INTERNET POLICY REVIEW
Journal on internet regulation

Volume 13 **Issue** 3

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Contesting the public interest in AI governance

Tegan Cohen *Queensland University of Technology (QUT)*
Nicolas P. Suzor *Queensland University of Technology (QUT)* n.suzor@qut.edu.au

DOI: https://doi.org/10.14763/2024.3.1794

Published: 30 September 2024
Received: 7 March 2024 **Accepted:** 31 July 2024

Funding: This research was funded through an Australian Research Council Future Fellowship (FT210100263) with support from the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005).
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Cohen, T., & Suzor, N.P. (2024). Contesting the public interest in AI governance. *Internet Policy Review*, *13*(3). https://doi.org/10.14763/2024.3.1794

Keywords: Artificial intelligence, AI governance, Democracy, Public contestability, Public interest

Abstract: This article argues that public contestability is a critical attribute of governance arrangements designed to align AI deployment with the public interest. Mechanisms to collectively contest decisions which do not track public interests are an important guardrail against erroneous, exclusionary, and arbitrary decision-making. On that basis, we suggest that efforts to align AI to the public interest through democratic participation will benefit substantially from strengthening capabilities for public contestation outside aggregative and deliberative processes. We draw on insights from democratic and regulatory theory to explore three underlying requirements for public contestability in AI governance: (1) capabilities to organise; (2) separation of powers; and (3) access to alternative and independent information. While recognising that suitable mechanisms for contestability will vary by system and context, we sketch out some possibilities for embedding public contestability in AI governance frameworks with a view to provoking further discussion on institutional design.

This paper is part of **AI systems for the public interest**, a special issue of *Internet Policy Review* guest-edited by Theresa Züger and Hadi Asghari.

1. Introduction

A small group of private companies are currently competing to develop highly capable general-purpose AI models.[1] Over the course of 2022-23, the public release of a number of chatbots built on large language models, including OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude, gave many a first glimpse of the massive potential of general-purpose AI to both serve and harm the public interest. Servers buckled under the weight of the demand of millions of users discovering creative and surprising ways to extract information, poetry, recipes, and advice from the chatbots. Inevitably, many users discovered problems as well. Chatbots hallucinated, spat out emotionally manipulative dialogue, and generated racist and sexist content after some moderately imaginative prompting (Deshpande et al, 2023; Weise & Metz, 2023). Insight into how these models could and would behave in the real-world inevitably sparked debates about how they *should* behave. The idea that AI should align with the "public interest" is one of the most common themes. Commitments to the public interest not only appear frequently in government AI policy documents but in the mission statements of companies at the frontier of development (Anthropic, 2023a; Brockman et al., 2015). The apparent consensus around commitments to the public interest, however, masks the lack of consensus about what exactly those commitments require.

This ambiguity arises, at least in part, from the open-textured nature of the public interest as a standard. It is broadly accepted that the public interest transcends "private" interests (be they individual or commercial). Beyond this tautological formulation, there is no universal definition. The concept is "situation dependent and dynamic" (Züger & Asghari, 2023, p. 816) and heavily contested. Values closely associated with the public interest, such as fairness, are equally susceptible to multiple, context-sensitive interpretations. Without consensus about what the public interest requires in AI regulation, meta-questions of governance become increasingly salient: who decides what kinds of AI behaviour and uses align with the public interest? How are disagreements resolved? And how is account for those decisions

1. The term "general-purpose AI" is used throughout this article to refer to an "AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained" (Gutierrez et al., 2023, p. 5). General-purpose models may be large language models, trained on enormous unlabelled data sets. These models will usually form the base infrastructure for specialised downstream applications.

rendered?[2]

In this article, we argue that public contestability is crucial for attempts to design corporate governance arrangements to align AI deployment with the public interest. It is now common for commercial firms and public policymakers to cite increasing civil society and public participation in AI governance as an aspiration on the assumption it will yield better understandings of the public interest and values. In line with broader trends in corporate social responsibility and legitimacy seeking (Caulfield & Lynn, 2024), practical efforts by corporations to incorporate public input draw heavily on deliberative accounts of democracy, which promote structured and orderly processes for reaching reasoned and rational consensus on the common good. We argue that, along with inclusive participation and deliberation, space for contestation is essential to the interminable processes of defining and defending the public interest.

With the term public contestation, we invoke notions of both competition and challenge. In electoral democracy, contestability is a measure of the degree to which incumbents face challenges from oppositional forces (Dahl, 1971). Outside the electoral contest, contestability describes the possibility that citizens may challenge political decisions which they consider do not track the public interest, through institutions such as courts and ombudsmen (Pettit, 1999, 2000). In each case, the spectre of contestation is a key impetus for powerholders to respond, and remain responsive, to constituent or stakeholder interests (Bottomley, 2007; Dahl, 1971). Public contestation can also serve as a backstop against the exclusion of marginalised or unimagined interests. Understood in this way, contestation is a stimulant for continuous negotiation, refinement, and alignment with the public interest. We use contestability to denote a property of governance systems that permit and are responsive to public contestation. Across different democratic theoretical traditions, public contestation is recognised as a critical check on error, exclusion, and hegemony in public governance. As AI companies continue to seek legitimacy by invoking democratic values, we consider the benefits of, and possibilities for, building public contestability into AI governance in the private sector. Our focus is on the institutional design choices that impact the ability of public stakeholders to contest corporate decisions about AI development and use. By public stakeholders, we mean those individuals and groups external to an AI firm that are affected by, depend upon, or have an interest in the deployment of its products. We

2. These core questions are inspired by the Institute on Governance's conception of the central concerns of governance, namely: "How are decisions made? Who has a voice in making these decisions? Ultimately, who is accountable?" (Institute on Governance, n.d.).

suggest that efforts to realise commitments to the public interest by "democratising" AI governance will benefit substantially from explicitly strengthening capabilities for public contestation.

We explore three interlocking conditions for public contestability: (1) capabilities to organise; (2) separation of powers; and (3) access to alternative and independent information. Using these categories to organise the discussion, we identify various mechanisms to support these conditions in nascent AI governance frameworks, including new facilities for collective bargaining and legal action in data governance regimes, power separation in "hybrid" non-profit/corporate structures, and proposed audit regimes. Observing shortcomings in existing approaches, we consider the possibilities for improving and securing the contestatory power of groups at different levels (meso and macro) and phases (setting limits, oversight, enforcement, etc.) of governance. While the focus of the article is on the development and use of AI by the private sector, we recognise that both private and public actors have a role to play in building capabilities for public contestability. As such, some of our suggestions for institutional design could be implemented by company directors and executives, while others will require lawmakers to intervene.

While many of the examples we draw on to illustrate the abstract requirements for contestability relate to AI generated content and moderation, the observations are neither exclusive nor tailored to those particular contexts. In fact, it is important to note at the outset that the optimal degree and mechanisms of contestability in a system of governance will vary depending on, among other things, the type of system and its context of use (Henin & Le Métayer, 2022). The proper mechanisms and degree of public contestation will be different for AI models tasked with evaluating humans and those tasked with generating text and images, or foundational models applied for medical diagnosis (Moor et al., 2023) and those deployed for political persuasion (Bai et al., 2023). The makeup of "the public" will also turn on the sources and types of knowledge that a model/application seeks to represent. Not all decisions and systems should be open to public contestation by the same publics or to the same degree, nor should formal rules and structures determine all forms and possibilities for contestation. Our aim is to offer some generalisable principles to guide but not prescribe institutional design.

The article proceeds in three parts. Part II considers the predominant approaches taken by AI firms to seeking input from public stakeholders, the democratic theories reflected in those approaches, and their limits. Part III then draws on insights from democratic and regulatory theory to explore three underlying requirements for public contestability in AI governance and options for practical implementa-

tion. Part IV offers some concluding remarks on the interdependencies between these requirements and future directions for research.

2. AI and the public interest: Who decides?

Over the last decade, practical consensus has emerged around the need for meaningful ethical safeguards in the development and deployment of AI. At the highest level, the ethical principles take the form of broad commitments to public interest values, including "fairness", "non-maleficence", "transparency", and "equality", among others. At this level of abstraction, these principles are largely unobjectionable; but settling on their meaning in a given context requires contestable value judgements and unavoidable trade-offs (Mittelstadt, 2019; Palladino, 2023). The problem with the public interest is that it can seem like an "empty vessel", waiting "to be filled with whatever values the user wishes" (Feintuck, 2004). Central topics in contemporary debates about generative AI governance (*What are the boundaries of acceptable content (or "speech")? What kinds of work (creative, knowledge, sex, emotional, or otherwise) should AI be allowed to perform? What acceptable role can it play in political persuasion?*) can each be argued by reference to competing interpretations of the public interest. The room for disagreement in these debates leads to clear concerns around *procedure*: who gets to make and enforce the rules, and what are the avenues for appeal and redress? Where technology companies had enjoyed relatively large degrees of control over these decisions since the birth of the commercial internet, they now struggle to maintain legitimacy (Suzor, 2019).

