digital platforms, whether through content moderation or changes to algorithm design, require the collaboration of companies. These approaches face a fundamental challenge: it is in the financial interest of digital platform companies to spread content that is engaging, divisive content is engaging, and platforms are not incentivized to increase their accuracy in detecting or reducing the distribution of polarizing content. Digital platform companies will never invest as much on content moderation as would be required, and are unlikely to make changes to algorithms that limit engagement.

We need a different policy approach, one that addresses the financial incentives that lead platform companies to build algorithms whose feedback loops and unintended consequences result in polarization.

We argue that online polarization is a negative externality of algorithmically mediated platforms that results from the incentives of surveillance capitalism and the attention economy. If we understand online polarization as a negative externality, then there is a policy argument in favor of creating a financial disincentive to its production.

This could be done by taxing polarization as a negative externality, and we explore five possible avenues for taxation: one based on a direct measure of polarization, four based on proxies that affect the production of online polarization. We reject two of these proxies as not viable or inappropriate.

Taxing polarization as a negative externality would create a financial disincentive to online polarization.

Recognising that much work remains to be done to make these taxation options viable, we propose two concrete next steps:

1. Piloting approaches to measure the "polarization footprint" of digital platforms, and the viability of proposing a taxation regime based on this measure;
2. Researching the connection between personalization / content targeting based on certain kinds of personal data and polarization outcomes, in order to understand whether introducing a non-linear tax on large databases might reduce polarization.

Overall, this paper sets out options for credible policy pathways to make surveillance capitalism more expensive in order to reduce the impact of digital platforms on societal divides. We invite challenge to the assumptions that lead us to argue for taxation, and discussion on the viability of concrete tax regimes proposed. This is the start of a longer and much needed debate.

01 The negative societal impact of algorithmically mediated platforms

"Like any rational entity, the algorithm learns how to modify the state of its environment — in this case, the user's mind — in order to maximise its own reward."
— Stuart Russell

1.1 Algorithms influence preferences

All digital platforms are mediated by algorithms: users browse content on the platform, and the order in which that content is presented is determined by a ranking algorithm. Whether a user is browsing social media, news, a shopping site, search engine results or streaming sites, an algorithm ranks the content they are viewing according to a set of rules and signals.

Design decisions determine what rules and signals an algorithm uses to rank content. Platforms use a variety of signals (sometimes hundreds or thousands), including chronological order, credibility scores, and more. The design of an algorithm determines the weight or importance a signal is given in ranking content, and most algorithms are designed to give a lot of weight to how likely an individual user is to interact with content, that is they rank primarily to maximize engagement. If the rule is to maximize engagement, then algorithms will look for signals that correlate with high engagement, and use those to recommend content. These signals may include past interactions of the user, known demographic characteristics of the user, and what content other similar users interact with. Critically, algorithms learn through feedback loops: an algorithm recommends based on user data, users react to what the algorithm presents, and that generates more data – all in the pursuit of engagement maximization.

Over time, the recommendations of algorithms influence our preferences.

There is growing evidence that, over time, the content consumption recommendations of algorithms influence our preferences [1]. In some ways, there is nothing wrong with engaging content that changes preferences: so does reading a good book. The difference is about agency and intent. First, the feedback loops that feed algorithms are largely hidden from most users, so that the way content is presented takes away agency from the user, amounting to something close to manipulation.

[1] https://www.tandfonline.com/doi/full/10.1080/23808985.2021.1976070 and https://dl.acm.org/doi/10.1145/3460231.3478849

Second, preference change is unintended, because the algorithm was not built with the intent to change our preferences, but rather to maximize another goal (engagement). In other words, this is not a deliberate use of automation or data-driven algorithms for marketing or propaganda intended to change preferences, but rather an unintended way that an automated system affects preferences in its algorithmic pursuit of maximization. [2]

There is an ethical question about whether manipulative, unintended preference changes respect individual human dignity and agency – that is not a question we address here. Instead, we are interested in whether the changes to preferences resulting from interacting with algorithmically mediated platforms, in the aggregate, have a negative impact on conflict dynamics in society. That is, we are not concerned with the ethics of an algorithm deepening a user's preference for complicated baking and affiliation to baking groups by presenting more videos on this subject, but we are concerned with the societal impact of an algorithm deepening a user's preference for armed violence and affiliation to armed groups by presenting more videos on this subject.

1.2 Algorithmically influenced preferences & societal divides

The negative societal impact of consuming algorithmically-curated content is at this point supported by both academic literature / experiments and by real world examples. Concretely, many academic studies show a positive correlation between digital media use and polarization [3], and there is ample evidence that polarization harms democracy [4], correlates with dehumanization [5], and leads to violent conflict [6]. Although many of these studies fall short of demonstrating causal evidence of this link [7], there is widespread agreement that the effect of digital media consumption on polarization needs to be taken seriously for policy purposes.