In response to increasing criticism, technology companies have explicitly relied on democratic rhetoric as a key source of legitimacy. Frontier AI companies have publicly declared a desire to cede some governance power and orient their technologies toward common interests (Seger et al., 2023).[3] Experiments with stakeholder democracy in the technology sector have taken both aggregative and deliberative forms. In terms of aggregation, technology companies have sought to "align" AI models with varied proxies of popular opinion, using techniques such as polling users on predetermined questions (Schrage, 2012) and A/B testing for preferences on chatbot responses. While aggregative models are good for identifying the majority choice or lowest common denominator, these are not the same as the public

3. For example, Stability AI's CEO has expressed discomfort at the prospect of a "centralised, unelected body" lording over "the most powerful technology in the world" (Roose, 2022, para. 15). OpenAI's charter includes a commitment to avoid "unduly concentrate power" (OpenAI, 2018, para. 3). Meta AI has adopted its open source approach with the stated hope that "democratising access" will "bring more voices to the frontier of large language model creation [and] help the community collectively design responsible release strategies" (Zhang et al., 2022, para. 10).

interest. An interest might be common (e.g. freedom from harassment or the right to speak one's own language) but a specific threat to that interest may not attract majority recognition when it is disparately experienced by a marginalised group (Pettit, 2000). In constitutional democratic theory, inalienable, entrenched, or derogable charter rights impose checks on the "tyranny of the majority". One of the most prominent responses from technology companies to this problem has, unsurprisingly, been to look to technical solutions for voting systems that can efficiently aggregate and prioritise preferences across a large, and geographically dispersed user-bases, such as Tools for Quadratic Voting (RadicalXChange, n.d.). These infrastructural innovations have not overcome the classic problems; quadratic voting methods, for example, seek to overcome majority tyranny by weighting the intensity of preferences (Lalley & Weyl, 2017) but they do not constrain majorities strongly opposed to equity.

In contrast to preference aggregation, some large commercial AI labs have launched more high-profile, ambitious experiments with discursive approaches to democratisation. Meta, for example, has piloted citizen assemblies (Clegg, 2023; Halpern & Costa, 2022). Anthropic recruited 1,000 US participants to vote and propose rules for Claude's "Constitution" (Anthropic, 2023b). OpenAI launched a funding programme for research on democratic processes for governing AI systems (Zaremba et al., 2023). These deliberative models are meant to encourage individual participants to transcend self-regarding interest through reason-giving and consensus. According to deliberation theory, discursive processes allow participants to engage in complex policy areas and build empathy. However, scepticism about whether company-controlled democratisation programmes will truly disperse decision-making power lingers.

Enthusiasm for the so-called "participatory turn" in AI design appears widespread, but real opportunities for public participation in commercial AI development are still scarce. In the past, technology firms have leant toward keeping participatory processes "centralized, with the company in control" (Caplan, 2023, p. 3458). In some cases, deliberative approaches can obscure the contingent, contested nature of the public interest. By de-emphasising dissent, deliberative procedures may conceal when arrival at a consensus "entails some form of exclusion" (Mouffe, 1999, p. 756) or reflects structural and material inequalities among participants (Banerjee, 2022; Sanders, 1997). Deliberative forums may alienate participants by privileging conformity to particular discourse norms and/or ignoring the significance of silence and refusal (Rollo, 2017). Later generation deliberative theories seek to acknowledge diverse speech cultures and clarify rather than avert dis-

agreement (Bächtiger et al., 2018; Rollo, 2017), but multiple forms of exclusion remain a risk in corporate implementations.

First, the pool of participants may not represent the full spectrum of values, interests and views held by public stakeholders. Putting together a fully representative "mini public" for a general purpose AI model is practically difficult. Some firms, such as Anthropic, are looking to technical solutions to widen participation (Anthropic, 2023b); "collective response systems" such as *Polis* and *Ramesh* seemingly hold promise for facilitating more productive deliberation at scale (Ovadya, 2023). Yet exclusion is likely to be a persistent problem in deliberative processes about AI behaviour, particularly for general purpose models, where the boundaries of "the public" (or impacted persons) are fluid, shifting and expanding over time as new use cases, system properties, and effects emerge. Without careful attention to the question of who gets to be in the room, deliberating citizens tasked with considering the common interests of all public stakeholders will often lack mandates and competence to make assumptions about the situation of those absent. These are not necessarily insurmountable problems, but the task of doing deliberation effectively at scale likely requires some radical new experimentation (Young et al., 2024).

Second, exclusion resulting from a restricted agenda is a risk (Young, 2001). Consultation is often hampered by tight project deadlines, resource constraints, and, for general-purpose technologies, a lack of clear understanding of contexts of use (Delgado et al., 2023; Groves et al., 2023). In practice, consultation is often limited to discrete system components, such as user interfaces, select tools, or specific policies (Delgado et al., 2023). Interview studies with AI practitioners suggest that commercial labs tend to involve the public in "narrowly scoped, rapid input consultations" rather than problem formulation and agenda-setting in the early stages of product development (Groves et al, 2023, p. 42). Firms may, intentionally or not, exclude contentious matters from the scope of deliberation, circumscribing actions and approaches to those least disruptive to their commercial objectives. Opportunities to participate, the questions and choices open to public input, and the "right to judge the legitimacy or feasibility" of recommendations remain within company discretion (Arnstein, 1969, p. 220). In these circumstances, many view participation as too easily co-opted by "powerholders to claim that all sides were considered" while maintaining "the status quo", functioning as a "public relations vehicle" rather than a social inquiry into the public interest (Arnstein, 1969, pp. 216, 218; Birhane et al., 2022).

Space for public contestation can mitigate these limits. Calls to account for public

contestation in the governance arrangements for transformative technologies are not new. Scholars have previously advocated and proposed a variety of approaches to embed contestability in AI design (Alfrink et al., 2022; Hildebrandt, 2019; Hirsch et al., 2017; Kluttz et al., 2020). Kars Alfrink and colleagues have advocated for contestability throughout a system's life cycle ("contestability-by-design"), reasoning that conflict is not only inevitable but "desirable as a means of spurring continuous improvement" (2022, p. 632). Earlier, Kate Crawford and Catharine Lumby argued for radical pluralism in digital platform governance by allowing the many and diverse actors invested in online content regulation to dissent and contest values (2013).

If we accept that the public interest is a contestable, contingent, evolving ideal, it follows that its meaning and achievement should be open to refinement, iteration, and falsification over time. The need to edit and iterate will be particularly significant in the case of general-purpose models, as new applications and externalities emerge over time, and maturing businesses drift into path dependency and group think. Contestation can serve a productive, disruptive function in this process, by giving voice to excluded parts of "the public" and agenda. Moreover, public contestation provides incentives for those in power to be responsive to articulated public interests, which may be low in centrally controlled participatory processes (Arnstein, 1969; Birhane et al., 2022; Sloane et al., 2020).

3. What institutions does public contestability of corporate AI governance require?

In the following sections, we explore three of the most salient requirements and associated challenges for increasing contestability in AI governance: (a) capabilities to organise; (2) separation of powers; and (3) access to independent and alternative sources of information. We extend these requirements, which democratic theorists commonly cite as necessary for democratic contestability, to commercial AI governance. These requirements are not exhaustive but rather a starting point for considering how public contestability can be embedded into AI governance frameworks.

3.1. Capabilities to organise

Contestability requires that stakeholders be able to collectively recognise, organise, and channel their common interests, concerns, and experiences into demands. The reasons echo some of the main rationales for associational autonomy in large-scale electoral democracy. Collective action helps empower citizens to influence

and contest the use and abuse of governing power. Forming independent organisations is not only desirable to amplify common concerns but necessary to accrue power and influence in large, complex governance systems (Dahl, 2005; Trantidis, 2017). Organising affords opportunities to acquire political skills and share knowledge and resources. In the context of technology governance, deficiencies in technical expertise and material resources are a common barrier to participation by civil society (Galvagna, 2023). Participatory democrats stress that "individuals learn to participate by participating" (Pateman, 2012, p. 10). But participation is costly and time intensive. Organising allows resource pooling, exchange, and capacity-building at scale. Firms looking to increase participation will likely need to invest heavily in institutional sustainability and capacity building as well as underwriting the direct costs of full participation for those not already in the room. Above all, collective action is critical to the pursuit of the public interest which, by definition, transcends the individual.

3.1.1. A lack of institutional routes for collective action

Capabilities to collectively contest commercial AI deployment across different contexts are deeply constrained. Most of the reasons are familiar. Organising can ameliorate resource deficiencies, but public interest groups still face economic barriers to entry (Galvagna, 2023). Systemic failures, which provide a common cause for association, are sometimes obscured by institutional or technical design (Pasquale, 2015). How to deal with the issues of resource scarcity and system design is an important question already the subject of useful research and reform. Our primary focus is on a third (interlocking) factor: institutionalised opportunities for groups to contest AI development and use. Of course, not all collective action will or should be shuttled through formal institutional routes. The efforts of activists who place traffic cones on the bonnets of driverless cars in San Francisco (Paul, 2023) or deploy occlusion and confusion techniques to evade facial recognition (Thomas, 2019) are incompatible with stabilised, "official" routines. Activism that deploys ephemeral, creative tactics and aims at remaking institutions will necessarily take place outside of them. Activists and public interest groups will understandably be wary of consultation theatre (Clement, 2023; Geist, 2021), efforts to depoliticise opposition (Urbinati, 2010) and co-opt activist energy (Young, 2001). Our goal is not to advocate a purely institutionalised and procedural approach, but to acknowledge that the absence of established routines, structures, and resources for contestation in AI governance structures can make it more cumbersome for many to express dissent, as groups need to expend more time and resources to have their viewpoints heard. The emotional labour of digitally mediated activism (Gleeson, 2016; Tufekci, 2017), gaining entry and participating in technology gover-

nance, is already borne unequally by marginalised groups (Mannell & Smith, 2022). While not solely determinative, the availability and design of formal avenues for contestation can help shape how resources are allocated, how power is distributed, and what types of public interests are pursued and are thus deserving of attention.

Generally speaking, avenues for public interest groups to challenge decision making are limited in AI governance structures. Procedures for review and appeal the outputs of AI systems are embedded in various platforms and legal frameworks but centre around individual complaints. More specifically, some digital platforms have dedicated procedures for appealing content moderation decisions, which have been extended by statute to incorporate review by external arbitration bodies in certain jurisdictions (see, e.g. Online Safety Act, 2021, ss30(4), 36(3); Regulation 2022/679, art. 18). Some jurisdictions have taken steps to codify due process rights for individuals in automated and digital decisional contexts by mandating a human-in-the-loop (see, e.g. LGPD, 2018, art. 20; Regulation 2016/679, art. 22; Regulation 2022/2065, art. 17). These avenues primarily afford individuals recourse against fallible automated decision-making systems, especially those that review or rate human performance (Henin & Le Métayer, 2022; Hirsch et al., 2017; H. Lyons et al., 2021a). Affordances for collective advocacy via these channels are limited, something users and academics have lamented (Vaccaro et al., 2020, 2021).

Similarly, class or representative legal actions in relation to AI deployment are bound by several limits. The need for a clear cause of action and proof of causal responsibility will often preclude claims based on novel problems or unprecedented threats to the public interest (Fraser & Suzor, 2024). In some jurisdictions, requirements for plaintiffs to particularise concrete injury will impede actions to prevent hypothetical harm. Attempts to adapt existing causes of action which have developed around *individual* rights, such as breaches of data protection statutes or misuse of private information, to address an injury to the *public* interest will be hampered by requirements for individualised assessments of damages (Lenz, 2020; *Lloyd v Google LLC*, 2021). Contractual grounds for action are also meagre, as platforms generally take on few obligations in their terms of service (Suzor, 2018). Low prospects of damages will, in turn, disincentivise plaintiff lawyers from taking on the time-consuming and costly task of bringing claims (Lenz, 2020). Finally, private AI firms may seek to erect additional barriers to class actions in their terms of service.[4]

4. OpenAI's Terms of Use dated 31 January 2024 include a class action waiver which states that "You

Avenues to collectively demand review and appeal of AI outputs are important for public contestability, as many will lack the resources and time to bring individual claims (Alfrink et al., 2022; H. Lyons et al., 2021b; Pettit, 2000). However, there need to be justiciable limits on AI deployment before courts and tribunals can intervene. As such, liberalising standing requirements to facilitate public interest litigation of AI deployment will only go so far without complementary opportunities for public stakeholders to contest the limits, norms, and restrictions on use.

3.1.2. Private sector initiatives and procedures

Avenues for groups to influence the behaviour of AI systems do exist within corporate governance structures but are limited and unequally distributed. Multistakeholder initiatives at the organisational (such as trust and safety councils) and industry level (such as the Partnership on AI) have multiplied in recent years. However, the success of these initiatives in increasing civil society influence over the objectives, values, and policies of technology companies has been mixed. In practice, engagement with public interest groups often takes place on an informal and ad hoc basis (Caplan, 2023; Suzor & Gillett, 2022). Expert and civil society participants themselves have lamented the lack of responsiveness by firms. External stakeholders frequently must leverage personal relationships or public scandal to exert influence. Activists harness ephemeral channels, informal alliances, and innovative and opportunistic tactics to eke out responses from technology companies. For example, #freethenipple campaigns leveraged multiple channels (petitions, virtual sit-ins, physical protests) and influence networks (media, celebrities, activist groups, artists) to procure incremental changes to content moderation policies (Myers West, 2017). Platform workers on Amazon Mechanical Turk and various food delivery apps have coordinated with networks of activists and researchers, solidarity unions, and grassroots community groups, to stage proto-strikes, demonstrations, sousveillance, letter-writing, brand-shaming, and public information campaigns to contest algorithmic control and other aspects of their working arrangements (Cini, 2023; Kellogg et al., 2020; Gray & Suri, 2019). Still, the successes of these organised efforts are limited. Influence is arbitrarily and unequally distributed, with companies more attentive to the concerns of groups that represent major sources of revenue (Griffin, 2023).

Open procedures for feedback, such as options to flag harmful or inaccurate con-

and OpenAI agree that Disputes must be brought on an individual basis only, and may not be brought as a plaintiff or class member in any purported class, consolidated, or representative proceeding. Class arbitrations, class actions, and representative actions are prohibited." Class action waivers are not enforceable in all jurisdictions, such as the EU, which may be why equivalent provisions do not appear in OpenAI's Terms of Use for the EU dated 15 February 2024.

tent,[5]typically do not explicitly acknowledge or facilitate coordination amongst individuals. Yet, individuals tend to find ways to coordinate, *to organise*, in order to maximise influence and achieve common goals. Recall Microsoft's short-lived release of the chatbot "Tay" on Twitter in 2016, now a cautionary tale of loosely coordinated malicious influence on AI behaviour.[6] Not all coordinated efforts to subvert rules or exploit system vulnerabilities will be malicious. In a study of *Decide Madrid* and *vTaiwan*, two initiatives which have fuelled recent enthusiasm for techno-deliberative democracy, Yu-Shan Tseng observed how "groups of highly networked citizens exploit game rules and mobilise online/offline networks" (Tseng, 2023, p. 1336). As Tseng notes, the initiatives exhibit "old problems with digital democracy – the tendency of participatory processes to be influenced by the same group of participants" (Tseng, 2023, p. 1336). But coordination is not itself the problem. Issues arise when coordinated influence is unanticipated and hidden from scrutiny. Variations in motivation and ability to organise across the polity, as well as malicious actors, are a consistent feature of political processes. Processes which do not explicitly recognise and foster groups will rarely preclude them from forming but can help *obscure* coordinated behaviour and nodes of power, while erecting additional barriers to collective action which only highly motivated and skilled participants can circumvent.

3.1.3. Public regulation

Various national and supranational AI regulation strategies contemplate opportunities for civil society participation in rulemaking and interpretation. For example, civil society groups are expected to play a role in representing the public interest during the development of harmonised standards and codes of practice to clarify the implementation of obligations in the EU's proposed AI Act. As previous critiques point out, capabilities for civil society groups to meaningfully shape rules in the public interest in standardisation forums are constrained by historical industry dominance and myopic focus on "technical" solutions (Galvagna, 2023; Gornet, 2023; Perarnaud, 2023). "Outsiders" to standardisation procedures will likely need to contend with rituals, rules, and norms of participation that advantage estab-

5. At the time of writing, feedback options for two major chatbots, Anthropic's Claude 3.5 Sonnet and Google's Gemini, are presented to users as "thumbs up" and "thumbs down" icons beneath the bot's responses. Thumbs up and down responses prompt users to classify the content (e.g. Google provides the options offensive, not factually correct, and other and "provide additional feedback" for a thumbs down response).

6. Within a few hours of launch, the chatbot was verbosely spouting racist and sexist language. Microsoft quickly removed the bot, blaming "a coordinated attack by a subset of people [which] exploited a vulnerability in Tay" (Lee, 2016) – a link to the bot's Twitter account had appeared on 4chan encouraging users to exploit its "repeat after me" function (Schwartz, 2024).

lished players in these forums.

3.1.4. Strengthening capabilities to organise

Institutional designers could seek to improve capabilities to collectively contest AI deployment in at least four respects. First, opportunities for contestation should extend across different phases of AI governance. To pursue the public interest, groups need opportunities to participate in *rulemaking*. In practice, this could mean lawmakers carving out formal pathways for public interest groups to negotiate the ground rules directly with commercial firms. A flawed precedent here is the state interventions in "inter-capitalist competition" (Flew et al., 2021, p. 2) between news publishers and digital platforms (see, e.g., Bill C-18; Treasury Laws Amendment (News Media and Digital Platforms Mandatory Bargaining Code), 2021). Tentatively, proposals to develop legal frameworks for data cooperatives to negotiate the terms and conditions of data collection and processing with AI developers more directly target the power asymmetries that insulate large corporations from the interests of public stakeholders (Ada Lovelace Institute & AI Council, 2021; Benthall & Goldenfein, 2021). To make *oversight* contestable, lawmakers could ensure that there are multiple channels for third parties to access information and participate in audits and impact assessments (see section 3.3).