This effect is clearly connected to the amplification of hate and manipulation on digital platforms. Many people and groups take to digital media to share hateful or manipulative content that deliberately aims to pit one group against another. These direct attempts at polarization are problematic, but no different to other forms of divisive, anti-democratic or violent propaganda. The issue specific to digital platforms is that they create a perverse incentive to produce divisive content because this content is more likely to go viral [8]. Content expressing hate towards out-groups [9] or political opponents [10] and content that expresses moral outrage [11] are all substantially more likely to engage users – and to eventually lead to hate and violence. Furthermore, actors wanting to spread divisive content have an incentive to coordinate because many algorithms reward networked sharing [12], thus rewarding manipulative tactics.

[2] https://medium.com/understanding-recommenders/is-optimizing-for-engagement-changing-us-9d0ddfb0c65e
[3] https://osf.io/preprints/socarxiv/p3z9v/
[4] https://journals.sagepub.com/doi/10.1177/0002716218818782
[5] https://link.springer.com/article/10.1007/s11109-019-09559-4
[6] https://journals.sagepub.com/doi/10.1177/0022343307087168
[7] https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57
[8] https://www.brookings.edu/techstream/how-social-media-platforms-can-reduce-polarization/
[9] https://osf.io/rhmb9/
[10] https://www.pnas.org/doi/10.1073/pnas.2024292118
[11] https://www.pnas.org/doi/10.1073/pnas.2024292118
[12] https://www.adl.org/resources/blog/reality-how-harassment-spreads-twitter

In recent years, Russian trolls have attempted to run a number of manipulation campaigns over social media, with the intent of sowing hate and stoking division in contexts as varied as the women's march of 2017 [13] and the Ukraine war [14], although evidence of the effectiveness of these tactics is not clear [15]. Not all actors benefiting from the amplification of hateful and manipulative content are ideologically motivated: some (perhaps most) are just out to make money. In 2019, the Global Disinformation Index said disinformation websites earned $250 million in ad revenue. By 2021, that amount had risen to $2.6 billion [16]. One example of how this impacts the spread of divisive content: of 30 German-language sites the EU DisinfoLab identified as consistent sources of false content, more than 30% earn money with Google ads. Many of these sites mix far right narratives with COVID19 disinformation. Another example: activists in Myanmar report that there has been a rise in financially-motivated actors who are creating disinformation to boost pro-military narratives in order to capitalize on YouTube's monetization options. These actors are primarily based in Cambodia and Vietnam, and some are also working to produce disinformation on Ukraine.

These examples illustrate the perverse and insidious way in which algorithms amplify the intent of divisive actors, and of financially-motivated actors. The amplification of hate and manipulation might not in itself shift user preferences – arguably, it could be that user preferences are driving the incentives that result in the algorithm amplifying this type of content, as was shown in a study of YouTube recommendations [17]. It's the presence of feedback loops that leads to a shift in user preferences, specifically through the effects of filter bubbles / echo chambers and of partisan sorting. These effects are explained below.

Filter bubbles

Filter bubbles happen when feedback loops reinforce ideological preferences: a user views or chooses ideologically aligned content, then an algorithm infers that the user has a preference for this type of content and increases the fraction of this type of content it presents to the user, which then strengthens the user's preference for this type of content. This feedback loop has been shown in simulations [18]. In effect, the algorithm nudges users towards beliefs that are increasingly extreme, largely because radical beliefs are more likely to engage and most algorithms are designed to optimize for engagement, so it is in the interest of the algorithm to influence users so they become more radical.

Several commentators have written about the "far right rabbit hole" that explains stories such as that of Caleb Cain, a liberal college dropout whose consumption of YouTube videos led him to align with alt-right views [19]. However, in experimental research, the operation of filter bubbles on YouTube has been shown to marginally nudge users towards increasingly narrow ideological content in the USA, but not towards "rabbit holes" [20]. It may be the case that we see larger effects on other platforms that have been under less scrutiny, but systemic evidence is not available.

[13] https://www.nytimes.com/2022/09/18/us/womens-march-russia-trump.html
[14] https://www.vice.com/en/article/wxdb5z/redfish-media-russia-propaganda-misinformation
[15] https://www.nature.com/articles/s41467-022-35576-9
[16] See: https://www.disinformationindex.org/
[17] https://www.pnas.org/doi/10.1073/pnas.2101967118
[18] https://dl.acm.org/doi/10.1145/3306618.3314288
[19] https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html
[20] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4114905

Partisan sorting

Partisan sorting explains the phenomenon of homophily on social media: people seek out others like themselves, algorithms pick up on this preference and reinforce it. Furthermore, people become more similar to the people they interact with, creating a reinforcing feedback loop. On social media, partisan sorting allows for more niche or marginal groups to find each other because of the long-range connections across society that it enables. This can be a social positive in some ways – for example, LGBTQ+ people may find support and community online in ways that were not possible offline – but the society-wide impact can result in greater polarization, especially in already divided societies. If people interact mostly offline, the majority of their contacts will be local, and each community will end up with its own local ideological flavor, resulting in a wide range of local community ideologies. When people can use social media to interact across a society, partisan sorting can lead to one-dimensional, bipolar structure of beliefs, which results in polarization.

The impact of digital media partisan sorting on affective polarization has been shown in simulations. [21] Over time, partisan sorting allows actors from peripheral groups to gain influence, boosting in-group solidarity and out-group animosity. [22] Partisan sorting in part explains how social media content has led to violence against ethnic and identity groups in conflict contexts from Myanmar (against Rohingya people) to Ethiopia (in Tigray).

1.3 Conclusion

"We care which means the algorithm used to solve the problem, but we only told it about the ends, so it didn't know not to cheat." — Krueger et al.

There is widespread agreement that algorithmically mediated platforms are degrading the digital space and the social fabric of societies. The scarcity of systematic, replicable, causal evidence [23] is outweighed by the abundance of experimental research showing correlation and real-world experiences demonstrating impact. The immediate problem presented by algorithmically mediated platforms is the generation of societal divides through patterns of content consumption and interaction online. These patterns largely result from changes to user preferences that are the unintended consequences of the incentives that are encoded in engagement-maximizing algorithms.

[21] https://www.pnas.org/doi/10.1073/pnas.2207159119
[22] https://journals.sagepub.com/doi/10.1177/14614448221137324
[23] https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57

02 Current policy approaches

2.1 Content moderation

Most digital platforms self-regulate content moderation, using their terms of service to outline a set of categories of prohibited content, and guidelines on content removal and account restrictions based on these categories. Although categories vary by platform, most prohibit hate speech, incitement to violence, synthetic or manipulated media, and any post discussing the exchange of regulated goods. Many platforms have been criticized both for lack of transparency in the application of these policies and for dedicating insufficient resources, especially in non-English speaking contexts. The failings of self-regulated content moderation have been evident in a number of high-profile cases, for example Facebook's well-documented failure to heed multiple warnings about hate speech and incitement to violence targeting the Rohingya in Myanmar [24]. In their report on the role of Facebook in the Rohingya crisis [25], Amnesty International argue that self-regulated content moderation is not a sustainable solution because digital platforms have no financial incentive to rein in hate. The balance on this calculus may be changing - increasingly hate speech drives away advertisers, generates criticisms and invites regulatory scrutiny - but the fact remains that overall, and especially in non-English regions, self-regulated content moderation is chronically under-resourced.

Some governments have attempted to introduce external regulation of content moderation on digital platforms. The German Network Enforcement Act requires all digital platforms to remove content that is "manifestly unlawful" under German law. Many countries have followed the example of Germany, but have too often used what was intended as a law to protect from hate and harm to censor and curtail public debate. [26] In most countries, laws that protect free speech make external regulation of content moderation legally complex. In the USA for example, Section 230 makes external regulation of content moderation very difficult by shielding intermediaries like Facebook or YouTube from liability for user-generated content, and providing additional protection to ensure that intermediary moderation doesn't invite new liability. Recent debates in the USA have indicated that digital platforms could be considered not to be a public square, but rather more similar to edited media, because algorithms are in effect the "editors" that present certain content to users. In the USA and elsewhere, such discussions could eventually lead to liability, but this will be a long road.

Moderation will never affect more than a small amount of content that explicitly violates policies.

[24] https://rh.myanmarinternet.info/
[25] https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/
[26] https://www.eff.org/deeplinks/2020/07/turkeys-new-internet-law-worst-version-germanys-netzdg-yet

Regardless of whether it is self-regulated or externally regulated, and how well regulations are implemented, moderation will never affect more than a small amount of content that explicitly violates platform policies. Attempts to expand content moderation policies to include additional categories have led to confusion (for example, in response to the Ukraine war [27]) or unfair over-enforcement (for example, in Israel & Palestine [28]). Furthermore, there is some evidence that automated methods to moderate content are unavoidably error-prone. [29] Even with the best content moderation policies, it may be impossible to remove all content that may drive polarizing or violent outcomes, both because of freedom of expression concerns and the practical realities of moderation at scale.