Second, access to contestation channels should be distributed widely. This does not mean that participation in all forums must be open access. Concerns about bad faith and vexatious actors should be addressed through objective criteria to gain, say, legal standing or special status as a data coop. However, formal opportunities for public interest groups to participate in AI governance at both the organisational and state level should not be divvied out on a purely invitational basis.

Third, institutional design choices should be made with existing power structures in mind. As discussed above, failing to acknowledge coordinated influence or the historical dominance of one interest group (e.g. industry) in procedural and institutional design will merely obscure and entrench existing power imbalances. Lawmakers could provide mechanisms to challenge power dynamics in new or existing institutions (such as standardisation bodies) such as by making key roles (like chair) contestable or allowing members to publish dissenting opinions.

Fourth, multiple, diverse channels for public contestation will increase opportunities for public interest groups to combine pressure from other actors (Hong & You, 2018). As noted, collective campaigns for platform worker rights and changes to content moderation have successfully exerted influence by leveraging a variety of tactics and channels. Activists and advocacy groups in public sector contexts have

leveraged public forums to build network pressure. For example, the errors and illegality of the Australian Government's robodebt scheme were exposed through a combination of direct complaints to the responsible government department, extensive social media campaigning, mass media coverage (first instigated by a whistle-blower), class action lawsuits, as well as reviews by institutional watchdogs – Ombudsman investigation, multiple Senate committee inquiries, and finally, a Royal Commission (The Royal Commission into the Robodebt Scheme, 2023). In the next section, we discussed how the targets for public pressure throughout governance networks could be multiplied through institutional design.

3.2. Pluralised separation of powers

The separation of powers has primarily been associated with theories of public power, namely, its distribution between the legislative, executive, and judicial branches of government. The main justification invoked in political theory for distributing the functions of government across multiple institutions is to prevent concentration and abuse of power. Through a system of checks and balances, separation guards against powerholders acting with impunity. In the absence of other democratic mechanisms, such as public control over company leadership, power division could help to render account for corporate decision-making.[7] Another rationale for separating power which resonates with current AI governance challenges is to break government action down into component phases "so that the various aspects of law-making and legally authorized action are not just run together into a single gestalt" (Waldron, 2013, p. 457). Instead of treating government action as something simple, the separation of powers induces an "articulated government through successive phases of governance each of which maintains its own integrity" (Waldron, 2013, p. 467). The articulation rationale, we think, chimes with the notion that separation gives rise to productive contestation between branches. This is the idea that dividing governing power between distinct functions, institutions, and personnel allows for scrutiny and counterviews to be aired out in the open.

In this way, separation of powers can provide a brake on decision making, whereas centralised governance structures enable swift action. In recent history, technology companies have placed great emphasis on the ability to act quickly. Several frontier AI companies are notable for their highly centralised governance structures

7. It is worth acknowledging that, while we use the terms interchangeable in this article, the "separation of powers" and "division of powers" is sometimes distinguished in constitutional theory, with the former being used to refer particularly to legislative, executive, and judicial separation (Waldron, 2013).

that give founders and company leadership special powers to defeat popular votes by ordinary shareholders (Aggarwal et al., 2022).[8] Various governance scandals in the technology industry illustrate the accountability deficits associated with centralisation. One stark example comes from Uber where, for eight years, co-founder Travis Kalanick held super-voting shares, controlled three board seats and, in his capacity as CEO, exercised a wide-ranging remit over operations. Under Kalanick's stewardship, the business underwent a period of rapid growth in value and reach, which many attributed to a risk-seeking *modus operandi*. It also became embroiled in a long list of regulatory, governance, and cultural controversies, which culminated in a PR crisis in 2017. The board's response was to initiate reforms to decentralise governance power within the company, adopting a series of measures which diluted Kalanick's control and increased contestability among shareholders (Larcker & Tayan, 2017).

A small group of AI firms with public interest objectives have adopted corporate structures designed to impose public interest checks on economic interests. At Anthropic, for instance, a "long term benefit trust" controls several board seats (Anthropic, 2023a). Mozilla.ai, like the Mozilla Corporation, is a wholly owned subsidiary of the non-profit Mozilla Foundation, established to support its open access, public benefit mission (Surman, 2023). Originating as a non-profit, by the time OpenAI kicked off the current AI race with the release of ChatGPT3.5, the outfit had morphed into a capped-profit company, overseen by the original non-profit board. According to its architects, the corporate structure was designed to preserve OpenAI's foundational commitment, while allowing it to fund the immense cost of computing and brain power – balancing the pursuit of profit and public interest (Brockman et al., 2019).

The implosion of OpenAI's non-profit board in late 2023 could be read as evidence of a failure of structural design (Dave, 2023). In late 2023, Sam Altman, the CEO of the capped profit, was temporarily ousted by the non-profit board before most of its members subsequently resigned. An independent review of the incident conducted by law firm WilmerHale cited a "breakdown in trust" as the prime reason for the board's decision to fire Altman (OpenAI, 2024). Altman's return was instigated by pressure from investors and staff, who threatened to resign if Altman was not reinstated. In theory, the organisation's structure separated the guardians of the firm's public interest mission from its for-profit operations in a formal hierarchy between the (public interest) parent and operational subsidiary. In reality, the non-profit board was subject to external checks by other stakeholder groups, most

8. Alphabet and Meta, for example, have multi-class share structures with different voting rights.

visibly employees and financial partners (particularly Microsoft), which ultimately persuaded it to capitulate.

Power division has the potential to create interstitial space for dialogue about public interest concerns. However, an arrangement where one organ is given the power to dominate the other (through veto or removal powers) will lack formal incentives to articulate the exercise of that power, to treat the power as complicated rather than simple. The OpenAI case is a good example. The board failed to articulate clear, specific, and persuasive reasons for its decision to the relevant stakeholder groups at the time, making the decision *appear* arbitrary to outsiders.

How might we implement separation of powers to increase the accountability of corporate executives and directors to public stakeholders? There is no clear best practice guide to institutional design that replicates the functions of the separation of powers in corporate governance while also preserving (and in some cases strengthening) centres of power and interdependencies between them. The separation of powers is a simplified abstraction in any political system; any workable attempt to overlay this abstraction to corporate governance arrangements likely has to recognise that, in messy reality, "domain over social issues is… dispersed across a number of actors within a system of societal governance" (Caulfield & Lynn, 2024, p. 38). In practice, an ideal and workable power separation arrangement involves "many semi-autonomous powers recursively checking one another" (Braithwaite, 1997, p. 312). Actors are "semi-autonomous" when they are not "so independent or autonomous that they are beyond effective control or scrutiny" by others but just autonomous enough to perform their roles and check other powers through partial participation in their functions (Bottomley, 2007, p. 93). These semi-autonomous powers could include external actors at different levels of governance and internal actors or "sub-firm constituencies" (Gorwa, 2024) guided by independent interests or allegiances outside the firm. Think of an internal auditor, OH&S officer, or quality manager with a statutory or professional duty to report wrongdoing, or a pharmaceutical CEO who must sign off on a production manager's decision to overrule the adverse findings of a quality control manager (Braithwaite, 1997, pp. 351-352). In each case, power division will, at minimum, encourage dialogue between the actors. An arrangement of many, semi-autonomous powerholders recursively checking one another could help build pressure to respond to public interests through dialogue, persuasion, and shaming, rather than coercion.

3.2.1. Many, semi-autonomous powers

What would a pluralised division of public and private powers look like in the context of AI governance? We can start by asking: which actors embedded within commercial AI development cycles are well-positioned to impose checks? Candidates might include red teams, compliance and safety auditors, and ethics boards, who are sometimes engaged by AI companies to identify and advise on the societal impact of their products. In other sectors, ethical duties monitored by professional boards and statutory disclosure duties have helped secure the independence and integrity of similar functions (Raji et al., 2022). Ideally, actors who have greater involvement in daily decision-making would join this list. Proponents of professionalising AI engineering argue that the threat of personal liability or licence revocation would incentivise engineers to think twice about failing to audit training data for bias or otherwise cutting corners to meet a commercial timeline or objective (Sharma, 2023). Commitments might also be taken up voluntarily through unionising or professional associations; witness, for instance, the progressive professionalisation of online trust and safety bolstered by initiatives like the Trust & Safety Professionals Association.

Outside AI firms, certain up- and downstream supply chain actors may be positioned to impose checks on development. Frameworks encouraging certain actors to utilise their leverage to make public interest demands are already forming. Ethically centred AI procurement guidelines encourage downstream government clients to harness their buying power to procure binding commitments to public values in AI development, which they proactively monitor and enforce through customary contractual mechanisms (Hickok, 2022) under the specter of executive or public action. Upstream, effectively functioning data cooperatives and data trusts would empower data subjects to negotiate contractual limits on data processing and rights to enforce those limits. While third parties are ordinarily precluded from seeking to enforce contracts to which they are not privy, limited exceptions could add an additional layer of external checks. Specifically, allowing affected data subjects to bring good faith claims on behalf of a data intermediary unwilling to exercise its contractual rights could help deter complacence.[9] Arrangements which enable public interest groups to enforce or seek declaratory relief in relation to public sector AI contracts could have similar effects.