2.2 Algorithm re-design

Some policy discussions in recent years accept that it may be impossible to catch all (or even most) hateful and manipulative content, and instead turn their attention to policies that might dampen amplification by regulating algorithm design. These discussions have become more mainstream as revelations from ex-employees at digital platforms show that algorithms are purposefully built to optimize for provocative and polarizing interactions in order to maximize user interaction and attention. [30] This is not to say that platforms want to be sensational as a business model, but rather that sensationalism has emerged as a side effect of trying to be useful by ranking content in ways that help platform users navigate vast amounts of content, because engagement is an ambiguous signal: it's not easy to separate "good" engagement from "bad" engagement. [31]

One way to tackle this issue is to make it difficult to create the feedback loops that feed algorithms: data minimization policy proposals suggest companies should collect only the data necessary to provide their product or service. [32] Data minimization is primarily a human rights issue: it protects the privacy of individuals and prevents the most egregious misuse of data (e.g. for surveillance or deliberate manipulation). Existing regulations in a number of countries, including the EU's General Data Protection Regulation (GDPR), already include a data minimization standard ("adequate, relevant and not excessive") with additional purpose limitations for data collected ("fair and legitimate purposes").

Data minimization does not stop digital platform companies from building algorithms if we accept that humans need a machine-supported way to sort through the amount of content available on digital platforms (and that these algorithms are necessary for this service provision). In fact, companies can still build algorithms applying data minimization standards [33], and there is no research (to my knowledge) on whether algorithms built along data minimization best practices would dampen the harmful feedback loops that lead to polarization outcomes. Furthermore, these standards are already challenging to enforce when it comes to how they impact privacy, and are even more so if we try to consider social harms [34]: some of the algorithm design choices will be unaffected by them.

[27] https://www.bsr.org/en/reports/meta-human-rights-israel-palestine
[28] https://www.washingtonpost.com/technology/2022/03/25/social-media-ukraine-rules-war-policy/
[29] https://cyber.fsi.stanford.edu/news/automated-content-moderation-primer
[30] https://www.theguardian.com/technology/2021/oct/24/frances-haugen-i-never-wanted-to-be-a-whistleblower-but-lives-were-in-danger
[31] https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851
[32] https://www.accessnow.org/data-minimization-guide/
[33] https://www.privitar.com/blog/better-machine-learning-through-data-minimization
[34] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3191231

Some policy proposals push beyond data minimization to suggest companies should not use any historical human behavioral data to train algorithms, in an attempt to address some of the more harmful unintended consequences of engagement optimization. For example, forgetful advertising proposes platforms could target content using information that can be gleaned from a single interaction between a user and a website, but could not store human behavioral data in order to remember any previous interactions to inform its targeting. [35]

Other policy proposals make direct suggestions about what algorithms should optimize for (rather than what data they should utilize). Proponents of bridging algorithms suggest building a recommender that rewards positive interactions across diverse audiences, including around divisive topics [36]. Another option is to downrank "borderline" content and uprank authoritative content – something a number of the larger social media platforms have implemented to some degree [37]. The challenge with these design solutions is that agreeing on the principles that guide them (what are positive interactions, what content is borderline, etc) can be difficult and contested – although arguably less so than content moderation decision that rely on content removal. One option that circumvents this challenge is to introduce some uncertainty into the predictions of an algorithm, some randomness and noise in the recommendations.

Even if policy makers could agree on key guidelines for algorithm design, their implementation requires the collaboration of digital platform companies. The machine learning processes used by digital platform companies are largely opaque. Some organizations, like the Integrity Institute [38], have been calling for companies to disclose details of the role that algorithms play in the distribution of harmful content, in an attempt to start a discussion about algorithm re-design. Work to conceptualize how algorithms could be designed to minimize harm is gaining traction [39], but at present, significant collaboration from platform companies on algorithm re-design is not forthcoming.

2.3 Conclusion

Current policy approaches to address the negative societal impact of digital platforms, whether through content moderation or changes to algorithm design, require the collaboration of companies. This is fundamentally flawed: it is in the financial interest of digital platform companies to spread content that is engaging, divisive content is engaging, and platforms are not incentivized to increase their accuracy in detecting and reducing the distribution of polarizing content.

It is in the financial interest of digital platform companies to spread content that is engaging, and divisive content is engaging.

35] https://law.yale.edu/isp/publications/digital-public-sphere/healthy-digital-public-sphere/forgetful-advertising-imagining-more-responsible-digital-ad-system
[36] https://www.belfercenter.org/publication/bridging-based-ranking
[37] https://journals.sagepub.com/doi/full/10.1177/20563051221117552
[38] https://integrityinstitute.org/
[39] See for example: https://knightcolumbia.org/blog/call-for-participation-optimizing-for-what-algorithmic-amplification-and-society

Digital platform companies will never invest as much on content moderation as would be required, and are unlikely to make changes to algorithms that limit engagement. We need a different policy approach, one that addresses the financial incentives to build algorithms that maximize engagement (and whose feedback loops and unintended consequences result in polarization). In order to argue this policy approach, we need to better understand the financial incentives of digital platform companies: surveillance capitalism and the attention economy.