9. Stephen Bottomley recognises the statutory derivative action as a key contestability mechanism in corporate governance (2007).

3.2.2. Instilling independence

These suggestions for power and labour division are by no means prescriptive or exhaustive. As noted at the outset, the degree and mechanisms for public contestability will vary across contexts. However, we note two generally applicable considerations. First, protection from arbitrary dismissal will be critical to ensuring that internal gatekeepers are autonomous enough to perform their functions. The short history of AI ethics boards is littered with examples of limited independence and institutional insecurity: the dissolution of Twitter's Trust and Safety Council after a change in ownership (Zakrzewski et al., 2022); the mass resignation of Axon's AI and Policing Technologies Ethics Board after the company failed to consult it on plans to develop taser-equipped and surveillance drones for schools (Friedman et al., 2022). Google's AI ethics board was dismantled (and remains so) after employees challenged the appointment of certain members (Piper, 2019). Internal ethics and safety functions are vulnerable to culls (De Vynck & Oremus, 2023; eSafety Commissioner, 2024) or being bypassed (Roose, 2024). Companies are liable to abandon measured strategies amid perceived "racing" dynamics (see Google's response to ChatGPT3.5's release). Power users who agitate against company policy can be blocked or deposed by the platform; see, for example, Reddit's decision to remove moderators from administrative roles after they staged "blackouts" (ie. turned subreddits private) to protest changes to the company's API pricing policy (Carlon & Leeftink, 2023). Taking inspiration from other regulatory fields, the tenure of critical gatekeepers could be better secured by statutorily mandated positions (in manner of data protection officers under the GDPR), vesting appointment powers in shareholders (as some company law regimes do for external auditors) or independent committees, and whistleblower protections (Bottomley, 2007; Schuett et al., 2024).

Second, safeguards against conflicts of interest are needed to protect gatekeeper independence. Under the "pure" separation of powers doctrine, conflicts are mitigated by barring members of one branch from membership of another (Vile, 1998). This principle is expressed in various regulations targeting private power. For example, restrictions on firms providing both financial audit and non-audit services to clients are a common feature of company laws and professional accounting standards. The burgeoning AI audit and risk assessment industry would benefit from similar safeguards to inculcate a culture of independence. The principle already informs aspects of the EU's data sharing framework. Under the *Data Governance Act,* data intermediaries are prohibited from processing data for any purpose other than providing intermediation services. Any additional data-driven services can only be performed through a separate legal entity (Regulation 2022/868, art.

12). Recognising that vertical integration could create competing incentives for data intermediaries, the Act deploys an old instrument of competition regulation – structural separation. Traditionally conceived to protect contestability in markets, competition law restrictions and remedies may also support the goals of public contestability by maintaining divisions of power and interests in supply chains. More so than market contestability, public contestability provides stakeholders with an alternative to "exit" to convey their dissatisfaction – "voice" (Bottomley, 2007). In recent years, upstream technology suppliers have played critical roles in pressuring problem actors to respond to public concerns; see, for example, when Apple, Amazon, and Google converged to remove social media app Parler from its platforms following its role in the coordination of the January 6, 2021 insurrection (K. Lyons, 2021).[10] Measures to limit vertical integration in AI supply chains could not only carry competition benefits but multiply targets and levers for public pressure.

3.3. Independent and alternative sources of information

Access to independent and alternative sources of information is a necessary condition for public contestability (Dahl, 2005; Pettit, 2000). Without a basis to evaluate policy and conduct, channels for collective action and checks on decision making are of little utility. Governance arrangements which cast AI firms as the sole narrators of the operation, effects, and externalities of their systems have significant implications for contestability. If "a single party, faction or interest" has a monopoly over all important information about a policy or system, the ability of stakeholders to independently evaluate and critique will be severely diminished (Dahl, 2005). Access to alternative and independent sources means opening AI systems up to independent examination and verification (Züger & Asghari, 2023), rather than allowing private firms and regulatory agencies to control the content and flow of information used to evaluate their conduct.

Transparency is a consistent if amorphous goal of platform and AI regulation across jurisdictions. At the individual-level, the regulatory emphasis has been on mechanisms which support informed interactions with technology firms and their products: data collection notices, data subject access requests, rights to explanation of automated decisions, and transparency about when individuals are interacting with AI. Various stakeholders have sounded calls for greater system-level transparency over time as well (Kaye, 2018; Access Now et al., 2021). External

10. On 6 January 2021, over 2,000 people stormed the US Capitol (the seat of the US Congress) following a rally led by then-President Donald Trump to protest the ratification of the 2020 US Presidential election results, which Trump had lost (Barrett et al, 2021).

pressure has yielded incremental improvements in the transparency practices of some technology firms, leading to publicly accessible ad databases and aggregate reporting on content moderation.

Despite progress, there is still a risk of firms using vague commitments to transparency to resist meaningful scrutiny and accountability (Suzor et al., 2019). One cause for concern is the discretion firms retain over what to disclose and how. The choices firms make about the type and format of information available can shape and divert scrutiny (Cohen, 2023). For example, there is evidence that the composition of political ad archives (which various platforms have maintained since 2018) has influenced the focus of journalistic coverage during elections (Dommett & Bakir, 2020; Leerssen et al., 2023). A further concern relates to discretion over access. Third party efforts to produce independent accounts have, at times, been frustrated by platform actions to block researchers (Merrill & Tobin, 2019; Vincent, 2021).

A shift from voluntary to statutory-backed transparency obligations in some jurisdictions has reduced the scope for discretion. However, heavy reliance on self-reporting remains a feature, including in the nascent certification and audit requirements which are emerging as a cornerstone of AI regulatory frameworks (see, e.g. Bill-C27; Regulation 2024/1689). The growing array of tools for AI ethics auditing are also predominantly built for use by actors involved in product development and not external stakeholders (such as users and clients) (Ayling & Chapman, 2022). Self-assessment and reporting as a preferred approach is often rationalised by internal concentrations of expertise and respect for intellectual property rights and trade secrecy. While concessions about what is "proprietary" should be open to public interrogation (Mendonca et al., 2023), some reliance on internal record-keeping and self-reporting to realise transparency objectives is unavoidable. External scrutineers across regulatory fields and industries must often rely on an internal function for information. For example, external auditors rely on internal audit functions to evaluate financial statements. Central banks and prudential regulators rely on modelling by banks to conduct bottom-up stress tests. Sectoral regulators rely on incident self-reporting across many industries (e.g. aviation, nuclear, medicine). Experience in these regulatory fields hold valuable insights and strategies for lawmakers and institutional designers about how to secure the production of independent information (Raji et al., 2020, 2022).

3.3.1. Auditor independence

First, auditor independence will be important. Drawing on lessons from financial

auditing scandals at the turn of last century, to reduce conflicts of interest in the burgeoning AI audit industry, it may be appropriate to segregate the provision of audit from other services (see section 3.2) and mandate periodic rotation. Instituting non-profit or public bodies to oversee AI audits with public interest implications may further promote independence (Falco et al., 2021; Raji et al., 2022). Requiring conformity assessments and audit reports to be openly published (subject to organisational rights to seek redactions on restricted grounds) would not only subject the findings but also the methods of scrutiny to third party evaluation.

3.3.2. Public incident reporting

Second, public incident reporting systems could supply an alternative source of data to help regulators and public interest groups identify systemic problems for attention (Raji et al., 2022). Arrangements under which firms self-report incidents, such as data breach notification schemes, can assist in this regard but are susceptible to under-reporting, especially where the thresholds that trigger disclosure obligations are contested. Data from public incident reporting can assist in triangulating and validating system impacts. Civil society and academic-led initiatives which rely on citizen reporting and data donation, such as the *Tracking Global Online Censorship* and the *Australian Ad Observatory*, offer inspiration for how public incident reporting of AI outputs might work. Forums for incident reporting can also help groups coordinate and mobilise around shared experience. Online forums for sharing information, experiences and tactics for resistance have been key to collective action by platform workers (Cini, 2023). *Turkopticon*, a website where Amazon's Mechanical Turk workers can review bad requests, whose organisers also advocate for improved conditions, is one example. Statutory frameworks and/or public funding could help with the institution and maintenance of similar incident databases and forums for, say, large generative-AI applications.

3.3.3. Freedom of information

Third, freedom-of-information ("FOI")-like mechanisms can also supplement gaps in public databases. Technology firms are accustomed to information access rights for individuals in their capacity as data subjects (i.e. data subject access request), which have previously been harnessed to uncover systemic issues (Case C-362/14). FOI on public interest grounds is a newer concept for the sector. However, Article 40(4) of the EU's Digital Services Act (which contains conditional rights for researchers to access information to investigate systemic risk) and public rights to access government documents, offer precedents for consideration. Company objections to FOI requests on the basis of IP and trade secrecy concerns should not be readily conceded. While there are limits on access to "sensitive" information in

public sector contexts (for e.g. embargoes on cabinet documents), the limits are transparent and publicly contestable. Random audits of the administration of FOI systems by firms could provide an effective external check to verify (and incentivise) compliance.

4. Conclusion

Governance is a broad concept; many different, dispersed actors impact, in a contingent and complex way, how societies are regulated (Black, 2001). Commercial AI firms face increasing scrutiny of the social impacts emerging from the design of their products, their policies governing access and use, and their commercial practices. In response, AI firms are seeking to shore up their perceived legitimacy – their social license – by employing democratic rhetoric to demonstrate "alignment" between their operations and the "public interest". At the same time, state regulators are looking for effective mechanisms to improve self-regulatory practices within and across industry participants. In this article, we consider the role of contestability as a fundamental component of democratic legitimacy that has, we think, been relatively under-emphasised in debates over AI governance to date. We provide some early reflections that we hope might be useful to regulators, public interest groups, and those companies that are genuinely seeking to experiment with more effective, thicker conceptions of democratic participation in AI governance.

Democratic contestability is a critical feature of governance systems that are responsive to the public interest. Mechanisms that allow many, diverse stakeholders to collectively articulate contested visions of public interests, assess and explain the causally complex aggregate impacts of choices by infrastructure providers, and challenge policies and decisions are important guardrails against erroneous, exclusionary, and arbitrary decision-making. We have argued that efforts to incorporate public input and legitimate corporate AI governance through "democratisation" will benefit substantially from strengthening capabilities for public contestation. We have argued that those capabilities include institutionalised opportunities for groups to contest AI deployment, power separation, and access to independent sources of information. These conditions are both interdependent and mutually reinforcing. Governance structures that distribute power amongst multiple, semi-autonomous gatekeepers increase targets for pressure by public interest groups, which in turn encourage intra-branch scrutiny and competition. Independent information provides a basis for public stakeholders to evaluate the conduct of commercial AI firms. It fortifies the independence of semi-autonomous gatekeepers

like auditors, which is necessary for a functioning system of checks and balances. Those gatekeepers, in turn, are able to contribute independent and alternative information used to scrutinise and render account for AI behaviour.

Adopting democratic contestability as a conceptual lens helps shed new light on the possible evolution of and relationships between existing ideas about how to govern AI, including ethics boards and data cooperatives, regulatory oversight and public interest litigation, and AI audit and professional licensing schemes. We have attempted in this article to sketch out tentative suggestions for how the conditions for contestability might be better met in practice. At this stage, we suggest that further experimentation is key, and further research will be required to better understand conditions for success in creating productive and beneficial tension between public and commercial interests. As a starting point, we hope that this initial reflection can help inform more specific demands from civil society and public authorities for firms to foster more effective avenues for more diverse contestation. If nothing else, we hope that advocates might use this often neglected but fundamental component of democratic theory to hold AI firms more directly to account to deliver on their democratic rhetoric. For states looking to encourage legitimate and effective roles for self-governance within the rapidly changing AI industry, we hope that the suggestions above provide inspiration for more detailed and practical regulatory measures that might build better opportunities for public participation in AI governance. As new experiments are deployed, further research on routes to implementation and measures of effectiveness and desirability of different institutional approaches to contestability in different contexts will be needed. There is still a relatively large conceptual gap to translate and adapt the insights of democratic theory to the polycentric webs of public and private regulatory stakeholders involved in the governance of emerging technologies. So far, the democratic experiments undertaken by AI firms have largely been informed by deliberative democracy research, a field which has devoted significant attention to translating theory into practice. Bringing democratic contestability into AI governance will require researchers to translate theory and lessons from public governance into guidance for exploration in real world settings.

References

Access Now, ACLU Foundation of Northern California, ACLU Foundation of Southern California, ARTICLE 19, Brennan Center for Justice, Center for Democracy & Technology, Electronic Frontier Foundation, Global Partners Digital, InternetLab, National Coalition Against Censorship, New America's Open Technology Institute, Ranking Digital Rights, Red en Defensa de los Derechos

Digitales, & WITNESS. (2021). *The Santa Clara principles on transparency and accountability in content moderation* [Version 2.0]. https://santaclaraprinciples.org/

Ada Lovelace Institute & UK AI Council. (2021). *Exploring legal mechanisms for data stewardship* (Data for the public good) [Final report]. https://www.adalovelaceinstitute.org/project/legal-mechanisms-for-data-stewardship/

Aggarwal, D., Eldar, O., Hochberg, Y. V., & Litov, L. P. (2020). The rise of dual-class stock IPOs. *Journal of Financial Economics*, *144*(1), 122–153. https://doi.org/10.1016/j.jfineco.2021.12.012

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2022). Contestable AI by design: Towards a framework. *Minds and Machines*, *33*, 613–639. https://doi.org/10.1007/s11023-022-09611-z

Anthropic. (October 17, 2023a). *Collective constitutional AI: Aligning a language model with public input* [Report]. https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

Anthropic. (September 19, 2023a). *The long-term benefit trust* [Announcement]. https://www.anthropic.com/news/the-long-term-benefit-trust

Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Institute of Planners*, *35*(4), 216–224. https://doi.org/10.1080/01944366908977225

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics*, *2*, 405–429. https://doi.org/10.1007/s43681-021-00084-x

Bächtiger, A., Dryzek, J. S., Mansbridge, J., & Warren, M. (2018). Deliberative democracy: An introduction. In A. Bächtiger, J. S. Dryzek, J. Mansbridge, & M. Warren (Eds.), *The Oxford handbook of deliberative democracy* (pp. 1–32). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198747369.013.50

Bai, H., Voelkel, J. G., Eichstaedt, J. C., & Willer, R. (2023). *Artificial intelligence can persuade humans on political issues*. Open Science Framework Preprints. https://doi.org/10.31219/osf.io/stakv

Banerjee, S. B. (2022). Decolonizing deliberative democracy: Perspectives from below. *Journal of Business Ethics*, *181*, 283–299. https://doi.org/10.1007/s10551-021-04971-5

Barrett, T., Raju, M., & Nickeas, P. (2021, January 7). US Capitol secured, 4 dead after rioters stormed the halls of Congress to block Biden's win. *CNN*. https://edition.cnn.com/2021/01/06/politics/us-capitol-lockdown/index.html

Benthall, S., & Goldenfein, J. (2021). Artificial intelligence and the purpose of social systems. *Proceedings of the 2021 AAAI/ACM Conference on AI Ethics and Society*. AIES '21, New York, NY, USA. https://doi.org/10.1145/3461702.3462526

Bill C-18. (2023). *An Act respecting online communications platforms that make news content available to persons in Canada (Online News Act S.C.)*. 1st Session, 44th Parliament, Canada. https://laws-lois.justice.gc.ca/PDF/2023_23.pdf

Bill-C27. (2022). *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*. 1st Session, 44th Parliament, Canada. https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI. *Proceedings of the 2nd ACM*

Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–8. https://doi.org/10.1145/3551624.3555290

Black, J. (2001). Decentering regulation: Understanding the role of regulation and self-regulation in a "post-regulatory" world. *Current Legal Problems*, *54*(1), 103–146. https://doi.org/10.1093/clp/54.1.103

Bottomley, S. (2007). *The constitutional corporation: Rethinking corporate governance* (1st ed.). Routledge. https://doi.org/10.4324/9781315614984

Braithwaite, J. (1997). On speaking softly and carrying big sticks: Neglected dimensions of a Republication separation of powers. *The University of Toronto Law Journal*, *47*(3), 305–361. https://doi.org/10.2307/825973

Brockman, G., Sutskever, I., & Open AI. (2015). *Introducing OpenAI* [Announcement]. https://openai.com/blog/introducing-openai

Brockman, G., Sutskever, I., & OpenAI. (2019). *OpenAI LP* [Announcement]. https://openai.com/blog/openai-lp

Caplan, R. (2023). Networked platform governance: The construction of the democratic platform. *International Journal of Communication*, *17*, 3451–3472. https://ijoc.org/index.php/ijoc/article/view/20035

Carlon, D., & Leeftink, D. (2023). *Burning the hooks: Wwhat happens when we lose our subreddits, APIs and exchanges?* (Automated Decision-Making and Society) [Opinion piece]. Medium. https://medium.com/automated-decision-making-and-society/burning-the-hooks-what-happens-when-we-lose-our-subreddits-apis-and-exchanges-1b72bd3aa49f

Case C-362/14. (n.d.). *Judgment of the Court (Grand Chamber) of 6 October 2015. Maximillian Schrems v Data Protection Commissioner*. The Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:62014CJ0362

Caulfield, M., & Lynn, A. (2024). Federated corporate social responsibility: Constraining the responsible corporation. *Academy of Management Review*, *49*(1), 32–55. https://doi.org/10.5465/amr.2020.0273

Cini, L. (2023). Resisting algorithmic control: Understanding the rise and variety of platform worker mobilisations. *New Technology, Work and Employment*, *38*(1), 125–144. https://doi.org/10.1111/ntwe.12257

Clegg, N. (2023, June 22). *Bringing people together to inform decision-making on generative AI*. Meta Newsroom. https://about.fb.com/news/2023/06/generative-ai-community-forum/

Clement, A. (2023). *AIDA's "consultation theatre" highlights flaws in a so-called agile approach to AI governance*. Centre for International Governance Innovation. https://www.cigionline.org/articles/aidas-consultation-theatre-highlights-flaws-in-a-so-called-agile-approach-to-ai-governance/

Cohen, T. (2023). *The datafied polity: Voter privacy in the age of data-driven political campaigning* [Doctoral dissertation, Queensland University of Technology]. https://doi.org/10.5204/thesis.eprints.242466

Crawford, K., & Lumby, C. (2013). Networks of governance: Users, platforms, and the challenges of networked media regulation. *International Journal of Technology Policy and Law*, *1*(3), 270–282. https://doi.org/10.1504/IJTPL.2013.057008

Dahl, R. (1971). *Polyarchy: Participation and opposition*. Yale University Press.