03 An alternative framing: online polarization as a negative externality

"Algorithms are opinions embedded in code." – Catherine O'Neill

3.1 The surveillance capitalism business model

Most digital platforms have a surveillance capitalism business model. [40] Their monetization strategy is to provide a free service (a search engine, a social network, etc.) and then collect large amounts of data from the users of this service. This data can then be used for commercial purposes, broadly in two areas. First, companies can use data to build models or profiles that help predict the actions of users, and can therefore be used to target ads – ad targeting is an important monetization strategy for Google, for example. Second, companies can re-sell the data to others who can use it to build their own models or profiles and target content.

Surveillance capitalist models have one thing in common with any part of the attention economy: they need high levels of engagement with their content. To get data, digital platforms need users to engage over and over again; and to drive this engagement, they use data they have previously collected to recommend content that their profiles / models predict will result in greater engagement. For digital platforms, algorithms are a way to solve the core problem of the attention economy: if human attention is a scarce resource that the platform needs to garner in order to collect data that can be monetized, then algorithms should be programmed to maximize attention, as expressed in engagement. In section 1.2, we explained that there is evidence that polarizing content is more engaging, which is why making polarizing content go viral serves the attention extraction model of surveillance capitalism.

[40] https://www.publicaffairsbooks.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395700/

Unchecked surveillance capitalism is a danger to individuals and society [41]. As discussed in section 2.2, content moderation policies and algorithmic design could address the spread of divisive content and polarizing interactions. The problem is that digital platforms do not have sufficient financial incentives to implement these solutions – reducing polarization is expensive, leaving it unchecked does not carry a sufficient financial penalty.

3.2 Online polarization as a negative externality

Digital platforms produce societal divides as a by-product of a profit maximizing strategy.

Digital platforms were not set up in order to distribute divisive content or create polarizing interactions. Rather, digital platforms produce these societal divides as a by-product of a profit maximizing strategy: that is, polarization is a negative externality of their business model. Shoshana Zuboff explains that where industrial capitalism exploits nature, surveillance capitalism exploits human nature. The concept of negative externalities is commonly used to explain the impact of industrial capitalism on the environment and on public health. Simply put, an externality is an uncompensated effect of production or consumption that affects society outside of the market mechanism. Where there is a negative externality, the private (or company) costs of production are lower than the social costs of production. In other words, there is no disincentive to produce this negative externality because it is not priced in the business model.

Applying this framework to digital platforms explains why the production of polarization goes unchecked. Producing hateful or polarizing content is not necessarily profitable to platforms: polarizing content would have to be a very large fraction of all content to change engagement metrics in a business-relevant way, and its presence drives away advertisers, damages reputation and invites regulation. The challenge is rather that polarization is a difficult-to-control side effect of optimizing for the type of engagement that platforms do want to maximise for. Pollution is not profitable to industrial capitalism – it's the manufacturing that is profitable, but reducing pollution is expensive. Similarly, polarization is not profitable to surveillance capitalism – it's the engagement that is profitable, but reducing polarization is expensive.

A critical aspect of negative externalities is that they are collective problems across the market economy – and therefore require societal policy changes rather than increases to individual rights. This is why carbon taxes are favored over individual carbon credits. Individual ownership of our carbon footprint cannot work if we do not have meaningful choices to change our carbon production, because these choices are determined by the profit-maximizing incentives of producers. Furthermore, we understand that trace amounts

[41] https://www.amnesty.org/en/latest/press-release/2019/11/google-facebook-surveillance-privacy/

of carbon dioxide are barely detectable and cause no environmental harm, but in the aggregate large amounts of greenhouse gasses cause fundamental damage to the environment. Carbon taxes, on the other hand, set a price that each polluting company must pay per tonne of greenhouse gas they emit, thus introducing incentives to reduce emissions by pricing in their social cost. [42]

If we can understand that polarization is a negative externality of surveillance capitalism, much like carbon emissions are negative externalities of industrial capitalism, then a polarization tax is akin to carbon taxes.

The corollary for surveillance capitalism is that some policy makers might propose individual data ownership policies could address the harms of data (mis)use: users could control how their personal data is used, and decide whether to share or sell it. [43] However, like with carbon, the creation and consumption of data reflects how power is distributed in a society – in most situations, you cannot refuse to consent to your data being collected because you don't have a meaningful choice of alternative services. [44] Furthermore, as with carbon, an incremental erosion of trust driven by polarizing content is hard to notice and does little harm to each individual, but a massive change in the nature of human communication causes fundamental damage to the social fabric.

Data "pollutes" the social good when the collection and use of human behavioral data affects society as a whole, beyond those whose data is collected, and separate from the harm to their individual privacy or agency. [45] If we can understand that polarization is a negative externality of surveillance capitalism, much like carbon emissions are negative externalities of industrial capitalism, then a polarization tax is akin to carbon taxes.

The other widespread policy option to address environmental harm is cap-and-trade. Where carbon taxes set a price on the environmental harm of carbon and allow the market to adjust to a quantity of carbon emissions that prices in that harm; cap-and-trade sets a quantity of carbon emissions and allows the market to determine the price of the environmental harm of carbon. Cap-and-trade are considered the more desirable taxation policy in the EU. [46] However, cap-and-trade policies work well for a narrowly defined set of producers who receive complex but well-specified initial allowances of pollutants. This is much harder to do for digital platforms, which are many and varied, and for polarization or data allowances, which run into greater complexity given all the possible types. [47]

[42] A further exploration of the history and current legislation for carbon taxes is offered in "Accounting for the Environmental Impact of Data Processes", a paper written by Eticas under in this Next Now report series.
[43] https://www.ft.com/content/a00ecf9e-2d03-11e8-a34a-7e7563b0b0f4
[44] https://www.technologyreview.com/2018/12/14/138615/its-time-for-a-bill-of-data-rights/
[45] Omri Ben-Shahar calls this concept "data pollution", see https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3191231
[46] See https://climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets_en
[47] See here for a further exploration of this argument: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3191231

3.3 Taxing online polarization as a negative externality

Understanding that online polarization is both expensive to avoid for platform companies and a negative externality to the social fabric sets the scene for arguing in favor of taxes that address it as a policy response to elicit behavior change.

3.3a Taxing the polarization footprint

The most direct way to introduce a tax that addresses the negative societal impact of algorithmically mediated platforms would be to agree on a measure of online polarization, and tax companies according to how much of this negative externality they produce. This might superficially seem similar to external content moderation policies that impose fines to penalize platforms that fail to adequately moderate certain categories of content. The main difference with these content moderation policies is that a taxation policy regulates viral, polarizing content and interactions in the aggregate. In other words, the policy and its enforcement mechanisms need to be set up not to identify individual violations for categories of content that must be removed but to measure a pattern of content consumption, distribution and interaction across the platform.

We could agree on a measure of online polarization - a "polarization footprint" - and tax companies according to how much of this negative externality they produce.

At a minimum, this would require developing a framework for a mathematical model that can score patterns of content consumption and interaction across platforms and contexts. This measure would have to be based on identifying a set of divisive patterns, and then determining their incidence across the platform. Although there have been many studies on the impact of algorithmically mediated platforms on users, [48] these studies mostly look at how overall changes in consuming digital media impact affective or political polarization, and are mostly observational. They do not disaggregate by specific patterns of content consumption or on-platform interaction, and do not measure polarization as an outcome.

In the peacebuilding and mediation field, Build Up has begun to categorize digital conflict drivers [49] and the archetypes of polarization on social media. [50] Based on these frameworks, derived from peacebuilding and mediation practice, a model to score online polarization would need to cover:

[48] https://gpai.ai/projects/responsible-ai/social-media-governance/responsible-ai-for-social-media-governance.pdf
[49] https://howtobuildup.org/wp-content/uploads/2022/10/2022_Book_FundamentalChallengesToGlobalP_Chapter.pdf
[50] https://howtobuildup.medium.com/archetypes-of-polarization-on-social-media-d56d4374fb25

1. Amplification of deliberate polarizing tactics, measuring the virality [51] of content that is:
    a. Coordinated harassment of individuals / institutions
    b. Coordinated harassment of identity groups
    c. Incitement to hate / violence
    d. Inflation of certain positions (consensus or division)
    e. Disinformation about key facts relevant to social or political issues

2. Amplification of network-wide polarizing interactions, measuring the prevalence of interactions that:

    a. Polarize attitudes towards groups
    b. Polarize group affiliation
    c. Break down intergroup interactions
    d. Polarize interest-based or group narratives
    e. Reinforce trust-degrading norms

An immediate challenge to a taxation policy based on this score of online polarization would be to define the patterns to be monitored under 1 and 2. This might require establishing a causal link between the prevalence of these patterns and other (societal or individual) measures of polarization, which is very hard to do in any robust way unless you can systematically manipulate algorithms to check for before / after effects. [52] Even if this could be overcome, there are significant technical challenges to building and implementing any kind of automated, systematic monitoring of this score. Critically, platform companies would have to be required to provide API access to their data and comprehensive transparency of their algorithms. The Integrity Institute has outlined a set of transparency requirements [53] that would allow public tracking of harms on social media platforms, and that could be a starting point to ensure the viability of measuring the "polarization footprint" of digital platforms externally, and taxing this negative externality much like carbon taxes are determined by a carbon footprint.

Although taxing a measure of online polarization would be the most direct way to apply the logic of taxing a negative externality to address the societal harm of surveillance capitalism, it may not be viable given the definitional and measurement challenges. In the remainder of this section we look into taxable proxies for online polarization, defined as aspects of the surveillance capitalism business model that are taxable and impact the production of online polarization.

3.3b Taxing databases

Digital platforms collect masses of human behavioral data that is then used to power recommender algorithms. An argument can be made that a tax on databases could be a way

[51] Further work would be needed to define how "virality" or "amplification" is defined and measured, especially against user engagement and sharing behaviors.
[52] https://gpai.ai/projects/responsible-ai/social-media-governance/responsible-ai-for-social-media-governance.pdf
[53] https://integrityinstitute.org/s/Ranking-and-Design-Transparency-EXTERNAL.pdf

to price the societal cost of using this data in a way that we know is linked to the production of a negative externality (online polarization). In effect, this would be similar to taxing the "data footprint" of digital platforms rather than their "polarization footprint" by imposing a tax on very large databases (or introducing a tax incentive to disincentive very large data hoarding). The tax rate could potentially be non-linear, imposing a higher rate the more data is collected (so that Facebook pays more for each additional unit of data than a small social app). The tax rate could be higher for data collected with no immediate use to the collector (i.e. collected for future brokerage purposes). [54]

There is one important counterargument to this policy: data use has both negative and positive externalities. [55] Concretely for digital platforms, not all uses of data to recommend content result in societal divides. In fact, there is evidence that personalization creates a lot of value for users. [56] In other words, personalized recommendations are more likely to result in polarization outcomes, but they also result in many other positive outcomes. This counterargument is somewhat mitigated by considering that the positive externalities of data use are likely to be at least to a certain degree priced in by digital platforms. Furthermore, the tax rate could apply only to specific categories of personal data or data use, in order to account for the trade off between genuine user value from the data and polarization outcomes. Implementing this policy would necessitate researching the connection between personalization / content targeting based on certain kinds of personal data and polarization outcomes, in order to understand whether introducing a non-linear tax on large databases might reduce polarization.

If implemented at a high enough rate, it is possible that digital platforms would choose to pass on the tax to users – effectively ending free accounts and online services. As Omri Ben-Shahar argues, people are currently paid for their data with services, and if we think that producing data is leading to social harm, then that is akin to being paid to pollute. [57] A tax on data would be radical, and would radically change the nature of our digital ecosystem.

A tax on data would be radical, and would radically change the nature of our digital ecosystem.

3.3c Taxing data centers

Data centers are the basic physical infrastructure that houses the components needed for cloud computing. As such, data centers are key infrastructure for surveillance capitalism, and taxing them could increase the cost of data use, and therefore impact the production of online polarization down the line. [58]

[54] It is worth noting that the same challenges to data classification (as being of immediate use to the collector) that were mentioned in the critique of content moderation / GDPR above would apply to establishing this differential tax rate.
[55] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3191231
[56] https://www.law.upenn.edu/live/blogs/78-quantifying-the-user-value-of-social-media-dat
[57] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3191231
[58] Data centers result in negative environmental and socio-economic externalities by their existence, and that alone calls for taxation. See "Data Centres as Taxable Property" in this Next Now report series for an exploration of these negative externalities.

There are two possible ways to tax data centers - applying a property tax or applying an environmental tax - which are explored in detail in "Data Centres as Taxable Property", a paper written by Eticas in this Next Now report series. The overarching conclusion of this paper is that taxes specific to data centers are unlikely to be viable because (i) property taxes apply at municipality level and implementing taxes on data centers could potentially drive competition among jurisdictions, and (ii) not all jurisdictions consider data centers as physical property, which could pose a major challenge in terms of applying property taxation. Even if these challenges could be overcome, at best this tax would incentivize data minimization, but it would not alter data use. Its impact on online polarization is unclear.

3.3d Taxing the carbon footprint of data processes

Data processes refer to any transfer or manipulation of data. Data processes are used in two key ways in surveillance capitalism: tracking users to collect their data (using cookies) and training the machine learning models that power algorithms. Taxing these data processes would change the financial incentives to train algorithms.

"Accounting for the Environmental Impact of Data Processes", a paper written by Eticas in this Next Now report series, states that data processing uses a lot of energy, and therefore results in large carbon footprints. Thus, it can be taxed by measuring these carbon footprints, and then ensuring data processes are included in carbon tax regimes. [59] The paper outlines the best practices and challenges in measuring different data processes. Of the data processes relevant to surveillance capitalism, browser cookies have a negligible carbon footprint, but training of machine learning models has a significant carbon footprint. It is worth noting that digital platforms are financially well-off and could invest in mitigation strategies - as many already do - to lower the environmental impact of data processes without reducing the amount of data processed.

Assuming that a carbon tax applied to data processes did create a financial incentive to apply data minimization and to use less data to train algorithms, the impact on online polarization is unclear. Digital platforms will still have to use algorithms to rank content on their platforms, and algorithms that rank for more complex measures that take into account polarization outcomes might use more computing power / data processes than current engagement-maximizing algorithms. A tax focused on data processes may have other (environmental) benefits, but it is unlikely to be an adequate proxy for taxing online polarization.

[59] Note that in the EU, the ICT industry is left out of the EU Emissions Trading System.

3.3e Taxing data brokerage

Data brokerage is the business of collecting and selling personal data to interested parties without the data owners being aware of this transaction. It is a multibillion dollar industry that is growing fast – and it is central to surveillance capitalism. Data brokers obtain data from third parties (primarily credit card providers and retailers), from public records available on the internet, and from social media. Brokers collect and aggregate this data into resellable packages of interest primarily to marketers and advertisers, but also to insurance brokers, law enforcement, and various government agencies (national or foreign). "Data Brokerage Tax", a paper written by Eticas in this Next Now report series, outlines the literature and policies relating to the implementation of a sales tax on data brokerage, and explains how this could incentivize companies to be more transparent about how they collect and use personal data.

The link between data brokerage and the training of algorithms is more tenuous, but the connection with online polarization is still clear. Digital platforms collect their own data, and thus are not the primary clients of data brokers. However, some digital platforms are also data brokers – for example, Facebook and Google collect data from their users and then allow anyone using their ads platform to pay to use it. In these transactions, digital platforms are potentially enabling the use of profiling by actors that deliberately set out to manipulate and harass. And as we saw in section 1, the amplification of manipulation and harassment can eventually lead to widespread online polarization and real life consequences. Overall, a sales tax on data brokerage would impact some financial incentives of digital platforms, but its impact on the spread of online polarization is far from clear.

We are not arguing here that a sales tax on data brokerage should apply only to digital platforms: there are good reasons to curtail the operation of all data brokers. Data brokerage enables surveillance by private and public bodies, which is not only a threat to individual privacy, but also to democratic values. [60] This negative societal impact of surveillance capitalism – how it threatens democracy and freedom by covertly enabling authoritarian practices – has not been the focus of our paper, but is of course also important. A sales tax on data brokerage – applied to all brokers – would significantly disrupt the operations of big and powerful data brokers, and cut to the heart of the surveillance capitalism model.

[60] https://www.wired.com/story/opinion-data-brokers-are-a-threat-to-democracy/

3.4 Conclusion

This section argues that online polarization is a negative externality of algorithmically mediated platforms that results from the incentives of the surveillance capitalism model. If we understand online polarization as a negative externality, then there is a policy argument in favor of creating a financial disincentive to its production. We argue that this could be done by taxing polarization as a negative externality, and we explore five possible avenues for taxation: one based on a direct measure of polarization, four based on proxies that affect the production of online polarization.

Although much work remains to be done to make these taxation options viable, this section sets out options for credible policy pathways to make surveillance capitalism more expensive in order to reduce the impact of digital platforms on societal divides. Based on our assessment of these options, we propose two concrete next steps:

Measuring the "polarization footprint" of digital platforms

Piloting approaches to measure the "polarization footprint" of digital platforms, and the viability of proposing a taxation regime based on this measure

Further research on whether a non-linear tax on large databases might reduce polarization.

Researching the connection between personalization / content targeting based on certain kinds of personal data and polarization outcomes, in order to understand whether introducing a non-linear tax on large databases might reduce polarization.

About the author

Helena Puig Larrauri is a peacebuilding professional with over a decade of experience advising and working with civil society actors and multi-lateral organisations in conflict contexts and polarized environments. She holds a BA in Politics, Philosophy and Economics from Oxford University and a Masters in Public Policy (Economics) from Princeton University. She specializes in the integration of digital technology and innovation processes to peace processes and civic dialogues, and in understanding how conflict drivers show up in digital spaces. She has published and spoken on these subject matters extensively. She is also an Advisor on Digital Technologies and Mediation to the United Nations Mediation Support Unit and an Ashoka Fellow.

Contact: helena@howtobuildup.org