Dahl, R. A. (2005). What political institutions does large-scale democracy require? *Political Science Quarterly*, *120*(2), 187–197. https://doi.org/10.1002/j.1538-165X.2005.tb00543.x

Dave, P. (2023, November 19). How OpenAI's bizarre structure gave 4 people the power to fire Sam Altman. *Wired*. https://www.wired.com/story/openai-bizarre-structure-4-people-the-power-to-fire-sam-altman/

De Vynck, G., & Oremus, W. (2023, March 30). As AI booms, tech firms are laying off their ethicists. *The Washington Post*. https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023). The participatory turn in AI design: Theoretical foundations and the current state of practice. *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23. https://doi.org/10.1145/3617694.3623261

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. *Findings of the Association for Computational Linguistics*, 1236–1270. https://doi.org/10.18653/v1/2023.findings-emnlp.88

Dommett, K., & Bakir, M. E. (2020). A transparent digital election campaign? The insights and significance of political advertising archives for debates on electoral regulation. *Parliamentary Affairs*, *73*(Supplement_1), 208–224. https://doi.org/10.1093/pa/gsaa029

eSafety Commissioner. (2024). *Basic online safety expectations: Summary of response from X Corp. (Twitter) to eSafety's transparency notice on online hate* [Report]. Australian Government. https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C., Jirotka, M., Johnson, H., LaPointe, C., Llorens, A. J., Mackworth, A. K., Maple, C., Pálsson, S. E., Pasquale, F., Winfield, A., & Yeong, Z. K. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence*, *3*, 566–571. https://doi.org/10.1038/s42256-021-00370-7

Feintuck, M. (2004). *"The public interest" in regulation*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199269020.001.0001

Flew, T., Gillett, R., Martin, F., & Sunman, L. (2021). Return of the regulatory state: A stakeholder analysis of Australia's Digital Platforms Inquiry and online news policy. *The Information Society*, *37*(2), 128–145. https://doi.org/10.1080/01972243.2020.1870597

Fraser, H., & Suzor, N. (In press). Locating fault and responsibility for AI harms: A systems theory of foreseeability, reasonable care and causal responsibility in the AI value chain. *Law, Innovation and Technology*. https://eprints.qut.edu.au/251116/

Friedman, B., Abd-Almageed, W., Brundage, M., Calo, R., Citron, D., Delsol, R., Harris, C., Lynch, J., & McBridge, M. (2022). *Statement of resigning Axon AI ethics board members* [Statement]. The Policing Project. https://www.policingproject.org/statement-of-resigning-axon-ai-ethics-board-members

Galvagna, C. (2023). *Inclusive AI governance. Civil society participation in standards development* (The future of regulation) [Discussion paper]. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/

Geist, M. (2021, February 12). *Afraid to lead: Canadian government launches timid consultation on*

implementing copyright term extension. https://www.michaelgeist.ca/2021/02/afraid-to-lead/

Gleeson, J. (2016). '(Not) working 9–5': The consequences of contemporary Australian-based online feminist campaigns as digital labour. *Media International Australia*, *161*(1), 77–85. https://doi.org/10.1177/1329878X16664999

Gornet, M., & Maxwell, W. (2024). The European approach to regulating AI through technical standards. *Internet Policy Review*, *13*(3). https://doi.org/10.14763/2024.3.1784

Gorwa, R. (2024). *The politics of platform regulation: How governments shape online content moderation*. Oxford University Press. https://doi.org/10.1093/oso/9780197692851.001.0001

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.

Griffin, R. (2023). Public and private power in social media governance: Multistakeholderism, the rule of law and democratic accountability. *Transnational Legal Theory*, *14*(1), 46–89. https://doi.org/10.1080/20414005.2023.2203538

Groves, L., Peppin, A., Strait, A., & Brennan, J. (2023). *Going public: The role of public participation approaches in commercial AI labs* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2306.09871

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C., & Franklin, M. (2023). A proposal for a definition of general purpose artificial intelligence systems. *Digital Society*, *2*, Article 36. https://doi.org/10.1007/s44206-023-00068-w

Halpern, D., & Costa, E. (2022, September 20). How can citizens shape the future of social media platforms? *The Behavioural Insights Team*. https://www.bi.team/blogs/how-can-citizens-shape-the-future-of-social-media-platforms/

Henin, C., & Le Métayer, D. (2022). Beyond explainability: Justifiability and contestability of algorithmic decision systems. *AI & Society*, *37*(4), 1397–1410. https://doi.org/10.1007/s00146-021-01251-8

Hickok, M. (2024). Public procurement of artificial intelligence systems: New risks and future proofing. *AI & Society*, *39*(3), 1213–1227. https://doi.org/10.1007/s00146-022-01572-2

Hildebrandt, M. (2019). Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law*, *20*(1), 83–121. https://doi.org/10.1515/til-2019-0004

Hirsch, T., Merced, K., Narayanan, S., Imel, Z. E., & Atkins, D. C. (2017). Designing contestability: Interaction design, machine learning, and mental health. *Proceedings of the 2017 Conference on Designing Interactive Systems*, 95–99. https://doi.org/10.1145/3064663.3064703

Hong, S., & You, J. (2018). Limits of regulatory responsiveness: Democratic credentials of responsive regulation. *Regulation & Governance*, *12*(3), 413–427. https://doi.org/10.1111/rego.12193

Institute on Governance. (n.d.). *What is governance*. https://web.archive.org/web/20240213114928/https://iog.ca/what-is-governance/

Kaye, D. (2018). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (Report No. A/HRC/38/35). United Nations General Assembly. http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35

Kellogg, K., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kluttz, D. N., Kohli, N., & Mulligan, D. K. (2020). Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In K. Werbach (Ed.), *After the digital tornado: Networks, algorithms, humanity* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108610018

Lalley, S. P., & Weyl, E. G. (2018). Quadratic voting: How mechanism design can radicalize democracy. *AEA Papers and Proceedings*, *108*, 33–37. https://doi.org/10.1257/pandp.20181002

Larcker, D. F., & Tayan, B. (2017). *Governance gone wild: Epic misbehavior at Uber technologies* (Stanford Closer Look Series) [Report]. Corporate Governance Research Initiative. https://www.gsb.stanford.edu/faculty-research/publications/governance-gone-wild-epic-misbehavior-uber-technologies

Lee, P. (2016, March 25). Learning from Tay's introduction. *Official Microsoft Blog*. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Leerssen, P., Dobber, T., Helberger, N., & de Vreese, C. (2023). News from the ad archive: How journalists use the Facebook Ad Library to hold online advertising accountable. *Information, Communication & Society*, *26*(7), 1381–1400. https://doi.org/10.1080/1369118X.2021.2009002

Lenz, J. J. A. (2020). Privacy class actions' unfulfilled promise. In I. N. Cofone (Ed.), *Class actions in privacy law*. Routledge. https://doi.org/10.4324/9781003080510

LGPD. (2018). *Lei Geral de Proteção de Dados Pessoais [General Personal Data Protection Act]* (No. Law No. 13.709). Government of Brazil. https://lgpd-brazil.info/

Lloyd v Google LLC (No. UKSC 50). (2021). The Supreme Court of the United Kingdom. https://www.supremecourt.uk/cases/uksc-2019-0213.html

Lyons, H., Velloso, E., & Miller, T. (2021a). Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–25. https://doi.org/10.1145/3449180

Lyons, H., Velloso, E., & Miller, T. (2021b, a). Designing for contestation: Insights from administrative law. *Proceedings of 2019 CSCW Workshop on Contestability in Algorithmic Systems*. Contestability'19, New York, NY, USA. https://doi.org/10.48550/ARXIV.2102.04559

Lyons, K. (2021, May 17). Parler returns to Apple App Store with some content excluded. *The Verge*. https://www.theverge.com/2021/5/17/22440005/parler-apple-app-store-return-amazon-google-capitol

Mannell, K., & Smith, E. T. (2022). Alternative social media and the complexities of a more participatory culture: A view from Scuttlebutt. *Social Media + Society*, *8*(3), 1–11. https://doi.org/10.1177/20563051221122448

Mendonca, R. F., Almeida, V., & Filgueiras, F. (2023). *Algorithmic institutionalism: The changing rules of social and political life*. Oxford University Press. https://doi.org/10.1093/oso/9780192870070.001.0001

Merrill, J. B., & Tobin, A. (2019, January 28). Facebook moves to block ad transparency tools—Including ours. *ProPublica*. https://www.propublica.org/article/facebook-blocks-ad-transparency-tools

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

Mouffe, C. (1999). Deliberative democracy or agonistic pluralism? *Social Research*, *66*(3), 745–758. http://www.jstor.org/stable/40971349

Myers West, S. (2017). Raging against the machine: Network gatekeeping and collective action on social media platforms. *Media and Communication*, *5*(3), 28–36. https://doi.org/10.17645/mac.v5i3.989

Online Safety Act. (2021). *An Act relating to online safety for Australians, and for other purposes* (No. Cth). Government of Australia. https://classic.austlii.edu.au/au/legis/cth/consol_act/osa2021154/

OpenAI. (2018). *OpenAI charter* [Charter]. https://openai.com/charter

OpenAI. (2024). *Review completed & Altman, Brockman to continue to lead OpenAI* [Press release]. https://openai.com/index/review-completed-altman-brockman-to-continue-to-lead-openai/

Ovadya, A. (2023). *'Generative CI' through collective response systems* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2302.00672

Palladino, N. (2023). A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices. *Telecommunications Policy*, *47*(5). https://doi.org/10.1016/j.telpol.2022.102479

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pateman, C. (2012). Participatory democracy revisited. *Perspectives on Politics*, *10*(1), 7–19. https://doi.org/10.1017/S1537592711004877

Paul, K. (2023, July 26). The rebel group stopping self-driving cars in San Francisco – One cone at a time. *The Guardian*. https://www.theguardian.com/us-news/2023/jul/26/san-francisco-stop-self-driving-cars-traffic-cone-safe-street-rebel

Perarnaud, C. (2023, April 25). With the AI Act, we need to mind the standards gap. *CEPS Brief*. https://www.ceps.eu/with-the-ai-act-we-need-to-mind-the-standards-gap/

Pettit, P. (1999). Republican freedom and contestatory democratization. In I. Shapiro & C. Hacker-Cordón (Eds.), *Democracy's values* (pp. 163–190). Cambridge University Press.

Pettit, P. (2000). Democracy, electoral and contestatory. In I. Shapiro & S. Macedo (Eds.), *Designing democratic institutions Nomos XLII* (pp. 105–144). New York University Press. https://doi.org/10.18574/nyu/9780814786628.003.0009

Piper, K. (2019, April 5). Exclusive: Google cancels AI ethics board in response to outcry. *Vox*. https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board

RadicalXChange. (n.d.). *Quadratic voting*. RadicalXChange Wiki. https://www.radicalxchange.org/wiki/quadratic-voting/#tools

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider oversight: Designing a third party audit ecosystem for AI governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571. https://doi.org/10.1145/3514094.3534181

Regulation 2016/679. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. European Parliament and Council. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Regulation 2022/868. (2022). *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)*. European Parliament and Council. https://eur-lex.europa.eu/eli/reg/2022/868/oj

Regulation 2022/2065. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. European Parliament and Council. https://eur-lex.europa.eu/eli/reg/2022/2065/oj

Regulation 2024/1689. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. European Parliament and Council. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Rollo, T. (2017). Everyday deeds: Enactive protest, exit, and silence in deliberative systems. *Political Theory*, *45*(5), 587–609. https://doi.org/10.1177/0090591716661222

Roose, K. (2022, October 21). A coming-out party for generative A.I., Silicon Valley's new craze. *The New York Times*. https://www.nytimes.com/2022/10/21/technology/generative-ai.html

Roose, K. (2024, June 5). OpenAI insiders warn of a 'reckless' race for dominance. *The New York Times*. https://www.nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html

Sanders, L. M. (1997). Against deliberation. *Political Theory*, *25*(3), 347–376. https://doi.org/10.1177/0090591797025003002

Schrage, E. (2012, December 11). *Our site governance vote* [Post]. Facebook. http://web.archive.org/web/20121215162151/https://www.facebook.com/notes/facebook-site-governance/our-site-governance-vote/10152304778295301

Schuett, J., Reuel, A.-K., & Carlier, A. (2024). How to design an AI ethics board. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00409-y

Schwartz, O. (2024, January 4). In 2016, Microsoft's racist chatbot revealed the dangers of online conversation. *IEEE Spectrum*. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

Seger, E., Ovadya, A., Garfinkel, B., Siddarth, D., & Dafoe, A. (2023). *Democratising AI: Multiple meanings, goals, and methods*. arXiv. https://doi.org/10.48550/arXiv.2303.12642

Sharma, C. (2023, December 12). Setting a higher bar: Professionalizing AI engineering. *Lawfare*. https://www.lawfaremedia.org/article/setting-a-higher-bar-professionalizing-ai-engineering

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). *Participation is not a design fix for machine learning* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2007.02423

Surman, M. (2023, March 22). Introducing Mozilla.ai: Investing in trustworthy AI. *The Mozilla Blog*. ht tps://blog.mozilla.org/en/mozilla/introducing-mozilla-ai-investing-in-trustworthy-ai/

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society*, *4*(3), 205630511878781. https://doi.org/10.1177/2 056305118787812

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press. https://doi.org/10.1017/9781108666428

Suzor, N., & Gillett, R. (2022). Self-regulation and discretion. In T. Flew & F. R. Martin (Eds.), *Digital platform regulation: Global perspectives on internet governance* (pp. 259–279). Springer. https://eprint s.qut.edu.au/227786/

Suzor, N. P., Myers West, S., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 1526–1543. https://ijoc.org/index.php/ijoc/article/view/9736/0

The Royal Commission into the Robodebt Scheme. (2023). *Royal commission into the robodebt scheme* [Final report]. https://robodebt.royalcommission.gov.au/publications/report

Thomas, E. (2019, February 1). How to hack your face to dodge the rise of facial recognition tech. *Wired*. https://www.wired.co.uk/article/avoid-facial-recognition-software

Trantidis, A. (2017). Is government contestability an integral part of the definition of democracy? *Politics*, *37*(1), 67–81. https://doi.org/10.1177/0263395715619635

Treasury Laws Amendment (News Media and Digital Platforms Mandatory Bargaining Code). (2021). *An Act to amend the Competition and Consumer Act 2010 in relation to digital platforms, and for related purposes* (No. Cth). Government of Australia. https://classic.austlii.edu.au/au/legis/cth/num_act/tlam adpmbca2021734/

Tseng, Y.-S. (2023). Rethinking gamified democracy as frictional: A comparative examination of the Decide Madrid and vTaiwan platforms. *Social & Cultural Geography*, *24*(8), 1324–1341. https://doi.or g/10.1080/14649365.2022.2055779

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press. https://doi.org/10.25969/mediarep/14848

Urbinati, N. (2010). Unpolitical democracy. *Political Theory*, *38*(1), 65–92. https://doi.org/10.1177/00 90591709348188

Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). 'At the end of the day Facebook does what it wants': How users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–22. https://doi.org/10.1145/3415238

Vaccaro, K., Xiao, Z., Hamilton, K., & Karahalios, K. (2021). Contestability for ccontent moderation. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–28. https://doi.org/10.1145/34 76059

Vile, M. J. C. (1998). *Constitutionalism and the separation of powers*. Liberty Fund.

Vincent, J. (2021, August 4). Facebook bans academics who researched ad transparency and misinformation on Facebook. *The Verge*. https://www.theverge.com/2021/8/4/22609020/facebook-b ans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in

Waldron, J. (2013). Separation of powers in thought and practice? *Boston College Law Review*, *54*(2), 433–468. https://bclawreview.bc.edu/articles/702

Weise, K., & Metz, C. (2023, May 9). When A.I. chatbots hallucinate. *The New York Times*. https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html

Young, I. M. (2001). Activist challenges to deliberative democracy. *Political Theory*, *29*(5), 670–690. https://doi.org/10.1177/0090591701029005004

Young, M., Ehsan, U., Singh, R., Tafesse, E., Gilman, M., Harrington, C., & Metcalf, J. (2024). Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday*, *29*(4). https://doi.org/10.5210/fm.v29i4.13642

Zakrzewski, C., Menn, J., & Nix, N. (2022, December 12). Twitter dissolves Trust and Safety Council. *The Washington Post*. https://www.washingtonpost.com/technology/2022/12/12/musk-twitter-harass-yoel-roth/

Zaremba, W., Dhar, A., Ahmad, L., Eloundou, T., Santurkar, S., Agarwal, S., & Leung, J. (2023, March 25). *Democratic inputs to AI*. OpenAI. https://openai.com/index/democratic-inputs-to-ai/

Zhang, S., Diab, M., & Zettlemoyer, L. (2022, May 3). Democratizing access to large-scale language models with OPT-175B [Research report]. *Meta AI Blog*. https://ai.meta.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/

Züger, T., & Asghari, H. (2023). AI for the public. How public interest theory shifts the discourse on AI. *AI & Society*, *38*(2), 815–828. https://doi.org/10.1007/s00146-022-01480-5

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies