PLATFORM Journal of Media and Communication

Vol 10.1

Digital Governance for a Human-Centred Society

PLATFORM

Journal of Media and communication

Vol 10.1

Digital Governance for a human-centred society

ISSN 1836-5132 online

Platform: Journal of Media and Communication Volume 10 has been edited by graduate students at The University of Melbourne. It is funded and published by the School of Culture and Communication, The University of Melbourne

Table of Contents

Editorial

Reimagining Digital Governance for a Human-Centred Society

Brittany Craig, Iván Kirschbaum, Jingxian You

Over the past years, digital technologies have significantly transformed how information flows across online spaces. The pervasiveness of digital communication technologies in many users' everyday lives, together with the rising power of giant platform companies, has raised concerns about digital governance. A growing body of literature recognises that non-state actors, such as Google, Apple, and Facebook, are becoming 'the new governors' who have 'mediated', 'constituted', and 'moderated' public discourse (Klonick, 2018; Gilllespie, 2018). From single applications to the emergence of 'super apps', the expansion of digital, data-driven platform economy has subtly shifted conventional national based regulatory practices towards a more global phenomenon.

In digital communication studies, an increasing amount of research pays attention to the practices and debates surrounding how globalising technologies should be regulated (Flew et al., 2019; Gillespie et al., 2020). The growing global 'techlash' – marked by strong resistance to and rising scrutiny of the negative impacts associated with giant technology companies – alongside the global nature of digital communication technologies, has influenced not only macro-level international digital regulatory practices but also micro-level interactions between individual users and technologies. Consequently, more studies have sought to identify the multiple discursive dimensions of digital governance. Platform and app scholarship, for instance, has examined major global platform companies' influences on content moderation (Gillespie et al., 2020; Gorwa et al., 2020), 'super app' conglomeration (van der vlist et al., 2024), and the acceleration of uneven global flows of digital capital (Nieborg et al., 2020; Joseph et al., 2023).

The rapid expansion of Artificial Intelligence (AI) in the 2020s, coupled with its perceived contributions to productivity and economic development, has been accompanied by escalating concerns about algorithmic bias, data privacy, online security, and public trust (Flew, 2024; Nah et al., 2024; Sahebi & Formosa, 2025). The increasing deployment of AI in diverse contexts has heightened the demand for more comprehensive digital and data regulation of AI technologies. AI governance, therefore, has become a focal point of attention across academic, industrial, and political spheres. Intergovernmental policy agendas, for example, have underlined 'responsible and human-centric AI' and the protection of human rights, as reflected in the updated OECD AI Principles and the European AI Act (OECD, 2024; EU, 2024).

One pressing issue within AI governance, however, is the lack of consensus on the ethical framework guiding AI regulation. How could we understand the changing relations between technology, human, and the natural environment in the context of AI? What different approaches to AI ethics might reshape our notions of 'justice' and 'fairness'? These are questions explored in the first article of this Special Issue. In *A Virtuous Ethics of AI: Conviviality as a Regulatory Framework*, Gavin Duffy examines John Rawls' theory of 'justice as fairness' and Ivan Illich's notion of 'conviviality' as applied to AI

regulation. Duffy argues that a 'convivial' perspective on AI offers a more sustainable regulatory approach for a human-centred society. Another central concern in contemporary AI discourse is how ordinary users regulate their informational privacy when encountering automated systems. In the following article, "*I'm a bit cautious of jumping in with both feet': exploring information ownership and negotiated control in AI chatbot users' communication privacy management,"* Mark Bo Chen illustrates how users negotiate information ownership, boundary regulation, and control when interacting with AI chatbots.

Shifting to the intersection of AI chatbots and urban digital governance, Juan Martín Marinangeli focuses on the AI chatbot Boti promoted by the Government of the City of Buenos Aires in Argentina. In the third article, *Coding Trust: The Promise and Perils of Digital Transformation in Buenos Aires' AI Governance*, he discusses how public trust is constructed – and concealed – through the official AI chatbot. The final article of this Special Issue brings us to postdigital arts practices in Europe. Focusing on an increasing technological opacity, Etienne Malecki in *Postdigital Art & Privacy: In Search of a Sensible Experience of Technology* reveals how a group of European multimedia artists engage with the politics of technology and challenge surveillance norms and digital control.

Overall, the articles collected in this Special Issue tap into discussions on how digital governance might be reimagined for a more human-centred, responsible, and trustworthy technological future. They provoke questions regarding the norms, values, and power asymmetries embedded in today's complex and globalised digital environment. Taken together, these contributions shed light on the complex power structures that shape digital infrastructures – from global platform firms to municipal AI initiatives and artistic participations – and demonstrate how such structures influence individuals' everyday interactions with advanced technologies. Connecting these works is a shared concern with how societies might find more balanced relationships between public value and private interest in an increasingly interdependent and rapidly digitalising world.

Acknowledgements

We would like to gratefully acknowledge the support of the School of Culture and Communication, University of Melbourne, Australia. We thank the Staff Editorial Advisory Committee for their guidance and Lauren Bliss for her support in the preparation of this issue. We are also thankful to all contributors for their insightful engagement with the theme of this issue, and to our peer reviewers, whose generous and thoughtful feedback greatly strengthened the final publication.

Editorial Team:

Brittany Craig is a PhD candidate at the University of Melbourne and the University of Potsdam. Focusing on the work of Lisa Robertson, Annette Messager, and Rei Kawakubo, her thesis examines how experimental aesthetics across literature, fashion design, and visual art are used to explore and contest feminist issues related to gendered embodiment and subjectivity.

Iván Kirschbaum is a PhD candidate at the School of Culture and Communication, University of Melbourne. His current research focuses on digital media usage, urban precarity, and cities, examining

how Latin American immigrants living in Melbourne use digital media to navigate everyday challenges in housing, work, and practices of belonging.

Jingxian You is a PhD candidate at the University of Melbourne. Her thesis investigates the globalisation of mobile applications and their role in the merging of virtual and physical spaces. Her research interests include digital governance, app globalisation, transnational digital policy, and digital city.

Reference

European Union. (2024). *EU Artificial Intelligence Act*. Retrieved November 6, 2025, from https://artificialintelligenceact.eu/

Flew, T., Martin, F., & Suzor, N. (2019). *Internet regulation as media policy: Rethinking the question of digital communication platform governance*. Journal of Digital Media & Policy, 10(1), 33-50.

Flew, T. (2024). *Mediated trust, the Internet and artificial intelligence: Ideas, interests, institutions and futures*. Policy & Internet, 16(2), 443–457.

Gillespie, T. (2018). *Platforms are not intermediaries*. *Georgetown Law Technology Review*, 198-216.

Gillespie, T. (2020). *Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates*. Internet Policy Review, 9(4), 1–29.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*. Big Data & Society, 7(1), 1–15.

Joseph, D., Nieborg, D. B., & Young, C. J. (2023). *One big store: Source diversity and value capture of digital games in national app store instances*. International Journal of Communication, 17, 7246–7264.

Klonick, K. (2018). *The new governors: The people, rules, and processes governing online speech*. 131 Harvard Law Review, 1598-1670.

Nah, S. (2024). *Mapping scholarship on algorithmic bias: Conceptualization, empirical results, and ethical concerns*. International Journal of Communication, 18, 548–569.

Nieborg, D. B., & Joseph, D. (2020). *App imperialism: The political economy of the Canadian App Store*. Social Media + Society, 6(2), 1–11.

Organisation for Economic Co-operation and Development. (2024). *AI principles*. Retrieved November 6, 2025, from https://www.oecd.org/en/topics/sub-issues/ai-principles.html

Sahebi, S., & Formosa, P. (2025). *The AI-mediated communication dilemma: Epistemic trust, social media, and the challenge of generative artificial intelligence*. Synthese, 205(3), 1-24.

van der Vlist, F. N., Helmond, A., Dieter, M. J., & Weltevrede, E. J. T. (2025). *Super-appification: Conglomeration in the global digital economy*. New Media & Society, 27(6), 3314-333

Gavin Duffy. A Virtuous Ethics of AI: Conviviality as a Regulatory Framework.
Goldsmiths, University of London.

g.duffy@gold.ac.uk

Abstract

In recent years, we have seen the AI industry grow astronomically, becoming a technology that will seemingly impact all elements of our daily lives in the near future. AI omnipresence is now treated by many as almost inevitable, leaving only the question of who should control this technology. This has, understandably, drawn much concern from regulators (both at the state and international level), as well as from many within the AI industry. Unfortunately, there has been less agreement on how we should regulate AI and what the ethical framework for such regulation should be. This article presents two contrasting ethical frameworks of justice in relation to AI: John Rawls (1999) theory of justice as fairness and Ivan Illich's (1973) notion of conviviality. This article critiques the Rawlsian approach as being too concerned with an abstract notion of a 'fair' playing field when establishing notions of justice (through its concepts of the original position and difference principle) and ignoring, or even embracing, injustice of outcomes. In contrast, this article argues in favour of the conviviality approach, presenting it as an ethical framework based in virtue and concerned primarily with outcomes and material reality, rather than hypothetical and semantic notions of fairness. This includes showing how conviviality can be applied practically, applying a comprehensive (or 'thick') notion of sustainability to AI. This thick sustainability considers the entire lifecycle of AI development in considering regulation, including the impacts on ecology as well as the impacts on people. Thus, the conviviality approach de-centres technology and re-centres both humans and our natural environment, providing a holistic ethical framework which must underpin any serious regulation of AI.

Keywords: AI, conviviality, Illich, Rawls, justice, degrowth

Introduction

In recent years, the rise of artificial intelligence (AI) has been astronomical. The public release of ChatGPT is already seen as a watershed moment in re-organising society around AI (Baker, 2024). Nvidia's (a GPU company now specialising in AI chips (Oi, 2024)) growth reflects the changes already being brought about by AI, both economically (briefly becoming the most valuable company in the world (Labiak, 2024)) and geopolitically (becoming a proxy for US-

China tensions (McMorrow and Olcott, 2024)). The presence of AI is felt in domestic governance too, with increasing numbers of countries developing national AI policies, generally with the outlook that AI will inevitably become a central element in our everyday lives (DSIT, 2025a). Combined with well-known and established 'Digital Lords' embracing the technology (Brevini, 2023), AI has seemingly become an unavoidable prospect for even the most technologically hesitant.

Naturally, this raises questions about the governance of AI. If this technology is to be such a terrific force across society, how should it be governed? This article addresses this question through advocating for an ethics of conviviality: a socially oriented form of AI governance, rooted in the notions of human flourishing and equity. In short, a conception of human-centred AI ethics. The following section will detail why such a conception of AI ethics is needed, examining what makes AI distinct from previous digital technologies for regulatory purposes. This article will then discuss two central approaches to AI ethics. The first is a liberal approach, inspired by John Rawls (1999) and dominant in the field of AI ethics (Franke, 2021; Barsotti and Koçer, 2024). This approach suggests a minimisation of harm caused by AI, stemming from a deontological judgement of what constitutes a 'fair' playing field. The subsequent section presents a contrasting, more critical and expansive view on AI ethics, based on Illich's (1973) notion of conviviality. This convivial approach takes a distinctly more outcome-driven approach than the Rawlsian viewpoint, exemplified through examining AI in the media sector. As such, this article argues that the conviviality approach to AI ethics is more practical, more comprehensive, and desirable of the two frameworks, even if (or possibly because) it is also more demanding.

Why do we need an ethics of AI?

One pressing issue in the governance of AI is defining what we mean by "artificial intelligence". Burkhardt and Rieder (2024) note that one of the difficulties in assessing AI is that it is not a single technology but that current AI models represent something 'new' due to their generative, pre-trained, transformer (GPT) capabilities. These models are intended to be domain and task agnostic, based on large-scale foundation models which are much less specialised than generative adversarial network (GAN) models (which use large but narrow datasets to generate convincing outputs of a specific concept). This move towards a generalisable model with generative capabilities has political and ethical implications for societies: if current AI models are believed to be applicable to all tasks, they in turn can influence how we understand the world and what is capable within it. Amoore et al (2024, 2) describe this as AI "instantiating a model of the world,

and with it a set of political logics and governing rationalities that have profound and enduring effects on how we live today". With AI claiming the ability to see underlying (or latent) trends in large datasets, these technologies are becoming powerful actors in shaping our world, often to the (economic) benefit of already wealthy companies, such as Palantir, and at the expense of those already in precarious positions. We are already seeing various AI systems being used in ways which reproduce and reify existing inequalities for refugees (Madianou, 2021), intensify biases in fraud detection with Sweden's welfare system (Amnesty International, 2024), and to identify potential Palestinian targets (including civilians) in Israel's attacks on Gaza (Birch, 2024). This demands questions around the acceptable uses of AI in our world.

Alongside questions about the political characteristics of AI, there are questions of infrastructure. In defining AI, Crawford (2021) includes the creation, maintenance, and disposal processes of AI, rather than focusing solely on the AI product or marketing as experienced by the end user. Thus, AI is also the process of mining rare earth minerals and metals. Crawford reflects on how these practices are not entirely new. Instead, they echo the colonial and extractive history of other technological developments, such as the use of gutta-percha (a natural white latex) for insulating transatlantic telegraph cables. Again, we see the current regime of AI as maintaining long-standing systems of oppression, at both the national and international level, often in pursuit of new media and communication tools favouring the Global North.

These issues point to the need for comprehensive regulation around AI, one which considers the human and environmental impacts of these products first and foremost. Simultaneously, the breadth of these issues points to the difficulty in creating such regulation: if AI is so generalisable and to be used across all elements of society, identifying a specific AI 'regulatory target' seems almost impossible. Indeed, this already appears to be the case for the media sector in many areas, with AI being enabled to undermine intellectual property (IP) laws and the labour of human workers. In the EU, for example, AI developers use an exemption for data scraping in the 2019 Copyright Directive to justify training their models through practices which would normally be considered copyright infringement (Rankin, 2025). Similarly, the British government intends to relax copyright restrictions for developers in training AI models, effectively permitting what would otherwise be considered IP theft (Milmo, 2025). Miltner (2024) points to several more news articles discussing the theft of content and predatory data practices by AI (as well as AI models creating discriminatory or biased outputs) from a range of countries, primarily the US, Mexico, the UK, and India. For media and communication sectors, AI (and the broad scope of GPT-based AI in particular) evidently threatens the ability for people to create new art or

content, both through AI dominating the generation of content and continuing to devour any new human-produced media. In this context, regulation is often framed as being either impossible to create effectively (i.e. AI companies will find a way around it) or is simply in favour of AI companies.

As Miltner (2024, 27) highlights, even in media which laments the predatory and biased data regimes of AI tends to frame this as "just the way things are". This does not simply have to be the case, however. This is a discursive technique which naturalises the power of AI through its supposed inevitability and the powerlessness of citizens to resist (Markham, 2021). This is exemplified in the UK-based NGO Tony Blair Institute for Global Change suggesting that workers should move "beyond narratives of unemployment and Terminator" through a "greater emphasis… on how human workers can be empowered by robots" (Macon-Cooney et al, 2024, 45). In discussing how AI could be *better* regulated in this article, it is therefore imperative to first examine these underlying ways in which AI is conceived of and understood. Subsequently, this article will employ the framework of conviviality to outline a more holistic ethics of AI, informing more effective AI regulation. However, it is first important to outline the current ethical framework used to understand AI, namely a liberal one.

Approaches to AI ethics and regulation

The liberal perspective

Despite numerous critiques of neoliberalism within academic literature, a good deal of research continues to promote a liberal view of AI regulation. In particular, John Rawls' (1999) *A Theory of Justice* (TJ) continues to be influential for political philosophy in general (Laden, 2003) and egalitarianism more specifically (Stone, 2022). Further, Rawls' text transcends academic spheres, finding commercial success in its own time and maintaining prominence in (neo)liberal movements since (Coman, 2020). TJ therefore makes enduring contributions to normative understandings of social issues, justice, ethics, and rationality, extending beyond political theorists and social scientists, influencing economists, lawyers, and even theologians (Richardson and Weithman, 1999). It is this width of influence, across domains and time, that makes Rawls and TJ relevant to AI, a technology that promises to be so generalisable that it will be central to all elements of society. Regulating such a comprehensive technology requires an equally comprehensive ethical framework, given the difficulty described above in regulating the technology in a more piecemeal fashion. This article will therefore examine how Rawls' theory of justice has been applied to AI, highlighting the theory's shortcomings generating comprehensive, effect regulation. First, however, it is important to set out TJ's central concepts.

In TJ, Rawls (1999, 10) articulates the idea of 'justice as fairness', or "the principles that free and rational persons concerned to further their own interests would accept in an initial position of equality as defining the fundamental terms of their association [with society]" and which "regulate all further agreements". This does not need to lead to 'fair' outcomes, only that the principles of justice are initially agreed upon in a fair situation. Rawls outlines two primary principles for achieving this 'fair' justice: distributive justice and the difference principle. Pogge (1982) describes the first principle as guaranteeing the basic liberties of all people (emphasising that this should be understood as global in scope), with these basic liberties only constrained if it promotes greater liberty overall (e.g. the basic liberties of the intolerant may be restricted if it ensures the liberty of those they target and, by proxy, all others). Secondary to this is the difference principle, which states a society should seek to maximise the state of the least advantaged citizens, without violating the first principle (Estlund, 1996).

These are laudable ideas that few would disagree with. Less generously, they may be seen as so vague that few *could* disagree with them. It is therefore worth returning to Rawls (1999) for more detail on these principles. Regarding the difference principle, Rawls states that society should arrive at a conception of fairness (represented through equal liberties) through the original position. Rawls (1999, 11) compares the original position as "[corresponding] to the state of nature in the traditional theory of the social contract… [i.e.] a purely hypothetical situation characterized so as to lead to a certain conception of justice". This hypothetical situation occurs as a contractual negotiation with the intended outcome that "the principles that would be chosen, whatever they turn out to be, are acceptable from a moral point of view" for all (Rawls, 1999, 104). To achieve such results, however, requires that all actors reason from the *original position* behind a *veil of ignorance*. The veil of ignorance is again a hypothetical situation in which one does not know their place in society, his conception of the good, or even the circumstances of their own society overall. Instead, "the only particular facts which the parties know is that their society is subject to the circumstances of justice and whatever this implies" (Rawls, 1999, 119). The intention of the veil of ignorance is therefore to ensure that no one will "design principles to favor his particular condition", meaning that the principles of justice established "are the result of a fair agreement or bargain" and so will be rational (Rawls, 1999, 11). As such, the original position is "a status quo in which any agreements reached are fair" (Rawls, 1999, 104).

Secondly, when discussing the difference principle, Rawls measures what constitutes working to advantage the most disadvantaged not through changed outcomes but through altered

expectations. Specifically, Rawls (1999, 69) recommends that "we simply maximize the expectations of the least favored position subject to the required constraints... [as] the estimated gains from the situation of hypothetical equality are irrelevant, if not largely impossible to ascertain anyway". The difference principle is fulfilled through a positive change in *expectations* of the most disadvantaged in a society, justifying actual material inequalities and "initial inequality in life prospects" (Rawls, 1999, 68). Further, Rawls (1999, 68) positions the greater expectations of the already advantaged as fair and even positive for society as "the greater expectations allowed to entrepreneurs encourages them to do things which raise the prospects of laboring class". Thus, when examining the Rawlsian framework more closely, we can see the ways in which 'justice as fairness' acts to permit and justify inequalities, allowing only for very restricted redress to these issues.

It is, at this point, worth asking: how have Rawls' concepts been applied to AI? Westerstrand (2024) uses the Rawlsian framework to promotes ethical design and use of AI. Regarding Rawls' first principle (on basic liberties), Westerstrand (2024, 5) states that "Rawls offers a preliminary list of basic liberties… to be equally distributed". This includes "liberty and integrity of the person (including freedom from psychological oppression and physical assault and dismemberment)" (Rawls, 1999, 53). Expanding on this, Westerstrand (2024, 8) posits that "AI systems should not harm but support the liberty and integrity of the person, including freedom from psychological oppression and physical assault and dismemberment". This is a pressing matter, according to Westerstrand (2024, 8) as "AI has already being [sic] used in military to automate warfare" which risks causing physical oppression and assault. Regarding Rawls' second principle (the difference principle), Westerstrand (2024, 10) raises concerns that AI "could also lead [to] discrimination of people working certain professions", such as freelance designers or writers, concluding that AI should not be used is it could "negatively impact people's opportunities to seek income and wealth". Again, these are hardly objectionable concerns; they are legitimate insofar as they are both real and material, with NATO investing in Palantir's Maven Smart System (an AI-powered tool that sifts through battlefield data to "scan for targets and speed up attacks") (Foy and Bradshaw, 2025) and AI already being slated to cause massive job losses (Robinson, 2025).

It is, however, unclear how useful Rawls' principles of justice are in either example. Westerstrand (2024) does caution against the use of AI systems which impinge on liberty through physical assault. However, citing Johansson (2018), Westerstrand (2024) also claims that the AI-driven weapons could reduce causalities and so may adhere to Rawls' (1999) notion of liberty (this,

however, appears to ignore Johansson's (2018) warning that this applies only to the possessor of such weaponry and may actually lower the threshold for instigating a war as a result). It is initially somewhat clearer how the difference principle relates to those made unemployed by the use of AI systems, particularly within the arts. Indeed, Westerstrand (2024, 13) states that "following Rawls' theory, AI systems should always thus encourage societal improvement when used in processes that lead to inequalities". As always, it is less clear what this would look like in practice, with Westerstrand (2024) simply suggesting private corporations include the difference principle in their ethical frameworks. Further, Rawls (1999, 68) states that entrepreneurs may be granted unequal benefit under the difference principle should they "do things which raise the prospects of" the least advantaged, including making economic processes more efficient and innovation more rapid. This is exactly the claim made by AI boosters, e.g. the UK government's AI Opportunities Action Plan (DSIT, 2025b), which views AI as a part of the creative industries. Applying TJ and its principles at the case-by-case level can therefore become little more than semantic negotiation around what constitutes an acceptable amount of inequality, rather than eliminating this inequality.

This does not mean that Rawls can have no salience for AI regulation. It may merely mean that it is more important (and productive) to apply the principles of TJ to underlying principles of AI, rather than specific use cases. Indeed, Bay (2023), in critiquing Ashrafian's (2023) notion of a Rawlsian AI agent, suggests that the veil of ignorance, the original position, and difference principle are decidedly macro-principles, rendering them of limited utility for assessing specific AI. Gabriel (2022, 218) utilises a macro-principle approach, stating that AI is now a part of the background justice of our societies, playing an important role in many major institutions and social practices. However, this amounts to little more than recommendations for a public rationale being provided when governments use AI, including "nontechnical explanations of their performance", greater research on antidiscrimination practices and outcomes, and consideration of privacy as a basic right (Gabriel, 2022, 223). These recommendations come with some broad and limiting stipulations: rationale requirements for AI merely apply to "certain public contexts", and solely objects to "purely private goals"; antidiscrimination remains exclusively a matter of discussion; and privacy is only a basic right unless there is an "adequate justification" to the contrary (Gabriel, 2022, 223, 224). This, ultimately, provides only vague suggestions that AI should be reasonably transparent and interfering in certain contexts, to some degree, provided there is not a justification to act otherwise.

Gabriel's limited recommendations point to a central issue with applying a Rawlsian framework to AI and, simultaneously, why Rawls' notion of fairness remains a common one amongst AI-related ethicists (e.g. Larson, 2017; Hashimoto et al, 2018; Heidari et al, 2019; Li et al, 2021; Franke, 2024). As Jørgensen and Søgaard (2023) draw out, the continued use of Rawlsian fairness is due to the permissive nature of TJ, providing a range of exceptions and loopholes to its two central measures of equality. For example, Jørgensen and Søgaard (2023, 1186) state that through "Subgroup Test Ballooning" (tailoring a technology specifically to early adopters, with the argument that it will eventually be adapted for all end users) and "Snapshot-Representative Evaluation" (taking a sample population from the current userbase, rather than an fully representative or even weighted population sample), AI developers can give their products the appearance of 'fairness' (and so 'justness') through ignoring inconvenient (and generally the most precarious) population groups. As such, Rawlsian fairness "is too permissive to prevent common AI/NLP practices that actively contribute to global and social inequality gaps", while purporting to do the opposite (Jørgensen and Søgaard, 2023, 1190).

As noted above, Rawls (1999) discusses such exceptions in TJ, justifying income inequality as fair, for example, provided expectations of workers are managed appropriately. Rawls' notion of justice as fairness is intended to legitimate (at least some of) the inequalities experienced in liberal democracies when examined as a whole system. Applying the Rawlsian approach to AI serves primarily to justify inequalities encoded within and executed by these technologies as *one piece* of the whole system, framing these inequalities simply because of this system alone, rather than as being reified by AI and its developers. This produces distinct negative outcomes e.g. the further centralisation of English as the *lingua franca* at the expense of all other languages (Jørgensen and Søgaard, 2023) and a specific form of standardised English at the expense of other less nondominant Englishes (de Roock, 2024). Such a focus on a specific type of English shapes the ways in which AI models can 'think', perpetuating (dominant) Anglophone understandings of the world, including that of fairness and justice (Tacheva and Ramasubramanian, 2023). When considering the generalisable promises of AI and the universal standards demanded by TJ (Pogge, 1982), it is difficult to see how these exceptions should be justified as fair. In reality, through the permissive broadness of TJ, the Rawlsian framework enables a rhetorically robust but practically loose regulation of AI. This threatens inclusivity in media in ways much broader than the freelancers described by Westerstrand (2024), legitimating an extremely narrow and already dominant understanding of the world through the apparent vastness and consequent omnipotence of AI, leaving room for little else.

As a result, this article suggests that an alternative understanding of justice and fairness is needed for understanding and regulating AI in a manner that is more human-centred. Due to the tension between the deontological Rawls and de-deontological AI, this alternative approach must be more considerate of AI's consequences. This approach is Illich's (1973) conviviality.

Conviviality as an alternative approach

Before making an argument for a convivial approach to AI ethics, it is essential to outline what is meant by "convivial" here, understood through Illich's (1973) definition and application of the term. Instead, Illich uses convivial as a technical term to describe a society in which there is a responsibly limited usage of tools, with modern technologies serving politically interrelated citizens, rather than solely serving managers. Illich (1973, 11) explains that conviviality is an "intrinsic ethical value", that of "individual freedom realized in personal interdependence". A convivial society is therefore one in which people act in creative and autonomous relations with one another and their natural environment. This is contrasted with industrial society in which the power of machines consistently increases at the expense of the individual person, who is degraded to being a mere consumer and subject to the demands of others within a man-made environment.

This is not a binary distinction. Instead, it is only when a society falls below a certain level of conviviality (and industrial productivity rises above a certain level) that the populace becomes plagued by a sense of amorphousness and meaninglessness. Thus, conviviality does not equate to a complete rejection of technology nor that there is an inherently negative quality to technology. Rather, Illich notes that societies and their technologies can either be variously convivial or industrial depending on how they are owned, controlled, and used. Convivial societies are those which ensure a just distribution of unprecedented power (manifest through new technologies), ensuring that the autonomy of one person does not necessitate the subjugation of another. As such, a convivial approach to ethics is one which is interested in full participatory justice. This is in resistance to the ongoing amassing of power by professional elites "who promise to build up the machinery to deliver" futures which are dependent upon high production levels via increasing inequality and energy slaves (Illich, 1973, 12).

It is in this sense that convivial regulation should be understood: rooted in the notion of human flourishing and as a shared virtue. This again stands in contrast to regulation created around a Rawlsian framework of "justice as fairness", in which outcomes are rendered secondary to the

imagined conditions in which they were created. Conviviality as a shared virtue can also be seen in the origins of the term, underpinning the suggested notion of convivial regulation in this article. Illich's definition of conviviality draws upon Aquinas' (1947) argument that austerity is a virtue but must exist in conjunction with pleasure, that neither should be inordinate, instead balancing one another. Such a balance is essential, Aquinas claims, to prevent one from becoming burdensome upon others (should they excessively lack mirth) or to becoming boorish and rude (should lack austerity). It is this balance of mirth and austerity that we see in Illich's (1973) definition of conviviality as personal freedom through mutual interdependence. It is therefore important to note that conviviality is neither negative nor admonishing, even if it does make arguments against the current regulatory regimes. Instead, conviviality is a normative approach rooted in virtue, around the question of the good life at both the individual and collective level.

The convivial approach to ethics thus shares a similarity with the Rawlsian view. Both seek to maximise societal fairness through justice and see individual-level justice as contingent upon the societal-level organisation of fairness. However, the conviviality and Rawlsian approaches differ significantly in what this fairness means and how it is reached. As outlined above, Rawls (1999) puts forward the original position as a means of judging fairness. Once again, this necessitates that, due to the veil of ignorance, no one will "design principles to favor his particular condition" meaning that the principles of justice established "are the result of a fair agreement or bargain" and so will be rational (Rawls, 1999, 11). Such a suggestion appears to be, in itself, irrational. Our understandings of the present and imaginaries of the future are influenced by structural powers, including shaping our perceptions of what a just society is at all (Lukes, 2005). Within an industrial society, industrial forms of justice are to be an expected outcome of the original position, not because of an unwillingness of participants to engage with the *idea* of the original position but because ideas of what is rational (e.g. what values should be prioritised over others and to what extent, to achieve fairness) are inherently shaped by ontological viewpoints. We no longer sacrifice animals to god(s) as a means of repenting for our sins (van Dijk, 2008) but this does not make such activities irrational *in toto*; they simply exist within older forms of rationality. Unless it is believed that the entirety of history was irrational and that the present will always be viewed as rational, any outcomes of the original position must be assumed to be influenced by the context of their place, time, and culture. Illich implicitly recognises this through making an argument for a different form of rationality (conviviality over industrial). TJ does not.

This is a vital point of contention in the context of AI. It is not difficult to see how the current discourse around AI parallels Illich's (1973) warning of professional elites shaping how we imagine the future and political institutions promoting the goal of increased output through conflating the idea of "the good" with what is good for powerful institutions. This logic of industrial society is clearly seen through both national and supranational governments competing to most successfully curry favour with the digital lords of AI, e.g. the British government's AI Opportunities Action Plan (DSIT, 2025b) or the US government's immediate courting of SoftBank and OpenAI for greater AI investment (Hammond, 2025). Further, this approach is rationalised as promoting a common good through notions of increased employment, economic productivity, and environmental regeneration. This is despite many of these claims being visibly untrue and, further, incompatible with one another (Latouche, 2009), particularly given AI's resource intensiveness (Li et al, 2023). As such, any justice derived from an original position under this logical framework could not rationally arrive at a convivial perspective on AI, regardless of how "rational" such an outcome may be. Instead, the outcome from this original position would rationally be one which promotes increased use of AI in all spaces and an increasing allocation of resources and priority to AI. This is, in fact, what many AI boosters suggest and what many governments are seeking to do (DSIT, 2025b; Hammond, 2025). Whether or not such decisions are correct is immaterial to whether or not they are rational; they are rational within the given framework of thinking. Rawls (1999, 11) states that justice as fairness "does not mean that the concepts of justice and fairness are the same, any more than the phrase "poetry as metaphor" means that the concept of poetry and metaphor are the same". Similarly, rationality and correctness are not the same, even if something can be correct under a certain rationality.

Conversely, Illich's (1973) conviviality framework has been influential for many degrowth-oriented approaches to contemporary digital technologies, including AI. In particular, Illich's conviviality framework has inspired means of testing for 'fairness' in ways which are decidedly more robust and less permissive than Rawls' (1999) TJ. In considering specific products, for example, Vetter (2018) establishes a matrix of convivial technology which can act as a guide for what human-centred AI regulation may privilege. This involves promoting technologies which: recognise that humans exist in a series of relations to one another and so seek to promote positive relations between people; consider both material (hardware) and immaterial (software, knowledge) accessibility, as well as accessibility across different groups (e.g. addressing the traditionally male biases in technological development); have clear utility in their ecologies,

including ethical plans for the product's end-of-life, rather than simply being 'less harmful'; and consider the appropriateness of the product, with serious consideration of where it may *not* be useful, including where technologies may be desirable but not necessary. This, evidently, goes beyond a Rawlsian notion of fairness through a strict, clearly articulated criteria by which technologies should be measured across their lifecycle and its chain of production, resulting in a substantially less permissive framework for justice.

Considering sustainability at a more macro level, Heilinger et al (2024) develop a framework for assessing and regulating for the "thick" sustainability of AI. Thick sustainability is an approach to sustainable AI which looks not just at how the technology is used *for* sustainability purposes but also sustainable *as* a technology. This includes not only the environmental sustainability of AI but its social sustainability as well, discussed in the context of media in the following section. Heilinger et al contrast this with thin sustainability, which only examines the direct impacts of AI's immediate ecological actions, e.g. identifying more efficient strategies to deal with climate change, while prioritising economic sustainability over social sustainability. It is this 'thin' sustainability which AI ethicists and developers appeal to through the Rawlsian framework to make claims toward thin sustainability, relying on 'fair' exceptions carved out in ambiguous regulations (Gabriel, 2022), statistically and rhetorically concealing their supply chains (Crawford, 2021) and those othered by AI (Jørgensen and Søgaard, 2023).

In contrast, conviviality-based approaches such as that of Heilinger et al (2024) avoid the permissiveness of TJ through making companies responsible for the whole lifecycle of their product, and particularly its impacts. Through focusing on the life of a product, rather than theoretical assessments of fairness enabled through the Rawlsian approach, frameworks inspired by Illich (1973) pro-actively and continuously seek a society in which people are able to exist with greater agency, living in conjunction with technology rather than subject to it, i.e. a more convivial society. Rather than being permissive of an unjust outcome due to the supposedly fair nature of the contractual bargaining process which created the injustice, a framework of conviviality demands an outcome-oriented approach to fairness and justice. In practice, this is likely to come at the expense of the economic 'sustainability" (i.e. perpetual growth) prized by thin sustainability, recognising that this economic growth is inequitable and undesirable for a majority of the world's population, yoking them to an unjust economy of AI to enable the flourishing of a few.

The conviviality framework therefore operates as a more human-centred approach to regulation through this systematic approach to AI, in contrast to the narrower frame often used to assess

what an AI "does" or "is". Conviviality resists technosolutionist or technologically deterministic regulation through maintaining a critical (but not cynical) disposition to new digital technologies, seeing AI as yet another tool to be regulated and managed rather than as a digital Leviathan. This distinction is important, as we already see how AI is often framed as being almost mythical (Leaver and Srdarov, 2025), as opposed to a new watershed in the timeline of digital technology. This demystification of AI de-centres the technology, and the economic sustainability associated with it, in favour of greater human (and environmental) sustainability.

It should be noted that this article primarily argues for the adoption of such a convivial framework, rather than suggesting that this framework is already entirely constructed. The approaches to convivial AI discussed here represent practical steps towards ethical regulation of AI. In particular, the focus on developers' responsibility for their products throughout their production, use, and end-of-life states ensure a less permissive, more demanding idea of just regulation for AI than is seen through the use of Rawls (1999) and TJ. However, there remains more to be done in establishing comprehensive regulation. The following section raises some of these concerns, focusing on the interaction of AI and the media, discussing already emerging issues and the inability for the current, Rawlsian view of ethics to properly address these problems. These are issues which must be dealt with by future research, with a convivial approach presenting the best framework for achieving a practical, humane, and ultimately fair outcome.

What does this mean for media and communication?

As has been noted throughout, a great deal of the issues around regulating AI impact media and communications. Perhaps the most well-known issue (mentioned above) is that of AI models scraping data from across news sources, often being made exempt from copyright laws or simply infringing upon them (Grynbaum and Mac, 2023). Large AI companies are not only interested in existing media, however, but in producing media as well. De-Lima-Santos and Ceron (2022) find that the use of AI in news media largely relies on news organisations purchasing AI models from third-party companies, particularly large technology companies such as Alphabet. De-Lima-Santos and Ceron do note that AI produced text is seen less frequently in non-English languages, due to the English-centric nature of these models. While this could be taken to mean that non-English news media is not under threat by AI, it is more likely that this means non-English media will see an indirect harm by AI by being made more peripheral (de Roock, 2024).

Local news is particularly vulnerable to this kind of economic interference of AI. In the UK, for example, Reach PLC (the nation's largest local news company and owners of national papers such as the *Mirror* and *Express*, cumulatively reaching 69% of the country's population online) have been using AI since 2023, focusing on replicating articles across sites in a manner favoured by AI's ranking system (Gupta, 2024; Tribune, 2025) and ensuring content is considered 'appropriate' for advertisers (IBM, 2019). Similarly, Google's Digital News Initiative Innovation Fund awarded a grant to PA Media (then the Press Association) to develop their RADAR-AI (Gregory, 2017). RADAR-AI uses national level data to generate local news, including on children in custody, welfare payments, and council spending on temporary accommodation for homeless households (Care, 2025). AI companies are increasingly embedding themselves within the production and dissemination of news media, shaping what is considered 'valuable' in a story (i.e. how well it appeals to search algorithms and digital advertisers), and increasingly financialising an already precarious sector. This is worsened by the inaccuracies repeatedly found within such tools (Rahman-Jones, 2025), a limiting of journalists' editorial freedom (Thäsler-Kordonouri, 2025) and simply a lack of real knowledge about local areas (Tribune, 2025). This is felt by news readerships as well, with AI journalism undermining the trust readers have in the news, even when the content itself is still seen as being accurate and fair (Toff and Simon, 2023).

The risks posed by AI in news media therefore go well beyond making freelance journalism more difficult (Westerstrand, 2024), instead posing issues for the sector at every point of production and reception. Without a strong regulatory framework, one which considers the ways in which people either can or must interact with technology, it is difficult to imagine how this phenomenon will not worsen. This poses an issue for the deontological Rawlsian framework. Unless the decreasing number of jobs in journalism is considered a fundamental impingement upon the basic liberty of all citizens (although it seems unlikely that an equal job-to-demand ratio is a fundamental freedom and, if so, Rawlsians should take issue with *all* technologies since the industrial revolution), the rise of AI does not appear to threaten TJ's primary principles of justice. Further, provided that a government provides a reasonable justification for allowing AI use in such a manner, the issue of publicity as set out by Gabriel (2022) is averted. A convivial approach, conversely, prioritises the relationship that citizens have with technology (and with the societal institutions which own and deploy these technologies).

This approach to news media is not an aberration but rather is indicative of the wider perspective taken toward communicative and creative media by the AI sector. This is perhaps best exemplified by OpenAI CEO Sam Altman's recent interview at TED2025 (Cadwalladr,

2025). During this interview, Altman was asked if ChatGPT was committing IP theft, to which the present audience applauded. Altman simply responded, "you can clap about that all you want, enjoy... I think that people have been building on the creativity of others for a long time... I think there are incredible new business models that me and others are excited to explore" (TED staff, 2025). There is a clear desire from AI developers to further the economic precarity established through the platformised economics of media creation and dissemination (Drott, 2024), with AI developers becoming central to the political economy of creative expression in media.

Further, during this interview, Altman made a statement that exemplifies the underlying perspective on AI developers around creativity: "if you can't tell the difference, how much do you care?". This is in reference to being unable to know if AI is 'thinking' or just repeating data from its training set, but speaks to the wider implications of AI produced content in general (Altman himself prefaces this statement by describing it as an "incredible meta-answer") (TED, 2025). This statement articulates a direct response to concerns over the consequences of AI for human-centred creative outputs and the displacement of professional media careers: who cares? Altman's statement belies the perspective of AI developers around creation, i.e. all that matters is the end product, devoid of its context for creation or reason for being. This, in a sense, is a coherent viewpoint. If AI is a machine built upon and generative of consequences, it logically follows that those who create AI would be consequence focused as well. AI's perspective does not originate from the void; it is reflective of the viewpoint of its creators (which are in turn influenced by the products they create and so on).

This again returns us to the need for a consequence focused idea of justice to act as a regulatory counterweight to the ongoing AI-ification of the world. Donahue (2025) argues that there is value to maintaining a burden of collective moral achievement amongst a populous, i.e. the opportunity for individuals to come to and make their own moral decisions over time, as well as being a part of a larger society that makes moral judgements over time. Without the opportunity to make poor moral judgements, making good moral judgements is rendered less meaningful. Similarly, for media and communications, this article argues there is a collective creative achievement which would be undermined by loose non-human-centred AI regulation. This includes the individual level of being able to create art poorly, which gives greater meaning to art which is created well; and the collective level in which there must be opportunity to create art with potentially limited mass appeal but substantial value to those whom it does appeal (in the

context of AI, this may include non-English language content, something which has substantially wide appeal but is not necessarily captured by AI).

In TJ, Rawls (1999) constructs a fluid framework for society, which makes few normative claims about what justice looks like beyond provided it adheres to an ex-ante agreement on the fairness of society (and so the fairness and justness of its inequalities). Conversely, Illich's (1973) conviviality offers a framework for society based on normative ideas of how justice should be experienced and what just relations should look like in society, primarily based on our relations with one another and with technology. This framework therefore continues to make human-centred demands of justice ad tempus, in which justice is less concerned about previous agreements of what, in an abstract sense, is a fair and contractual agreement but instead sees justice as something to be constantly renegotiated in the face of new sociotechnical and material conditions. In order to preserve a thick sustainability of creative media output (and, indeed, improve current conditions), such an approach is necessary to counter the entirely outcome-driven ideology of AI. Without this, we risk an even greater enclosure of media creation, one which does not see an intrinsic value in the creation process (and the processes preceding creation, such as learning), instead seeing value only in quantifiable metrics such as data created and economic value. Seemingly, all that the Rawlsian approach can offer here is a demand for 'publicity', that we be made aware that AI is used and given justifications for this use, managing the expectations of citizens and so meeting TJ's criteria for fairness but evidently failing any measure of collective creative achievement.

Conclusion

With AI currently occupying such a large space in public discourse, particularly around how ubiquitous it should be in everyday life, it is vital to consider how this emerging technology should be regulated. It is for this reason that this article presents two opposing views when considering what constitutes a human-centred, ethical approach to AI regulation. The first is the liberal, Rawlsian view of justice as fairness. This position begins with the idea that justice should be distributive, established through the original position and difference principle. This is not to say that all should be equal. Rather, there *is* acceptance of an "appropriate division of advantages" by Rawls (1999, 15), provided that this distribution is generally acceptable to all when considered from the original position. Thus, the Rawlsian view is a deontological ethical framework and has been popular with many AI ethicists.

The alternative approach suggested in this article is based in Illich's (1973) notion of conviviality. Conviviality, as it is used here, is distinct from the Rawlsian view in that it is concerned with outcomes, rather than a more abstracted ethical position. Fundamentally, a convivial approach to regulation is based in the notion that technologies should exist to serve people, rather than people existing to serve technologies (for the benefit of a small number of people). In viewing AI as a technology, existing in a genealogy of other digital technologies, the conviviality approach emphasises that AI is malleable to human agency, rather than seeing AI as somehow inevitable. Conviviality therefore operates as a distinctly human-centred position, seeing the 'technology' itself as secondary to the social relations which surround it. This is considered in the context of AI through the matrix of conviviality and thick sustainability, which consider the importance of social and cultural sustainability alongside environmental sustainability. These two perspectives on justice are finally applied to news media, discussing the need for a comprehensive means of regulating AI in the news media and media more generally. Thus, through providing a more outcome-oriented framework that is interested in promoting the greatest level of virtue within society, the conviviality approach provides a practical and impactful starting point for regulating AI. This stands in contrast to the Rawlsian approach of seeking out the 'least bad' outcome and a hoped-for minimisation of disadvantage: in any human-centred ethics, we must demand more than this.

References

Amnesty International. (2024). Sweden: Authorities must discontinue discriminatory AI systems used by welfare agency. Retrieved 22 January, 2025 from https://www.amnesty.org/en/latest/news/2024/11/sweden-authorities-must-discontinue-discriminatory-ai-systems-used-by-welfare-agency/

Amoore, L., Campolo, A., Jacobsen, B., & Rella, L. (2024). A world model: On the political logics of generative AI. Political Geography, 113, https://doi.org/10.1016/j.polgeo.2024.103134

Aquinas, T. (1947). Summa Theologica. Einsiedeln, Switzerland: Benziger Bros.

Ashrafian, H. (2023). Engineering a social contract: Rawlsian distributive justice through algorithmic game theory and artificial intelligence. AI and Ethics, 3(4), 1447-1454. https://doi.org/10.1007/s43681-022-00253-6

Baker, S. (2024). Rise of ChatGPT and other tools raises major questions for research. *Nature*, *633*(8030), 5-5. https://doi.org/10.1038/d41586-024-02984-4

Barsotti, F., & Koçer, R. G. (2024). MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias. AI & Society, 39(3), 961-974. https://doi.org/10.1007/s00146-022-01577-x

Bay, M. (2024). Participation, prediction, and publicity: avoiding the pitfalls of applying Rawlsian ethics to AI. AI and Ethics, 4(4), 1545-1554. https://doi.org/10.1007/s43681-023-00341-1

Birch, M. (2024). Who did that? AI assisted targeting and the lowering of thresholds in Gaza. Medicine, Conflict and Survival, 40(2), 97-100. https://doi.org/10.1080/13623699.2024.2364937

Brevini, B. (2023). Global Digital Lords and Privatisation of Media Policy: The Australian Media Bargaining Code. Javnost-The Public, 30(2), 268-283. https://doi.org/10.1080/13183222.2023.2207427

Burkhardt, S., & Rieder, B. (2024). Foundation models are platform models: Prompting and the political economy of AI. Big Data & Society, 11(2), 1-15. https://doi.org/10.1177/20539517241247839

Cadwalladr, C. (2025). It's not too late to stop Trump and the tech broligarchy from controlling our lives, but we must act now. The Guardian. Retrieved 29 April, 2025 from https://www.theguardian.com/global/2025/apr/20/carole-cadwalladr-ted-talk-this-is-what-a-digital-coup-looks-like-its-not-too-late-to-stop-trump-and-the-silicon-valley-broligarchy-from-controlling-our-lives-but-we-must-act-now

Care. A. (2025). RADAR round-up: A look at the stories delivered by the RADAR-AI editorial team in the month of March. PA Media. Retrieved 29 April, 2025 from https://pa.media/blogs/editorial-data/radar-round-up-a-look-at-the-stories-delivered-by-the-radar-ai-editorial-team-in-the-month-of-march-2/

Coman, J. (2020). John Rawls: can liberalism's greatest philosopher come to the west's rescue again? The Guardian. Retrieved 30 April 2025 from https://www.theguardian.com/inequality/2020/dec/20/john-rawls-can-liberalisms-great-philosopher-come-to-the-wests-rescue-again

Crawford, K. (2021). Atlas of AI. New Haven, CT: Yale University Press.

de Roock, R. S. (2024). To become an object among objects: Generative artificial "intelligence," writing, and linguistic white supremacy. Reading Research Quarterly, 59(4), 590-608. https://doi.org/10.1002/rrq.569

de-Lima-Santos, M. F., & Ceron, W. (2021). Artificial intelligence in news media: current perceptions and future outlook. Journalism and media, 3(1), 13-26. https://doi.org/10.3390/journalmedia3010002

Department for Science, Innovation and Technology (DSIT). (2025a). Prime Minister sets out blueprint to turbocharge AI. Gov.uk. Retrieved 30 April 2025 from https://www.gov.uk/government/news/prime-minister-sets-out-blueprint-to-turbocharge-ai

Department for Science, Innovation and Technology (DSIT). (2025b). AI Opportunities Action Plan. Gov.uk. Retrieved 22 January, 2025 from https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

Donahue, S. (2025). AI rule and a fundamental objection to epistocracy. AI & Society, 1-13. https://doi.org/10.1215/9781478027874

Drott, E. (2024). Streaming music, streaming capital. Durham, NC: Duke University Press. https://doi.org/10.1215/9781478027874

Estlund, D. (1996). The survival of egalitarian justice in John Rawls's political liberalism. The Journal of Political Philosophy, 4(1), 68-78. https://doi.org/10.1111/j.1467-9760.1996.tb00042.x

Foy, H. & Bradshaw, T. (2025). Nato acquires AI military system from Palantir. Financial Times. Retrieved 30 April, 2025 from https://www.ft.com/content/7f80b1bc-114c-4a00-ad06-6863fb435822

Franke, U. (2021). Rawls's original position and algorithmic fairness. Philosophy & Technology, 34(4), 1803-1817. https://doi.org/10.1007/s13347-021-00488-x

Franke, U. (2024). Rawlsian Algorithmic Fairness and a Missing Aggregation Property of the Difference Principle. *Philosophy & Technology*, *37*(3), 1-19. https://doi.org/10.1007/s13347-024-00779-z

Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. Daedalus, 151(2), 218-231. https://doi.org/10.1162/daed_a_01911

Gregory, J. (2017). Press Association wins Google grant to run new service written by computers. The Guardian. Retrieved 29 April, 2025 from https://www.theguardian.com/technology/2017/jul/06/press-association-wins-google-grant-to-run-news-service-written-by-computers

Grynbaum, M. M., & Mac, R. (2023). The Times sues OpenAI and Microsoft over AI use of copyrighted work. The New York Times. Retrieved 29 April, 2025 from https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Gupta, N. (2024). How UK's Reach is using AI to help produce more content faster. WAN-IFRA. Retrieved 29 April, 2025 from https://wan-ifra.org/2024/10/how-uks-reach-is-using-ai-to-help-produce-more-content-faster/

Hammond, G. (2025). SoftBank and OpenAI back sweeping AI infrastructure project in US. Financial Times. Retrieved 22 January, 2025 from https://www.ft.com/content/48eb53a1-67ca-4509-8c62-401f0cf8b099

Hashimoto, T., Srivastava, M., Namkoong, H., & Liang, P. (2018, July). Fairness without demographics in repeated loss minimization, International Conference on Machine Learning, 1929-1938. https://doi.org/10.48550/arXiv.1806.08010

Heidari, H., Ferrari, C., Gummadi, K., & Krause, A. (2019). Fairness behind a veil of ignorance: A welfare analysis for automated decision making. Advances in neural information processing systems, 31, 1-17. https://doi.org/10.48550/arXiv.1806.04959

Heilinger, J. C., Kempt, H., & Nagel, S. (2024). Beware of sustainable AI! Uses and abuses of a worthy goal. AI and Ethics, 4(2), 201-212. https://doi.org/10.1007/s43681-023-00259-8

IBM. (2019). Reach and IBM launch Mantis, using IBM Watson to make brand safety smarter. IBM. Retrieved 29 April, 2025 from https://uk.newsroom.ibm.com/2019-10-17-Reach-and-IBM-launch-Mantis-using-IBM-Watson-to-make-brand-safety-smarter

Illich, I. (1973). Tools for Conviviality. London, UK: Calder & Boyars.

Jørgensen, A. K., & Søgaard, A. (2023). Rawlsian AI fairness loopholes. AI and Ethics, 3(4), 1185-1192. https://doi.org/10.1007/s43681-022-00226-9

Labiak, M. (2024). AI frenzy makes Nvidia the world's most valuable company. BBC. Retrieved 22 January, 2025 from https://www.bbc.co.uk/news/articles/cyrr40x0z2mo

Laden, A. S. (2003). The house that Jack built: Thirty years of reading Rawls. Ethics, 113(2), 367-390. https://doi.org/10.1086/342855

Larson, B. (2017). Gender as a Variable in Natural-Language Processing: Ethical Considerations. 3, 1-11. https://doi.org/10.18653/v1/W17-1601

Latouche, S. (2009). Farewell to Growth. Cambridge, UK: Polity Press.

Leaver, T., & Srdarov, S. (2024). Generative AI and children's digital futures: New research challenges. Journal of Children and Media, 1-6. https://doi.org/10.1080/17482798.2024.2438679

Li, M., Namkoong, H., & Xia, S. (2021). Evaluating model performance under worst-case subpopulations. Advances in Neural Information Processing Systems, 34, 17325-17334. https://doi.org/10.48550/arXiv.2407.01316

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. Communications of the ACM, 1-10. https://doi.org/10.48550/arXiv.2304.03271

Lukes, S. (2005). Power: A Radical View (Second Edition). Basingstoke, UK: Palgrave Macmillan.

Macon-Cooney, B., Mökander, J., Stanley, L., & Decorte, R. (2024). A New National Purpose: The UK's Opportunity to Lead in Next-Wave Robotics. Tony Blair Institute for Global Change. Retrieved 22 January, 2025 from https://assets.ctfassets.net/75ila1cntaeh/55qX0nXXRWUSe7q4x4xntd/e93f86c94f9d7b1910005a1ca6d45bd5/2Xu2ib2C6zxbYJzxch9L1R--152708102024

Madianou, M. (2021). Nonhuman humanitarianism: when 'AI for good' can be harmful. Information, Communication & Society, 24(6), 850-868. https://doi.org/10.1080/1369118X.2021.1909100

Markham, A. (2021). The limits of the imaginary: Challenges to intervening in future speculations of memory, data, and algorithms. New media & society, 23(2), 382-405. https://doi.org/10.1177/1461444820929322

McMorrow, R., & Olcott, E. (2024). Nvidia's AI chips are cheaper to rent in China than US. Financial Times. Retrieved 22 January, 2025 from https://www.ft.com/content/10aacfa3-e966-4b50-bbee-66e13560deb4

Milmo, D. (2025). UK copyright law consultation 'fixed' in favour of AI firms, peer says. The Guardian. Retrieved 29 April, 2025 from https://www.theguardian.com/technology/2025/feb/11/uk-copyright-law-consultation-fixed-favour-ai-firms-peer-says

Miltner, K. M. (2024). "AI is holding a mirror to our society": Lensa and the discourse of visual generative AI. Journal of Digital Social Research, 6(4), 13-33. https://doi.org/10.33621/jdsr.v6i440456

Oi, M. (2024). Nvidia: US tech giant unveils latest artificial intelligence chip. BBC. Retrieved 22 January, 2025 from https://www.bbc.co.uk/news/business-68603198

Pogge, T. (1982). The Interpretation of Rawls' First Principle of Justice. Grazer Philosophische Studien, 15, 119, 119-147.

Rahman-Jones, I. (2025). AI chatbots unable to accurately summarise news, BBC finds. BBC. Retrieved 29 April 2025, from https://www.bbc.co.uk/news/articles/c0m17d8827ko

Rankin, J. (2025). EU accused of leaving 'devastating' copyright loophole in AI Act. The Guardian. Retrieved 29 April, 2025 from https://www.theguardian.com/technology/2025/feb/19/eu-accused-of-leaving-devastating-copyright-loophole-in-ai-act

Rawls, J. (1999). A Theory of Justice (Revised Edition). Oxford, UK: Oxford University Press.

Richardson, H. & Weithman, P. (1999). The Philosophy of Rawls: A Collection of Essays. New York, NY: Garland Publishing.

Robinson, B. (2025). Fears about AI job loss: New study answers if they're justified. Forbes. Retrieved 30 April, 2025 from https://www.forbes.com/sites/bryanrobinson/2025/02/09/fears-about-ai-job-loss-new-study-answers-if-theyre-justified/

Stone, P. (2022). In the Shadow of Rawls: Egalitarianism Today. Ethical Theory and Moral Practice, 25(1), 157-168. https://doi.org/10.1007/s10677-022-10272-1

Tacheva, J., & Ramasubramanian, S. (2023). AI Empire: Unraveling the interlocking systems of oppression in generative AI's global order. Big Data & Society, 10(2), 1-13. https://doi.org/10.1177/20539517231219241

TED Staff. (2025). What's next for AI? A conversation with OpenAI's Sam Altman – Live at TED2025. TEDBlog. Retrieved 29 April, 2025 from https://blog.ted.com/whats-next-for-ai-a-conversation-with-openais-sam-altman-live-at-ted2025/

Thäsler-Kordonouri, S. (2025). News automation in UK newsrooms, in N. Thurman, I. Henkel, S. Thäsler-Kordonouri, & R. Fletcher (Eds.), UK Journalists in the 2020s: Who they are, how they work, and what they think, 27-31.

The Tribune. (2025). AI is taking over local journalism. Does it matter? Sheffield Tribune. Retrieved 29 April, 2025 from https://www.sheffieldtribune.co.uk/ai-is-taking-over-local-journalism-does-it-matter/

Toff, B. & Simon, F. (2023). "Or they could just not use it?": The Paradox of AI Disclosure for Audience Trust in News. https://doi.org/10.31235/osf.io/mdvak

Van Dijk, J. (2008). In the shadow of Christ? On the use of the word "victim" for those affected by crime. Criminal Justice Ethics, 27(1), 13-24. https://doi.org/10.1080/0731129X.2008.9992224

Vetter, A. (2018). The matrix of convivial technology–assessing technologies for degrowth. Journal of cleaner production, 197, 1778-1786. https://doi.org/10.1016/j.jclepro.2017.02.195

Westerstrand, S. (2024). Reconstructing AI Ethics Principles: Rawlsian Ethics of Artificial Intelligence. Science and Engineering Ethics, 30(5), 1-21. https://doi.org/10.1007/s11948-024-00507-y

Mark Bo Chen. "I'm a bit cautious of jumping in with both feet": exploring information ownership and negotiated control in AI chatbot users' communication privacy management.

School of Culture and Communication, The University of Melbourne.
Email: markchan0814@gmail.com

Abstract

Advances in artificial intelligence have garnered significant attention, with user privacy emerging as a focal point. Guided by a privacy management perspective, this exploratory study investigates how users make sense of informational privacy when interacting with their AI chatbot counterparts, drawing from Reddit data (submissions, n=193) that represent unsolicited user vignettes of chatbot-related privacy experiences. Situated in Human-Machine Communication (HMC), the study applies Communication Privacy Management (CPM) theory to analyse how information ownership and control are understood and negotiated as part and parcel of privacy management strategies in user-chatbot communication. Findings reveal users' struggle to grapple with boundary regulations in automated systems; their situational strategies of boundary making are shaped not only by users' disclosure intention and privacy concerns, but also the techno-social features of chatbots that limit the extent to which users' tactics of privacy management are practised. With a user-centric approach, this study extends CPM to HMC and contributes to our understanding of how ordinary users perceive and negotiate informational privacy in the context of everyday AI use. Theoretical and practical implications are discussed.

Keywords

Chatbot; Privacy; Artificial Intelligence; Communication Privacy Management; Human-Machine Communication; ChatGPT

Introduction

Recent years have witnessed an explosion in artificial intelligence (AI) technologies, including chatbots powered by large language models. Broadly, AI are complex techno-social assemblages (Eynon and Young, 2021), constructed through social processes that encapsulate not only the technicality, but also the knowledge, practices and negotiation in handling these systems (Guzman and Lewis, 2019). In everyday life, how users engage with AI technologies is fundamentally grounded in communication practices as relational collaborations (e.g., using natural language to communicate with chatbots) (Gunkel, 2012; Guzman, 2018). On the other hand, communication privacy is relevant to nearly all human activities (Altman, 1975; Petronio, 2002), and poses challenges in the context of AI use, particularly due to the opacity of algorithmic systems and the dynamic ways in which user data can be inferred, stored and repurposed beyond the original context (Gorwa and Veale, 2024; Lutz et al., 2019). Therefore, it is not surprising that while popular AI chatbots like ChatGPT are widely embraced in daily lives (e.g., Westfall, 2023), sentiments of uncertainty prevail, with one of the heated topics being loss of control on data and informational privacy (e.g., Sher and Benchlouch, 2023). As the hype around AI continues, communication research is required to understand, beyond the current hyperbole surrounding technological progressions, how ordinary people make sense of AI and manage privacy when they communicate directly with these machines.

AI chatbots, as epitomised by OpenAI's ChatGPT, are a type of narrow AI. Narrow AI is designed to perform a particular task, and in this sense, is seen as having limited capacity. Chatbots can extract information from user inputs and create outputs sensitive to the inputs and comprehensible to humans (Allen, 2003). Their functionalities rely on datafication (Hepp, 2020), that is, the collection and processing of large amounts of data to learn relationships between words and remember conversations and contextual dependencies to personalise responses to users. Personalisation sustains various utilitarian and social needs that motivate users to interact with AI chatbots (Brandtzaeg et al., 2022; Skjuve et al., 2024). Given the vast amounts of user data involved, these data-driven benefits can also lead to anxieties around what data are collected, how the data are processed, with whom the data are shared, and what measures are in place to protect user privacy.

Empirical research on user privacy and AI chatbots remains limited, with much literature (Ischen et al., 2019; Lim and Shim, 2022; Liu et al., 2023; Sannon et al., 2020) relying on experimental designs to measure privacy intentions in isolated environments and as numeric metrics. These designs risk priming participants to inflate their privacy concerns and overlook the relational and negotiated nature of communication privacy management (Palen and Dourish, 2003; Petronio, 2002). Furthermore, while some studies (Ischen et al., 2019; Lim and Shim, 2022) suggest that anthropomorphic design can reduce privacy fears, other perspectives (Liu et al., 2023; Sannon et al., 2020; Sundar and Kim, 2019) highlight persistent tensions in how users trust and manage information with chatbots. This underscores the need for deeper investigation into how communication privacy is understood and negotiated when users interact with conversational AI systems in the wild.

This exploratory study elucidates how users articulate their privacy experiences in everyday interactions with their chatbot counterparts, based on a Reddit-sourced dataset (n=193) from five sub-reddit forums (*r/ChatGPT*, *r/ClaudeAI*, *r/perplexity_ai*, *r/GeminiAI, r/CharacterAI*). Using Communalytic[1], submissions were collected in two phases, screened for relevance and then analysed thematically. In doing so, the present study moves beyond laboratory settings and evaluates how ordinary chatbot users understand information ownership and negotiated control, two key facets of privacy management. Findings were contextualised in the domain of Human-Machine Communication (HMC; Guzman, 2018; Guzman and Lewis, 2020), and interpreted using the Communication Privacy Management theory (CPM; Petronio, 2002) which considers how individuals develop rules to manage information disclosure, and (re-)negotiate these rules when boundary turbulence arises in episodes of privacy breakdown. This study extends CPM, a theory traditionally applied in interpersonal communication, to HMC, arguing that communication privacy behaviours are results of situational negotiations between users and chatbots, shaped by both technical affordances and interactional dynamics. The sections that follow begin with a review of relevant literatures and detail the methodological approach and data sources. Then, key findings are presented, followed by a discussion of their implications, a reflection on limitations, and an outline for future research.

Literature Review

Privacy Management and Communication Privacy Management Theory

Contemporary privacy scholarships draw on inspirations from diverse domains including sociology, psychology and law (Altman, 1975; Westin, 1967). Different perspectives have produced numerous tomes of insightful research but also complicate a universally applicable understanding of privacy (Solove, 2006). In communication research, a widely adopted definition comes from Westin's work (Lutz, 2023) where privacy is conceptualised as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (Westin, 1967, p. 7). Implicit to this definition is the informational dimension of privacy, which frames privacy as a matter of information management. While other dimensions of privacy are crucial to discussions on AI technologies more broadly (see Lutz et al., 2019), this research focuses on AI chatbots and echoes Lutz's argument (2023) that privacy implications of user-chatbot interactions primarily concerns the exchange of information. This can range from metadata (e.g., IP address, timestamps) to interactional content (e.g., chat logs, uploaded documents), as part of accessing and using chatbot services.

Digital technologies mediate not just information flow, but also emotional and affective relations (e.g., Bucher, 2017). This contributes to rendering boundaries between human and technology increasingly ambiguous (Turkle, 2005), giving rise to emerging forms of human-technology intimacy (Li and Zhang, 2024) and privacy implications (Lim and Shim, 2022). For AI chatbot users, privacy concerns may be sourced from a perceived loss of control over private information. Simultaneously, utilitarian and social benefits—such as productivity (Skjuve et al., 2024), personalisation and social connectedness (Brandtzaeg et al., 2022)—motivate continued use. As some degree of disclosure is required to use technology (Palen and Dourish, 2003), users face a tension between privacy fears (pushing factors) and the benefits (pulling factors). In this light, privacy in a human-chatbot dyad is not simply about a dichotomy between disclosure and concealment, but rather the selective control of access to personal information (Altman, 1975) and "the continual management of boundaries between different spheres of action and degrees of disclosure within those spheres" (Palen and Dourish, 2003, p. 131). Given the push-and-pull dynamics as described, it can be argued that the management of personal information flow and varied degrees of disclosure undergird individual users' privacy management practices in user-chatbot interactions.

To govern information flow, the tension between various pulling and pushing forces need to be mitigated. Petronio's Communication Privacy Management Theory (2002) provides a framework to make sense of such dialectical tension between privacy and disclosure. As a rule-based system, the theory posits that there are both risks and benefits to disclosure, and as such, individuals in dyadic relationships erect communication boundaries and establish privacy management rules for the disclosure and protection of privacy information, based on the belief that they are the owner of such information. According to CPM, these rules emerge from the "dialectical tension between openness and closedness" (Child et al., 2009, p. 2082), and are aimed at striking a balance between solitude and sociality in relational contexts. At its core, CPM rejects dichotomous thinking and recognises that disclosure and control of information are distinct user privacy management tactics, which has been extended to different technology-mediated environments including online blogging (e.g., Child et al., 2009), social media (e.g., Kang et al., 2022), e-commerce (e.g., Metzger, 2007) and smart technologies (e.g., Vitak et al., 2023). Therefore, although initially developed in

the domain of interpersonal communication, these existing cases showcase CPM's versatility and applicability in analysing technology- and privacy-related issues.

This research is inspired by CPM key principles to move beyond treating privacy as mere disclosure-withdrawal juxtaposition. It explores privacy practices as negotiated efforts of boundary management in everyday user-chatbot interactions. The next section builds on existing applications of CPM in technology-mediated communication and examines the theory's relevance to HMC. It then contextualises CPM within HMC's key focus on direct user engagement with communicative machines like chatbots.

Communication Privacy Management in User-chatbot Communication

CPM has informed various recent studies on digital technology and privacy (e.g., Child et al., 2009; Kang et al., 2022; Metzger, 2007). However, most of these cases are grounded in the computer-mediated communication (CMC) paradigm; as Lutz (2023) contends, a CMC perspective places its investigative locus on privacy relations either between individual users, or between the user and other stakeholders in the digital network (e.g., digital service providers). In contrast, HMC views machines as social actors that users communicate directly with, instead of as a mediator (Gunkel, 2012; Guzman, 2018). This perspective entails that user-chatbot communication poses different privacy implications from those explored in CMC studies, as it involves direct interactions with an autonomous system that functions as a conversational partner and a data collection interface.

Andrea L. Guzman (2018, p. 17) defines HMC as the "creation of meaning among humans and machines". Communication with machines as meaning-making endeavours echoes earlier scholarships (Gunkel, 2012; Reeves and Nass, 1996; Turkle, 2005) that interacting with human-like technologies is indeed a collaborative matter unfolding in situational communication contexts. Text-based communicative modalities are the primary interactive functions of AI chatbots, with whom users communicate directly through an interface using natural language (Hepp, 2020). To this extent, communication between the human user and the chatbot mimics that of interpersonal communication, as both parties occupy their legitimate spots in a two-way communication structure (Gunkel, 2012). This relational perspective inherent to HMC thinking acknowledges both the human user's active role in making sense of the technological other, and the machine's role in shaping the user's communication practices. CPM is premised on a relational view of privacy management as negotiated decisions and continual assessment of communication boundaries between partners (Petronio, 2002). The negotiated nature of privacy proposed by CPM suggests that privacy management strategies and rules to govern boundaries between closedness and openness are results of situational two-way collaborations that define these strategies and rules. This conceptual alignment between CPM and HMC, reinforces CPM's relevance to understanding how users develop and adapt privacy rules when interacting with relational machines.

Recent theoretical explorations (Spence, 2019) have proposed that human-human communication theories can offer productive jumping-off points to understand communication between human and machine. However, such a pragmatic approach is not without its risks (Guzman and Lewis, 2020); machine as a communicator is not the same as its human counterpart, as they lack clear social cues and contextual awareness. Furthermore, AI chatbots are complex automated systems of communication involving different techno-social layers (Hepp, 2020). To communicate with a

chatbot, users need to conceptualise a source which communication hinges upon (Guzman, 2019; cf. Reeves and Nass, 1996). In HMC it is not always straightforward what information sources (e.g., interface, hardware, software, developers, service providers) users orient themselves to (Solomon and Wash, 2014). This complicates the negotiation of privacy boundaries, as users' source orientation—whether toward the chatbot's interface or its broader system—shifts dynamically (Guzman, 2019). Consequently, privacy management in HMC involves user-driven and machine-augmented efforts that vary depending on which communication sources users believe they are engaging with.

Therefore, key CPM concepts such as ownership and control require explication to account for the contextual dynamics in HMC. First, CPM differentiates primary ownership and co-ownership, where privacy information becomes shared after disclosure (Petronio, 2002). However, given different orientations that may exist in user-chatbot interactions, the idea of co-ownership may be perceived differently when users' source orientation shifts. In addition, "private information changes in degrees of risk based on perceived repercussions for revealing and concealing" (Petronio, 2002, p. 67). These perceived repercussions can shift when users "peel back" the layers of the chatbot that reveal how different components—from interface to backend infrastructure—are involved in collecting, storing and processing data. For example, when the chatbot is perceived primarily as a conversational partner on screen, users may feel less risky and assume that information remains within that immediate interaction. In contrast, when the source is perceived as the service provider (e.g., OpenAI), users may feel that ownership has been transferred or diluted due to a perceived change in risk degree, leading to new expectations of co-ownership and privacy management strategies. It is also important to note that user perceptions of the source do not necessarily alter the actual parameters of ownership as defined by the technical architecture surrounding data governance, meaning that their data are still subject to broader system-level processing and retention.

Second, implicit to CPM is a relational understanding of control (Petronio, 2002; Petronio et al., 2022). Boundary coordination describes the dynamic process of negotiation between relational partners determining rules around 1) whether and who to include/exclude as information co-owners; and 2) the actual content of information divulged. In user-chatbot interactions, the relationality of control lies primarily in users' proactive attempts to manage information flow in relation to constraints or possibilities entailed by the chatbot system, rather than a clean-cut negotiation with the service provider (cf. Vitak et al., 2023). Drawing inspirations from existing studies (Metzger, 2007), chatbot users may perform a kind of "soft control" by withholding or falsifying information to obfuscate personal details (Brunton and Nissenbaum, 2015) and interfere with data collection. Moreover, having some information about the relational partner is crucial to privacy management (Petronio, 2002) as it aids assessment of the perceived consequentiality of privacy disclosure. Thus, information seeking (e.g., reviewing privacy policies and regulations) can also be a control strategy that guides boundary coordination.

The present study bridges CPM with HMC thinking, as well as updates and applies CPM's core concepts—including ownership and control—to understand the possible dynamics emerging from informational privacy management in user-chatbot communication that comprises multiple communication sources users may orient to. The empirical component of this study provides rich

user perspectives on how (co-)ownership and control are made sense of and practised, which serves to address gaps in the literature outlined below.

Existing gaps and research question

Empirical studies on chatbot and user privacy adopting a CPM perspective are relatively scarce. In a between-subject factorial design experiment, where participants were exposed to one of several chatbot conditions varying in interactivity and data-sharing protocols, Sannon et al. (2020) discover that chatbots disclosing user chatlogs to third-party advertisers elicit greater privacy concerns than those sharing data only with the service provider. Liu et al. (2023) employed a similar experimental method and find that information sensitivity moderates privacy concerns: compared to a low sensitivity condition, users asked to disclose highly sensitive information reported elevated privacy concerns and lower willingness to share. These findings support CPM's premise that users view themselves as owners of private information, and violations of user privacy expectations, especially in contexts involving sensitive data, lead to increased concerns and decreased disclosure intentions. Yet, what remains less understood is how users form and negotiate privacy boundaries in everyday interactions with chatbots, as neither study provides an in-depth account of user strategies nor meaning-making practices related to privacy management in real-world settings.

In addition, as HMC is an emerging field (Guzman and Lewis, 2020), scholars have only started to explore privacy issues through an HMC lens (e.g., Ischen et al., 2019; Lutz et al., 2019). On the topic of chatbot and privacy, Ischen et al. (2019) manipulated design choices to test user responses across 3 interface types: a human-like chatbot (with a name and social cues), a machine-like chatbot (with robotic visuals and tone), and an interactive website (with no agent presence). Their finding shows that higher perceived anthropomorphism in chatbots leads to lower privacy concerns and increased disclosure intention (see also, Lim and Shim, 2022). However, this finding sits somewhat paradoxically alongside Sundar's Machine Heuristic (Sundar and Kim, 2019), which posits that users may place greater trust in systems perceived as mechanical, believing them as more neutral and therefore safer for sensitive disclosure. This misalignment warrants further studies to disentangle disclosure intentions from actual privacy behaviours, and to explore how information disclosure is practised as part and parcel of chatbot users' relational privacy management practices.

More broadly, a recent review of conversational agents and privacy finds that much of the research focuses on how user privacy concerns influence self-disclosure to chatbots, with surveys and experimental methods—often relying on isolated variables and artificial conditions—dominating the field (Gumusel, 2024). This suggests that existing studies tend to treat privacy concerns as a static, individual-level variable, rather than as part of an ongoing process of privacy management and negotiation (Palen and Dourish, 2003; Petronio, 2002). As a result, findings are limited to quantitative insights, overlooking the situated and relational nuances of how privacy is negotiated in user–chatbot interactions. Furthermore, while these methods are valuable for hypothesis testing in controlled environments, they may lack ecological validity when applied to everyday HMC (see Spence et al., 2023), where users engage with chatbots in diverse, fluid and context-dependent ways.

Moving beyond quantitative insights and controlled conditions, the current research applies CPM's relational thinking to understand informational privacy in HMC, focusing on how users conceptualise ownership and control in their negotiated decisions around information disclosure to chatbots. It asks: **how do AI chatbot users understand and negotiate information ownership and privacy boundary control in everyday user-chatbot communication?** In addressing this question, this exploratory study contributes to the growing field of HMC and enriches existing scholarships on AI chatbot and privacy through a user-informed approach. It also provides empirical evidence to argue for the applicability of CPM in user-chatbot communication in particular and adds to our understanding of privacy disclosure and management in HMC in general.

Method

This study deploys qualitative thematic analysis (Clarke and Braun, 2017) to investigate how AI chatbot users understand informational privacy and practise privacy management strategies. Data were sourced from Reddit, a social networking platform with forums (sub-reddit) dedicated to specific topics or communities (Proferes et al., 2021). Users' active sharing of privacy-related experiences with chatbots can be seen as a form of community-driven audit that produces lay knowledge and surfaces the (in)capabilities of AI technologies in everyday contexts (Li et al., 2023, cited in Li and Zhang, 2024). A thematic analysis of such narratives contributes to uncovering detailed user perspectives surrounding privacy management in user-chatbot communication and showing how people make sense of AI chatbots in the everyday, which is key to HMC research (Guzman and Lewis, 2019).

Reddit data were chosen over direct user engagement methods (e.g., interviews) because it captures how users naturally articulate their concerns and privacy management strategies. However, it is important to note that online spaces like Reddit are socially shaped; users may tailor their posts for visibility (Shepherd, 2020). Furthermore, Reddit's user base is predominately male, skewing young (Proferes et al., 2021) and may also be over-represented by individuals with higher socio-economic status (Hargittai, 2018). Nevertheless, given the exploratory nature of this study, it is considered an acceptable trade-off. Limitations and their implications for future research are discussed in the conclusion.

Data collection

Data were retrieved via software tool Communalytic. "Privacy" was used as the keyword to retrieve relevant textual materials (called submissions[2]). Phase One was conducted in July 2024 to gather data from sub-reddit *r/ChatGPT*; a key purpose was to assess data quality and evaluate the alignment between theoretical framework and data. This phase yielded 200[3] submissions, which I read through and filtered manually, resulting in 84 relevant submissions. Irrelevant ones were excluded, such as promotional messages, news re-posts, and incomprehensive submissions. Research notes were taken to document preliminary findings. I also conducted a preliminary review of user replies associated with these filtered submissions to assess if they offered additional nuances. Findings suggested that they repeated themes present in the submissions or contained unrelated information. Therefore, replies were excluded for methodological consistency and data quality considerations. Phase Two was conducted in December 2024 to retrieve data from five

sub-reddit forums (see *Table 1* for additional details). All retrieved submissions were reviewed and filtered following the same criteria practiced in July 2024. In total, 193 submissions were included in the thematic analysis.

Table 1. Number of submissions before and after filtering

Sub-reddit name	Number of submissions retrieved	Number of submissions included in analysis
ChatGPT	200~	84
	200^	16*
ClaudeAI	129	32
perplexity_ai	28	12
GeminiAI	98	10
CharacterAI	200	39

~July 2024 dataset.
^December 2024 dataset.
*The final quantity was 100; these were then cross-checked with data from July 2024, resulting in the removal of 84 duplicates.

The 5 chatbot services were chosen for their public accessibility, popularity and active user communities[4]. As conversational systems, they represent a specific sub-set of chatbots underpinned by large language models (Guo et al., 2023) which require vast amount of data for training and iteration (Hepp, 2020). Public documents[5] show that model training draws on three main data types, including Internet content, third-party licensed datasets, and user-/crowd worker-provided information. All 5 services offer users basic privacy safeguards such as data deletion options, privacy settings, and published data policies. Limited protective measures reflect an institutional emphasis on data accessibility and value (Gorwa and Veale, 2024). In data analysis, these operational features of the selected chatbots were considered when examining how users referenced and navigated specific privacy settings and data policies in their submissions.

Procedure of analysis

To conduct the analysis, data (n=193) were compiled and uploaded into NVivo 14. Qualitative thematic analysis (Clarke and Braun, 2017) serves as a flexible methodological tool, as it facilitates both a deductive approach guided by the theoretical framework and an inductive approach to uncover emerging themes specific to the research context. First, I developed an initial coding scheme based on two sources: 1) key CPM concepts such as ownership, control, boundary coordination (Petronio, 2002) and key HMC concepts such as source orientation (Guzman, 2019); 2) notes taken during Phase One. The data were then coded iteratively through constant comparative analysis (Corbin and Strauss, 2008). This means that codes were continually revised and elaborated: new codes were added when necessary, and existing codes were refined or collapsed to address overlaps. Second, submissions containing rich, detailed descriptions of user experiences were exported into Excel for further analysis. Patterns were identified and linked to the research question.

Ethical considerations

I followed established internet research guidelines (British Psychological Society, 2021; Franzke et al., 2020) and assessed ethical issues related to Reddit data (Proferes et al., 2021). A consensus is that online platforms like Reddit are "informal spaces that users often perceive as private but may strictly speaking be publicly accessible" (Franzke et al., 2020, p. 69). Sub-reddit forums like those outlined above do not generally include sensitive information, nor do they bear significant risks of exposing vulnerable individuals or pose immediate harm towards a particular group. Given the number of submissions involved, it was not practical to gain informed consent from each user. These ethical considerations shaped my practices where several strategies were adopted to protect user privacy.

First, after data retrieval, files were downloaded and removed from Communalytic. Second, when reviewing, filtering and analysing submissions, I only looked at the titles and actual content. Any information identifiable to a user (e.g., username/Reddit ID) or a submission (e.g., URL links) was stored in a separate file. This file was used only for verification on Reddit, when submissions contained rich user perspectives and were selected for detailed analysis. Third, I used composite accounts (Markham, 2012) that blended similar statements and themes from multiple users. These accounts, designed to replace direct quotations and to prevent re-identification, are italicised in text.

Findings

Ownership boundaries and associated uncertainties

A prominent theme emerging from the data was users' sense of ownership towards their information. The scope appeared to have significant breadth, covering 3 major domains:

1) access pre-requisites like email address, date of birth, and credit card specifics.
2) tracked information like location, interaction session duration, and Internet Protocol address and other cookie-related details.
3) interaction details that users and chatbots co-create, such as chatlogs and conversation history.

Despite an overall perceived sense of ownership, users tended to express uncertainty in grappling with the extent to which private information is shared with what/whom. Some speculated that their information *might be retained on the server-side or linked to hidden identifiers*, while others feared that *uploaded content could be accessed by anyone with a URL*. These uncertainties were described as *major privacy concerns* and *security failures in the design of the systems*.

One repeated theme in relation to uncertainty of ownership was the opaque and layered nature of chatbot systems. Users raised concerns about whether their interactions with chatbots *were ephemeral*. Some questioned whether it was possible to engage with the system *without leaving a data trace*, asking if their inputs could be *excluded from training datasets*, or if the system could *remain unchanged after their sessions*. What also stands out is that some users demonstrated a notable degree of technical literacy, referencing servers, URLs and training pipelines, suggesting they were not passive users, but actively engaged with and questioned the technological structures shaping their interactions.

In these cases, it seemed that users initially set up privacy boundaries with the chatbot as an information co-owner (thus granting co-ownership), which was the immediate communication source. Data exchange and processing was deemed acceptable to the extent that information remains within the given communication context. This also helps to explain why users considered interaction details such as chatlogs and conversation history as privately owned, even though private information is not necessarily always disclosed. However, as other layers beyond the immediate source manifested (e.g., the system, the language model, the company, other third parties), users began to perceive that their information had moved beyond the original expected scope of interactions with the interlocutor. This triggered a sense of violated ownership rights—a form of boundary turbulence (Petronio, 2002)—leading to discomfort and unease.

However, not all users shared the same level of uncertainty in their understanding of ownership violations. The most notable case was—echoing existing studies (Draper and Turow, 2019; Hargittai and Marwick, 2016)—the resignation trope. These users tended to disregard the importance of data sensitivity as they felt little ability to control their own. This mentality led to lower privacy concerns and an overt focus on benefits to rationalise the lack of clarity around the system's data practices. For example, some users acknowledged privacy risks associated with chatbot user, such as *data retention and third-party access,* but they also expressed *a willingness to accept these risks in exchange for functionality or innovation*. For some, the potential of *real-time internet access* or *personalised assistance* outweighed such risks. Others normalised data sharing, comparing it to *everyday practices like location tracking or app permissions*. As one user put it, *privacy is important, but the possibilities are just too exciting to ignore*.

CPM posits that people engage in a mental risk-benefit calculus to determine the degree of privacy disclosure as an inherent part to privacy management practices (Petronio, 2002). As these cases suggest, in user-chatbot communication users may engage in tilting the balance towards benefits gained by downplaying risks, so that privacy disclosure is justified on an intrapersonal level. In this light, primary ownership becomes a personal sacrifice and obscured by the multiple layers of information exchange that a chatbot systems entails.

Negotiating control through privacy boundary making

Information control is fundamental to active privacy management practices (Altman, 1975; Palen and Dourish, 2003) and is viewed as tactics to balance the dialectical tension between openness and closedness (Petronio, 2002; Child et al., 2009). Uncertainties around ownership boundaries emerged as a key characteristic of data privacy management in user-chatbot interactions. CPM tenets suggest that risk and uncertainty perceptions contribute to amplifying such tension and subsequently motivating people to develop mitigation strategies to restore the balance. However, while uncertainty served as a motivation that prompted some users to introduce protective measures to maximise benefits gained while minimise risks of privacy loss, technological restrictions also interfered with users' information management intentions and practices. Boundary coordination in user-chatbot interactions became a negotiated effort and interplay between human and machine agency.

To start with, users engaged in information seeking as a strategy to aid disclosure decision-making, as gathering adequate information about the relational partner helps to assess risks and inform

disclosure depth (Petronio, 2002); for example, *going through privacy policies before setting up the account, for peace of mind*. In fact, privacy policies of chatbots were frequently referred to in users' articulation of privacy management, which formed part of users' knowledge base. Yet existing studies (Ragab et al., 2024) suggest the purpose of privacy policies is not always aligned with chatbot users' interests; terms and condition of data usage is left intentionally vague and open to interpretations. This observation is also evident in the current study. Some users welcomed recent improvements to privacy controls—such as *clearer opt-out options* or *data retention limits*—and expressed *a newfound willingness to use chatbots for highly specific tasks*. However, this optimism was tempered by 1) the ambiguity in *policy definitions of data collection and processing* or *incomplete explanations in FAQs*; and 2) the lack of *sufficient alternatives to opt out without giving up certain benefits*. Therefore, users were "a bit cautious of jumping in with both feet".

The proactive approach to reading privacy policies echoes CPM's concept of boundary ownership, in the sense that it involves users' sense-making of the rules and terms that govern the control and management of personal information (Petronio et al., 2022). However, users often found themselves at the mercy of intentionally vague policies, highlighting a mismatch between user expectations and system realities. This led a limited number of users to adopt protective measures ranging from the use of virtual private networks (VPN) and alternative payment methods (e.g., virtual debit cards) to active adjustments of privacy settings, use of chatbot-specific features like ChatGPT's temporary chat function and information deletion request to the organisation.

However, the effectiveness of these reported strategies was largely hindered because of system restrictions and updates, thus creating frictions in these user-initiated practices to negotiate privacy boundaries. Some users noted that opting out of data collection *came at the cost of losing core features like chat history or voice-to-voice interaction*. Others described *having to manually adjust settings for each session* – a burdensome process that *discouraged consistent privacy protection*. There was also dissatisfaction with *restrictive system-wide measures*, such as VPN blocks, which was perceived to *penalise legitimate privacy practices*.

CPM's metaphors of thick and thin boundaries (Petronio, 2002) provide the basis to understand such frictions between the user and the chatbot. Thick boundaries allow less permeability, meaning that less information is permitted to pass, whereas thin ones, with a higher degree of permeability, grant relatively easier information access. Users' tactics to manage data collection and processing could be viewed as attempts to thicken privacy boundaries by either opting out completely (e.g., adjusting privacy settings) or "confusing" the system (e.g., using VPN), which reflects a desire to control information permeability. The chatbot system, on the other hand, may be seen as thinning out the boundaries; not through negotiation with users, but through creating obstacles, limiting usability or disabling user solutions in the name of data safety. These user perspectives capture the frictional nature of privacy boundary coordination that emerges and intensifies as users practise their tactical agency while the chatbot system exerts its restrictions.

Discussion

Through a qualitative thematic analysis of user submissions from five sub-reddit forums, this study explores how AI chatbot users manage their data and negotiate communication privacy boundaries in human-machine communication. The exploration reveals that in user-chatbot communication,

privacy control is an unstable process of boundary negotiation; while some users attempt to assert ownership and protect their information, others resort to resignation or pragmatism. Users' privacy management strategies are met with system-imposed constraints, resulting in interactional frictions and privacy boundary turbulence. The study extends communication privacy research in HMC by presenting users' diverse perspectives on privacy boundary making as meaning creation between human and machine.

A key finding is users' struggles with uncertainties as they navigate information ownership. This uncertainty emerges as users orient to different communication sources, reflecting the layered communication structure of chatbot systems which distribute communicative agency across both visible and invisible components (Hepp, 2020). Upon initial encounters, users share information with their chatbot counterparts and regard data processing and storage acceptable with 'something' immediate on the other side of the interface that showcases communicative capabilities. This tendency, according to the classic Computers-Are-Social-Actors (CASA) tenet (Reeves and Nass, 1996), suggests how users readily apply social scripts to machines displaying enough social traits, such as natural language production. Building on experimental studies (Ischen et al., 2019; Lim and Shim, 2022), this orientation towards the chatbot as a responsive communicator may help to explain why some users initially disclose personal information, without considering privacy implications like information ownership violations.

The present study also builds on source orientation literature (Guzman, 2019; Solomon and Wash, 2014) and presents empirical evidence of deliberate user efforts to assess communication sources and adopt intentional approaches to privacy management with chatbots. The evidence is exemplified where users' initial orientation to the chatbot as a social actor is disrupted by uncertainties – particularly when they become aware of underlying operational layers (e.g., language model; service provider). The perceived inclusion of additional co-owners external to the initial privacy boundaries triggers a tightened desire for primary information ownership and amplifies privacy anxieties – an observation echoing Sannon et al.'s conclusion (2020).

CPM (Petronio, 2002) helps to contextualise the privacy implications of source orientation in HMC, as it provides a useful framework to understand how users' information ownership is challenged and negotiated in and through communication with chatbots of a perceived dual identity: social actor and technological assemblage. Relational partners in interpersonal settings negotiate rules regarding ownership and control of information and re-negotiate such rules to stabilise boundary turbulence when privacy breakdowns occur (Petronio et al., 2022). One's relationship with an AI chatbot—and by extension the algorithms, software, hardware, developers and the company that manages that chatbot—is structurally one-sided with limited user freedom and system-level transparency to determine the exact boundaries of data privacy. This is partially why perceived lack of control leads to privacy cynicism (Draper and Turow, 2019) and apathy in networked environments (Hargittai and Marwick, 2016). Hence unsurprisingly, to cope, some users rationalise their disclosures, downplaying privacy risks in favour of perceived benefits – a cognitive dissonance reduction strategy.

Another key finding is that users' desire to achieve relational control over private information is typified by situational tactics to regulate privacy boundaries with chatbots. CPM (Petronio, 2002)

explains that boundary thickness and thinness are determined by the degree of relational control over information flow. These user-initiated ways of boundary making showcase user obfuscation strategies (Brunton and Nissenbaum, 2015), defined as deliberate attempts to interfere with data collection, which can be seen as demonstrations of user agency to fortify boundaries by increasing thickness and thus resist unintended information flow. Yet, our contemporary digital ecosystems favour increasingly thinner boundaries to facilitate information collection, processing, and accumulation (Vitak et al., 2023). For AI technologies, data governance prioritises data accessibility and sharing, with limited platform-level guardrails for privacy invasion or user control (Gorwa and Vaele, 2024). These contradictory forces create interactional tensions between users' privacy management practices and chatbots' techno-social affordances. As Floridi (2013, p. 228) notes, "informational privacy is a function of the ontological friction in the infosphere, that is, of the forces that oppose the information flow within the space of information". This means that to enhance user privacy regarding information, ontological frictions must increase between the user and the chatbot. However, as illustrated by user vignettes in this study, the onus of introducing frictions falls on users who need to devise ways of resistance that are often countered by constant system updates to limit user control, or risk losing chatbot features.

User-chatbot communication introduces burning privacy challenges to resolve. Scholars (Natale and Depounti, 2024) have cautioned against the deceitful nature of AI chatbots, not because they are necessarily capable of deceiving users into something sinister but that their appearance as a communicator able to make sense in natural language invites social reactions from users who may feel a sense of continuity in their user-chatbot relationships. Although there is no direct proof in this study, this deception may have worked to encourage users to disclose more than they knew. From this perspective, the present study bears practical implications that can inform chatbot design practices to ensure transparency and data governance policies to serve users' interests. Designers and developers should consider including clear in-situ signposts (e.g., disclosure statement on the interface) to inform users of chatbots' role in data collection, processing and storage. Guardrails informed by the privacy-by-design principles (Cavoukian et al., 2010) can be inscribed into design choices to increase ontological frictions between the user and the chatbot, which can ease the burden of privacy management on users. As communication privacy is context-dependent and no one-time consent is adequate to ensure stable privacy boundaries (Petronio, 2002), policymakers should explore, in addition to the current informed consent framework, the feasibility of dynamic consent mechanisms (e.g., periodic re-confirmation of consent) to prevent risks of unwarranted over-disclosure from users.

Conclusion

AI technologies are increasingly becoming part of the social fabric of everyday life (Guzman and Lewis, 2020). By extending CPM to HMC, this study explores how human communication behaviours, such as the disclosure of information and the management of communication privacy, are shaped by situational interactions between users and their chatbots. With a user-centric approach, this exploration contributes to scholarship in communication privacy research in HMC (Lutz, 2023), specifies practical implications that can benefit the design of socio-technical systems, and provides an initial assessment of boundary regulations of AI chatbot data as users continue to explore these technologies. CPM's emphasis on ownership and control entails responsibility for

each relational partners involved (Petronio, 2002). To ensure the healthy and productive growth of AI that can benefit all, we must prioritise ethical AI development, establish robust data protection measures to safeguard user privacy, and hold AI systems accountable to foster informed decision-making in data-related practices.

This research has several limitations. First, it relies on Reddit data which only capture a fraction of users' experiences. As explained, the dataset was possibly over-represented by young male users. Findings also suggest a notable level of technical literacy, which is related to a higher socio-economic status (Hargittai, 2018). Furthermore, platform features like user-directed content moderation and algorithmic sorting and ranking can impact how narratives gain (in)visibility (Shepherd, 2020), which could subsequently impact the way Communalytic retrieved the data. For example, all five sub-datasets had less than half of the total retrieved submissions deemed relevant after review. Therefore, results of this study must be approached as an initial exploration and interpreted with caution. Future research is encouraged to engage human participants of diverse demographic backgrounds, obtain first-hand user perspectives of privacy management with chatbots, and identify shifts in disclosure patterns over time.

Second, this study only focuses on the informational aspect of privacy as it is most relevant to chatbot use (Lutz, 2023). The chatbots selected for the study represent only a sub-set of privately owned, publicly accessible AI technologies powered by large language models. Privacy is a complex concept irreducible to a single dimension (Solove, 2006), and different types of AI technologies entail different privacy implications in HMC (Lutz et al., 2019). For example, privacy research into social robotics needs to consider their spatial implications given its physical embodiment in domestic contexts like at home with users. Future scholarships should extend CPM to include other AI types and adopt a comparative angle to understand similarities and differences in user perceptions and privacy management behaviours.

Acknowledgements

I would like to thank the two anonymous reviewers for their thoughtful and constructive feedback, which helped to improve this article. I am also grateful to the Editorial Team for their support throughout the review process.

Special thanks to Wonsun Shin, Xin Pei and Zi Lin for their encouragement and insightful conversations throughout the development of this work.

This research was supported by the Melbourne Research Scholarship.

Notes

1. Communalytic is a no-code computational social science research tool developed by Gruzd and Mai (n.d); for more information, please visit: https://communalytic.org/frequently-asked-questions/.

2. Reddit has 5 ways to categorise submissions. Given the exploratory nature of this study, only the newest/most up-to-date submissions were retrieved for analysis. For more information on submission sorting, please visit: https://support.reddithelp.com/hc/en-us/articles/19695706914196-What-filters-and-sorts-are-available.

3. When a keyword is used, the maximum number of submissions Communalytic can retrieve is 200.

4. As of 23 January 2025, the approximate numbers of subscribers (as shown on Reddit) are 8.8 million (*r/ChatGPT*), 134 thousand (*r/ClaudeAI*), 44 thousand (*r/perplexity.ai*), 13 thousand (*r/GeminiAI*), and 2.2 million (*r/Character.AI*).

5. For more information, please refer to data and privacy policies: 1) ChatGPT: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed; 2) Claude: https://privacy.anthropic.com/en/articles/10023555-how-do-you-use-personal-data-in-model-training; 3) Perplexity: https://www.perplexity.ai/hub/technical-faq; 4) Gemini: https://cloud.google.com/gemini/docs/overview; 5) Character.AI: https://character.ai/privacy.

References

Allen, J. F. (2003). Natural Language Processing. In A. Ralston, E. D. Reilly, & D. Hemmendinger (Eds.), Encyclopedia of Computer Science (4th Edition) (pp. 1218-1222). Chichester: Wiley.

Altman, I. (1975). The environment and social behavior: Privacy, personal space, territory, crowding. Monterey, CA: Brooks/Cole Publishing.

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI Friend: How users of a social chatbot understand their human–AI friendship. Human Communication Research, 48, 404-429.

British Psychological Society. (2021). Ethics guidelines for Internet-mediated research. Leicester: British Psychological Society.

Brunton, F., & Nissenbaum, H. (2015). Obfuscation: a user's guide for privacy and protest. Cambridge, Massachusetts: The MIT Press.

Bucher, T. (2017). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. Information, Communication & Society, 20(1), 30-44.

Cavoukian, A., Taylor, S., & Abrams, M. (2010). Privacy by design: essential for organizational accountability and strong business practices. Identity in the Information Society, 3(2), 405-413.

Child, J. T., Pearson, J. C., & Petronio, S. (2009). Blogging, communication, and privacy management: development of the blogging privacy management measure. Journal of the American Society for Information Science and Technology, 60(10), 2079-2094.

Clarke, V., & Braun, V. (2017). Thematic analysis. The Journal of Positive Psychology, 12(3), 297-298.

Corbin, J., & Strauss, A. (2008). Basics of qualitative research: techniques and procedures for developing grounded theory. Los Angeles: Sage.

Draper, N. A., & Turow, J. (2019). The corporate cultivation of digital resignation. New Media & Society, 21(8), 1824-1839.

Eynon, R., & Young, E. (2021). Methodology, legend, and rhetoric: The constructions of AI by academia, industry, and policy groups for lifelong learning. Science, Technology, & Human Values, 46(1), 166–191.

Floridi, L. (2013). The ontological interpretation of informational privacy. In The ethics of information (pp. 228–260). Oxford, UK: Oxford University Press.

Franzke, A. S., Bechmann, A., Zimmer, M., & Ess, C. (2020). Internet research: Ethics guidelines 3.0. Retrieved July 2, 2024, from https://aoir.org/reports/ethics3.pdf

Gorwa, R., & Vaele, M. (2024). Moderating model marketplaces: Platform governance puzzles for AI intermediaries. Law, Innovation and Technology, 16(2), 341–391.

Gruzd, A., & Mai, P. (n.d.). Communalytic: A computational social science research tool for studying online communities and discourse. Retrieved July 2, 2024, from https://communalytic.org

Gumusel, E. (2024). A literature review of user privacy concerns in conversational chatbots: A social informatics approach—An Annual Review of Information Science and Technology (ARIST) paper. Journal of the Association for Information Science and Technology, 76(1), 121–154.

Gunkel, D. J. (2012). Communication and artificial intelligence: Opportunities and challenges for the 21st century. Communication +1, 1(1). https://doi.org/10.7275/R5QJ7F7R

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Supryadi, Yu, L., Liu, Y., Li, J., Xiong, B., & Xiong, D. (2023). Evaluating large language models: A comprehensive survey. arXiv. https://doi.org/10.48550/arXiv.2310.19736

Guzman, A. L. (2018). What is human–machine communication, anyways? In A. L. Guzman (Ed.), Human-machine communication: Rethinking communication, technology, and ourselves (pp. 1–28). New York, NY: Peter Lang.

Guzman, A. L. (2019). Voices in and of the machine: Source orientation toward mobile virtual assistants. Computers in Human Behavior, 90, 343–350.

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. New Media & Society, 22(1), 70–86.

Hargittai, E. (2018). Potential biases in big data: Omitted voices on social media. Social Science Computer Review, 1–15.

Hargittai, E., & Marwick, A. (2016). "What can I really do?" Explaining the privacy paradox with online apathy. International Journal of Communication, 10, 3737–3757.

Hepp, A. (2020). Artificial companions, social bots and work bots: Communicative robots as research objects of media and communication studies. Media, Culture & Society, 42(7–8), 1410–1426.

Ischen, C., Araujo, T., Voorveld, H., van Noort, G., & Smit, E. (2019). Privacy concerns in chatbot interactions. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, O.-C. Granmo, E. Luger, & P. B. Brandtzaeg (Eds.), Chatbot research and design (pp. 34–48). Cham, Switzerland: Springer.

Kang, H., Shin, W., & Huang, J. (2022). Teens' privacy management on video-sharing social media: The roles of perceived privacy risk and parental mediation. Internet Research, 32(1), 312–334.

Li, H., & Zhang, R. (2024). Finding love in algorithms: Deciphering the emotional contexts of close encounters with AI chatbots. Journal of Computer-Mediated Communication, 29(5). https://doi.org/10.1093/jcmc/zmae015

Lim, S., & Shim, H. (2022). No secrets between the two of us: Privacy concerns over using AI agents. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 16(4). https://doi.org/10.5817/CP2022-4-3

Liu, Y., Yan, W., Hu, B., Lin, Z., & Song, Y. (2023). Chatbots or humans? Effects of agent identity and information sensitivity on users' privacy management and behavioral intentions: A comparative experimental study between China and the United States. International Journal of Human–Computer Interaction, 40(19), 5632–5647.

Lutz, C. (2023). Privacy and human–machine communication. In A. L. Guzman, R. McEwen, & S. Jones (Eds.), The SAGE handbook of human–machine communication (pp. 310–317). London, UK: SAGE.

Lutz, C., Schöttler, M., & Hoffmann, C. P. (2019). The privacy implications of social robots: Scoping review and expert interviews. Mobile Media & Communication, 7(3), 412–434.

Markham, A. (2012). Fabrication as ethical practice. Information, Communication & Society, 15(3), 334–353.

Metzger, M. J. (2007). Communication privacy management in electronic commerce. Journal of Computer-Mediated Communication, 12(2), 335–361.

Natale, S., & Depounti, I. (2024). Artificial sociality. Human–Machine Communication, 7, 83–98.

Palen, L., & Dourish, P. (2003). Unpacking "privacy" for a networked world. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 129–136). New York, NY: ACM.

Petronio, S. (2002). Boundaries of privacy: Dialectics of disclosure. Albany, NY: State University of New York Press.

Petronio, S., Child, J. T., & Hall, R. D. (2022). Communication privacy management theory: Significance for interpersonal communication. In D. O. Braithwaite & P. Schrodt (Eds.), Engaging theories in interpersonal communication (3rd ed., pp. 314–327). New York, NY: Routledge.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. Social Media + Society, 7(2). https://doi.org/10.1177/20563051211019004

Ragab, A., Mannan, M., & Youssef, A. (2024). "Trust me over my privacy policy": Privacy discrepancies in romantic AI chatbot apps. In 2024 IEEE European Symposium on Security and Privacy Workshops (pp. 484–495). New York, NY: IEEE.

Reeves, B., & Nass, C. (1996). How people treat computers, television, and new media like real people and places. Cambridge, UK: Cambridge University Press.

Sannon, S., Stoll, B., DiFranzo, D., Jung, M. F., & Bazarova, N. N. (2020). "I just shared your responses": Extending communication privacy management theory to interactions with conversational agents. Proceedings of the ACM on Human–Computer Interaction, 4, 1–18.

Shepherd, R. P. (2020). Gaming Reddit's algorithm: r/the_donald, amplification, and the rhetoric of sorting. Computers and Composition, 56. https://doi.org/10.1016/j.compcom.2020.102572

Sher, G., & Benchlouch, A. (2023). The privacy paradox with AI. Reuters. Retrieved December 15, 2024, from https://www.reuters.com/legal/legalindustry/privacy-paradox-with-ai-2023-10-31/

Skjuve, M., Brandtzaeg, P. B., & Følstad, A. (2024). Why do people use ChatGPT? Exploring user motivations for generative conversational AI. First Monday, 29(1). https://doi.org/10.5210/fm.v29i1.13541

Solomon, J., & Wash, R. (2014). Human-what interaction? Understanding user source orientation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58(1), 422–426.

Solove, D. J. (2006). A taxonomy of privacy. University of Pennsylvania Law Review, 154(3), 477–560.

Spence, P. R. (2019). Searching for questions, original thoughts, or advancing theory: Human–machine communication. Computers in Human Behavior, 90, 285–287.

Spence, P. R., Westerman, D., & Luo, Z. (2023). Observing communication with machines. In A. L. Guzman, R. McEwen, & S. Jones (Eds.), The SAGE handbook of human–machine communication (pp. 220–227). London, UK: SAGE.

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–9).

Turkle, S. (2005). The second self: The human spirit in a computer culture. Cambridge, MA: MIT Press.

Vitak, J., Kumar, P. C., Liao, Y., & Zimmer, M. (2023). Boundary regulation processes and privacy concerns with (non-)use of voice-based assistants. Human–Machine Communication, 6, 183–201.

Westfall, C. (2023). New research shows ChatGPT reigns supreme in AI tool sector. Forbes. Retrieved January 5, 2024, from https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/

Westin, A. F. (1967). Privacy and freedom. New York, NY: Atheneum.

Juan Martín Marinangeli. Coding Trust: The Promise and Perils of Digital Transformation in Buenos Aires' AI Governance.
Universidad de San Andrés, Argentina.
Email: jmarinangeli@udesa.edu.ar

Abstract

This article examines Boti, the official chatbot of the City of Buenos Aires, as a sociotechnical intervention that reveals the political, infrastructural, and affective tensions shaping AI-driven public services. Promoted by the Government of the City of Buenos Aires (GCBA) as a symbol of digital transformation, Boti is framed as an affable, efficient, and accessible interface, seamlessly integrated into citizens' lives through WhatsApp. Drawing on the concept of technological domestication and recent literature on affective trust and platform governance, this study analyzes how the GCBA intends to construct public trust in Boti, and what that trust conceals.

While the GCBA foregrounds Boti's usability and emotional proximity, findings from audit reports, legal resolutions, interviews, and media analysis reveal a contrasting reality: weak transparency, opaque data governance, unregistered databases, and reliance on privately owned infrastructures. These tensions illustrate a central paradox: Boti fosters emotional trust through design and interface, yet lacks the institutional trustworthiness required for democratic legitimacy.

Rather than measuring user satisfaction, this paper interrogates how trust is narratively produced, institutionally unsupported, and politically consequential. It explores how Boti configures a specific type of digitally fluent 'citizen-user'; what risks emerge from platform-dependent public service models; and what institutional conditions are necessary for AI tools to enhance—not erode—democratic participation.

By situating Boti within broader trends in urban digital governance, this study contributes to critical debates on AI, trust, and citizenship, arguing that chatbots must be understood not merely as technical tools, but as political infrastructures shaped by contestable design choices.

Keywords: AI governance - Technological narratives - Citizen participation - Digital democracy - Public values - Chatbot.

Introduction

In an era marked by rapid technological advancement and the digital transformation of public services, artificial intelligence (AI) has emerged as a powerful force reshaping the interface between citizens and the state (Amodei et al., 2016; Benaich & Hogarth, 2020; Dwivedi et al., 2023). Chatbots and virtual assistants are increasingly deployed as tools to facilitate citizen interaction, reduce administrative burden, and project an image of innovation and proximity. Yet, as governments across the globe implement these systems, the promises of efficiency and accessibility (Bekkers and Homburg, 2007) often obscure deeper tensions around algorithmic opacity,

democratic accountability, and the reconfiguration of civic participation (Nemitz, 2018; Yeung & Lodge, 2019).

This is especially relevant in non-central urban contexts, where digital experimentation frequently advances faster than institutional reform. The case of Boti—the official chatbot of Buenos Aires—offers a compelling entry point to analyze these dynamics. Launched in 2019 by the Government of the City of Buenos Aires (GCBA), Boti has become the city's flagship initiative in AI-enabled public service delivery. Designed as a WhatsApp-based virtual assistant, Boti allows users to perform a range of administrative tasks, from booking appointments to requesting official documents, all through a conversational interface.

The chatbot, while "allowing several tasks to be automated through a conversational platform, either from the telephone or through a web page", acts as a digital representative of the government (Secretaría de Innovación y Transformación Digital, 2024, p.2). As the first government initiative to utilize WhatsApp as a channel for citizen interaction (Benegas, 2022), Boti exemplifies the growing trend of leveraging popular communication platforms to enhance public service delivery and citizen engagement (Androutsopoulou et al, 2019; Brandtzaeg and Følstad, 2017; van Noordt and Misuraca, 2019).

In order to understand the significance of Boti as a case of AI governance, it is essential to briefly contextualize the digital landscape of Buenos Aires. With a population of over 3.1 million inhabitants, Buenos Aires is Argentina's largest and most densely populated city (INDEC, 2025). It stands at the forefront of the country's digital infrastructure: internet penetration in households reaches 95.7%, while access to computers is at 84.1%. On a national scale, 89 out of every 100 individuals in urban areas use the internet, and 90 out of 100 use a mobile phone, although only 37% report regular use of a computer or tablet (INDEC, 2024).

This connectivity is further reflected in platform preferences. According to the GCBA, WhatsApp is installed on 92% of smartphones in Argentina and is used by 80% of mobile phone users in Buenos Aires (Secretaría de Innovación y Transformación Digital, 2024). This makes WhatsApp a near-universal interface for digital interaction. The city's level of digitalization is also internationally recognized: in 2024, Buenos Aires ranked 27th in the UN's Local Online Service Index (LOSI), placing it among the "very high" category of digitally enabled cities.

The implementation of chatbots in public administration has been driven by their potential to overcome traditional limitations of e-government initiatives. While earlier digital governance efforts often struggled with issues of integration, resource allocation, and information overload, chatbots promise to deliver more efficient, accessible, and responsive public services (Adnan et al., 2021; Souter, 2021). These AI-powered assistants can process natural language, handle complex tasks, and maintain conversations that approximate human interaction, potentially reducing administrative burden while improving communication with citizens (Adnan et al, 2021; Hoyer et al, 2020).

Boti is framed by the GCBA as a transformative tool that bridges the gap between citizens and state institutions. Its design rests on 3 core narratives: affect, through a curated personality that builds emotional trust; access, via seamless integration with WhatsApp; and efficiency, through the

automation of public services. These discursive strategies together construct Boti as an affable, effective, and inclusive mediator of civic life. However, this study reveals important discrepancies between this optimistic narrative and the institutional, legal, and infrastructural conditions underpinning the chatbot's implementation.

Our analysis approaches Boti as a sociotechnical intervention that not only mediates public service delivery but also constructs a specific vision of citizenship and trust. We combine theories of technological domestication (Silverstone & Haddon, 1996), affective trust and emotional design (Gordon & Guarna, 2022), and critiques of digital platformization (Barns, 2020; Funes, 2024), with Caputo's (2023) insights into the depoliticizing logics of automated participation. We argue that Boti fosters performative trust through interface design and affective cues, while lacking the institutional safeguards that would ensure democratic oversight, contestability, and transparency.

Our research is guided by 3 core questions: 1: How does the GCBA frame Boti as a trustworthy and transformative civic interface?; 2: What institutional tensions emerge between this narrative and the empirical realities of Boti's implementation?; 3: What broader lessons can be drawn from Boti about the role of AI in digital governance—particularly in urban contexts of the Global South?

To answer these questions, we employ a qualitative, interpretive methodology grounded in thematic analysis. Drawing on government reports, legal resolutions, civil society audits, press coverage, and a semi-structured interview with a key GCBA official, we reconstruct the GCBA's public narrative and contrast it with findings from oversight bodies and independent evaluations. Our methodological approach is discussed in detail in the next section.

The article proceeds as follows. First, we present our methodology, including the materials analyzed and our coding approach. Second, we outline the theoretical framework, integrating insights from science and technology studies and platform governance. Third, we examine how the GCBA constructs Boti as a model of digital transformation and trust. Fourth, we explore the empirical tensions between this narrative and the findings of external evaluations. Finally, the discussion and conclusion reflect on what this case reveals about the politics of trust and AI-driven governance in urban contexts.

Methodology

This study adopts a qualitative, interpretative approach grounded in science and technology studies (STS), critical discourse analysis, and digital governance research. Rather than evaluating Boti's effectiveness[1] as a technological solution, our aim is to critically examine how the chatbot is discursively constructed by the Government of the City of Buenos Aires (GCBA), and how this construction configures citizen–state relations within a broader platformization context. We focus on the public meaning-making practices that accompany technological interventions and analyze how institutional narratives about Boti frame democratic participation, trust, and civic subjectivity in the digital era.

To reconstruct the GCBA's public narrative about Boti, we conducted a qualitative review of official documents, interviews, and public statements. Our primary source was the technical report "Boti: The City's Chatbot" (in spanish, "Boti: El chatbot de la Ciudad"; Secretaría de Innovación y Transformación Digital, 2024), which presents the chatbot as a flagship initiative aimed at enhancing digital inclusion and modernizing government–citizen interaction. This was complemented by a semi-structured interview with Pedro Pérez, former Undersecretary of Smart City and one of the main architects of Boti's development, which provided insight into both the strategic vision and operational choices surrounding the chatbot. To further contextualize and update our analysis, we examined six news articles published between 2023 and 2025 across national and international outlets (e.g., Computer Weekly, La Nación, iProUP, Infobae), which include direct quotations from key GCBA officials and describe recent innovations in Boti's design and deployment.[2]

Specifically, our corpus of materials comprised eight key sources: 1: the GCBA's official technical report Boti: El chatbot de la Ciudad (Secretaría de Innovación y Transformación Digital, 2024); 2: a semi-structured interview with Pedro Pérez, former Undersecretary of Smart City; and 3: six pieces of media coverage that feature direct quotations from GCBA officials and describe innovations in Boti's design and deployment. These include Fernández (2023, Computer Weekly), Fernández (2024, iProUP), Torres (2024, Infobae), La Nación (2024), Buenos Aires Ciudad (2025), and Blasi (2025, Microsoft Customer Stories). Taken together, these materials constitute the discursive record through which the GCBA has articulated Boti's personality, accessibility, efficiency, and role in digital transformation.

Through thematic coding of these materials, we identified 4 recurring narrative dimensions. First, accessibility, which frames Boti as an intuitive and inclusive channel that "meets users where they already are," particularly on WhatsApp (Fernández, 2023; Torres, 2024). Second, efficiency, which is presented as both a bureaucratic and cultural achievement: GCBA officials describe Boti as part of a broader effort to "debureaucratize the state" and offer "quick and simple solutions" to over 1,100 public procedures (Fernández, 2023; Torres, 2024). Third, empathy and personalization, conveyed through Boti's personality design and its capacity to generate tailored interactions thanks to generative AI (Secretaría de Innovación, 2024; Blasi, 2025; La Nación, 2024). Fourth, trust, which is performatively constructed through a blend of technological sophistication and emotional warmth, positioning Boti as a dependable everyday companion capable of handling everything from health appointments to cultural recommendations (Buenos Aires Ciudad, 2025).

We conceptualize these discursive constructions as institutional narratives—that is, strategically crafted representations that aim to legitimate a technological intervention by linking it to broader public values such as modernization, participation, and proximity. These narratives are not merely informative; they perform political work. As Caputo (2023) notes, the discourse surrounding Boti interpellates citizens not as deliberative actors, but as data subjects whose preferences can be detected, anticipated, and satisfied through frictionless interfaces. By analyzing these narratives in relation to Boti's technical infrastructure and governance dynamics, our study interrogates the gap between aspiration and implementation, rhetoric and institutional practice.

Given the institutional limitations encountered during data collection—most notably, the lack of response from key officials and restricted access to internal documentation—our approach foregrounds the public discursive construction of Boti rather than its internal development processes. Only 1 first-hand interview was conducted, with Pedro Pérez, who requested anonymity. Despite these constraints, we argue that valuable analytical insight can be gained through publicly available interviews and statements issued by GCBA officials in various media outlets and institutional documents. These secondary interviews are not treated as direct testimonies about implementation, but as discursive artefacts through which the GCBA actively constructs legitimacy, narrates success, and shapes citizen expectations.

This approach aligns with interpretative traditions in STS and critical policy studies, which understand public communication not as a neutral reflection of practice, but as a constitutive element of governance itself (Hajer, 2009; Fischer, 2003). By treating these interviews and official declarations as intentional acts of meaning-making, our analysis focuses on how the state frames the role of AI in public administration, and how this framing configures citizens as particular types of users, subjects, and publics. In this sense, our research privileges the study of institutional narratives over technical audits or ethnographic access—while also incorporating independent evaluations, such as the 2023 Audit Report and the Public Defender's Office Resolution No. 2536/22, to contrast rhetoric with practice.

Our analytical strategy involved a thematic coding of the collected material—official documents, interview transcripts, public statements, and media coverage—using an inductive approach guided by the theoretical concepts discussed above. We identified recurring narrative motifs related to accessibility, efficiency, empathy, trust, and digital transformation, which we treated as entry points for deeper conceptual interpretation. The codes were interpreted in light of our theoretical framework, which draws on Caputo's concept of discursive interpellation, Liste and Sørensen's user configuration, and Silverstone's domestication model to understand how institutional narratives shape civic subjectivity in AI-mediated governance. In other words, these codes were not treated as neutral descriptors, but as manifestations of broader ideological frames about the role of AI in governance and the configuration of citizen–state interaction.

We then contrasted these institutional narratives with the empirical findings from independent audits (e.g., AGCBA 2023), civil society reports (e.g., Ferreyra, 2024), and public accountability mechanisms (e.g., Public Defender's Resolution No. 2536/22). This contrastive reading enabled us to examine the tensions between discourse and implementation, promise and infrastructure, affect and opacity. Rather than measuring narrative fidelity, our aim was to surface the performative and strategic uses of discourse in constructing institutional legitimacy and shaping citizen subjectivity.

Through this interpretative, document-based method, the study positions Boti not merely as a technological artefact, but as a discursive and institutional interface through which the GCBA enacts its vision of digital governance, configures civic participation, and negotiates trust in the age of automated public services.

Theoretical Framework: Domestication, Trust, and the Platformization of Urban AI

Our analytical framework draws from multiple traditions within science and technology studies, critical platform studies, and communication theory. We begin with the concept of technological domestication (Silverstone & Haddon, 1996), which describes how new technologies are integrated into everyday routines, habits, and emotional landscapes. In this view, users are not passive adopters, but active negotiators who interpret and embed digital tools within existing social structures. In the case of Boti, the official chatbot of Buenos Aires, this domestication involves not just familiarity and usage, but the cultivation of affective proximity and emotional trust.

Trust in AI systems, as recent studies argue, is not simply a rational evaluation of performance, but a socio-affective construct shaped by interface design, tone, and responsiveness (Brandtzaeg & Følstad, 2017; Gordon & Guarna, 2022; Shin, 2022). Chatbots are designed not only to deliver services, but to "feel" human—warm, friendly, available. In Boti's case, the GCBA crafted a deliberate personality inspired by figures such as the Dalai Lama, Marie Kondo, and Alfred Pennyworth to produce a relational interface that evokes empathy and intimacy. Through this design, trust is performatively constructed: citizens do not necessarily trust the institution, but they trust Boti. This process enacts what Gordon and Guarna (2022) describe as performative trust—a trust based on perceived affability, not procedural accountability.

Yet domestication does not occur in a vacuum. Following Liste and Sørensen (2015), we understand digital tools like Boti as instances of user configuration: the implicit and explicit ways technologies script their users. Boti does not merely assist citizens—it subtly instructs them on how to behave, what to expect, and what kinds of interactions are considered legitimate. This shaping of user subjectivity is not neutral; it encodes political choices about who counts as a citizen, how participation is structured, and what forms of feedback are acceptable.

To deepen this reading, we turn to Caputo (2023), whose discourse analysis of Boti as a tool of "citizen attention" provides a critical lens to interrogate the ideological underpinnings of this platform. Drawing on Althusserian theory and the concept of discursive formations, Caputo argues that Boti constitutes a digital device of interpellation, where citizens are hailed not as political agents, but as manageable users of pre-scripted services. Through a logic of curation, the chatbot offers a narrow, depoliticized menu of interactions that effaces deliberation and dissent. Participation is reconfigured as interaction: the citizen becomes a producer of data rather than a subject of rights or a participant in co-governance.

Caputo's analysis allows us to see how Boti enacts a form of technocratic governance, in which affective design, data capture, and automated response combine to simulate attentiveness while excluding genuine political engagement. In this model, trust is not institutional but instrumental. Citizens are represented algorithmically, their preferences mapped and processed without transparency, contestation, or reflexivity. As Caputo warns, the chatbot operates within a feedback system that naturalizes structural inequalities, offering "solutions" while masking the ideological work of framing problems in apolitical terms.

This ideological work is materially sustained by the platformization of public services (Van Dijck et al., 2018; Barns, 2020; Leszczynski, 2020). Although GCBA officials claim to be "platform-

agnostic," Boti's implementation rests on infrastructures operated by private corporations—namely WhatsApp (Meta), AWS, and Botmaker. This delegation of infrastructural control undermines democratic oversight and reinforces a mode of governance that privileges convenience and adoption over accountability and sovereignty. In such a model, proximity is simulated, but political distance is deepened.

Finally, drawing on Ananny and Crawford's (2018) work on algorithmic accountability and Eubanks's (2018) analysis of digital exclusion, we argue that trust in public AI cannot be reduced to affective design or usage metrics. It must be rooted in transparency, contestability, and meaningful inclusion. Boti offers no space for citizens to contest how services are structured, how data is used, or how priorities are set. It personalizes bureaucracy but does not democratize it. What emerges, then, is a narrow vision of participation: one that celebrates interaction while foreclosing deliberation.

By combining the lenses of domestication, user configuration, ideological critique, and platform studies, this framework illuminates the tensions embedded in Boti's implementation. It helps us interrogate not only how trust is produced, but also how power is exercised—subtly, affectively, and infrastructurally—within AI-mediated governance.

Framing Boti: Affect, Access, and the Rhetoric of Digital Transformation

Boti, the official chatbot of the Government of the City of Buenos Aires (GCBA), is widely promoted as a symbol of the city's digital transformation. Public reports, interviews with city officials, and institutional communications describe Boti not only as an administrative tool, but as a trusted companion for navigating the state. Across these materials, the GCBA constructs a multifaceted narrative that presents Boti as emotionally intelligent, ubiquitously accessible, highly efficient, and increasingly personalized through the use of artificial intelligence. This narrative, while compelling, performs a crucial rhetorical function: it positions Boti as both the face and the infrastructure of a reimagined digital state.

Affective Design and Emotional Trust. At the core of this narrative is the affective dimension. From its inception, Boti was not presented as a neutral interface but as a character—designed to be affable, trustworthy, and emotionally engaging. This emotional strategy is not incidental. As part of the government's trust-building efforts (Brandtzaeg and Følstad, 2017; Shin, 2022), Boti was imbued with a personality that blends traits from recognizable cultural figures: honesty from the Dalai Lama, decisiveness from Marie Kondo, didacticism from Merlí, and empathy from Alfred Pennyworth (Secretaría de Innovación y Transformación Digital, 2024).

The intent behind this design was to humanize bureaucratic interaction and reduce the psychological distance between citizens and government. Melisa Breda, in an interview with Gordon and Guarna (2022), noted that Boti's empathetic tone allowed people to shift from talking to abstract ministries to having a "conversation" with a relatable figure. Emotional design thus becomes a vehicle for perceived institutional proximity and a key mechanism for cultivating what we call performative trust; trust that is felt and constructed (through tone, language, and style), even in the absence of structural transparency. Boti *feels* trustworthy because it is affable.

Domestication and Strategic Ubiquity. At the heart of the GCBA's narrative is the notion of technological domestication (Silverstone and Haddon, 1996), through which Boti is framed not only as a service but as an everyday companion. The chatbot is presented as seamlessly integrated into citizens' daily routines, operating through WhatsApp, the "space where people already are," as former Undersecretary of Smart City Pedro Pérez emphasized in an interview: "The success of the product is not to force people to do what we want, but to be where citizens already are." Thus, the decision to use WhatsApp as one of digital pragmatism, not technological dependency: "The key was to be where people already spend time talking with friends, their partner, their family. And while doing that, if they need a birth certificate, they should be able to do the procedure via chat," added Pérez. In this formulation, domestication is equated with ubiquity and comfort: Boti becomes a familiar figure, one that blends public service with conversational intimacy.

This aesthetic of trust underpins a broader narrative of proximity, which presents Boti as the "first channel of contact" between citizens and the city. The COVID-19 pandemic accelerated this centrality, with Pérez noting that "what for us was a dream—being the first contact channel—became a reality." Boti is positioned not just as a technical solution but as a relational agent, able to "solve citizens' pain through conversations," a phrase that crystallizes the chatbot's emotional promise.

Efficiency and Digital Modernization. Beyond emotional design, efficiency plays a central role in the GCBA's narrative. In public statements, officials repeatedly describe Boti as a solution to state inefficiencies and bureaucratic inertia. "Today, more than 1,100 procedures can be resolved virtually, simply and quickly," which "not only saves time for the administration but also for citizens," stated Diego Fernández, Secretary of Innovation and Digital Transformation (Computer Weekly, 2023).

Efficiency, here, is framed as a win-win: the state optimizes its workflows while citizens receive faster, easier services. This logic mirrors global trends in e-government and smart city rhetoric (Androutsopoulou et al., 2019; van Noordt and Misuraca, 2019). Boti is not presented as experimental or incomplete, but as a mature, reliable interface capable of scheduling appointments, processing document requests, and even receiving complaints. As Fernández put it, "The pandemic accelerated the transformation we were already leading," positioning Boti as a key interface during health crises and as a permanent fixture in the digital delivery of services.

Boti can handle tasks such as scheduling appointments, providing information about public services, and even processing complaints—functions that would otherwise require significant human resources and time (Secretaría de Innovación y Transformación Digital, 2024). This focus on efficiency aligns with broader trends in digital governance, where AI technologies are increasingly used to optimize resource allocation and enhance service delivery (Benaich and Hogarth, 2020).

Such configuration aligns with the logic of urban platformization (Barns, 2020; Funes, 2024), where the infrastructure of public service is delegated to private platforms. In Boti's case, WhatsApp, Amazon Web Services, and Botmaker serve as its operational backbone. While Pérez

described the GCBA as "agnostic of platform," the reliance on WhatsApp was—as outlined before—strategic.

Personalization and the AI Promise. A more recent addition to the GCBA's rhetorical strategy is personalization through AI. In 2024, Boti was integrated with GPT-4o, allowing it to engage in natural language conversations and generate tailored responses. In an article published by La Nación, GCBA officials noted that this upgrade enabled Boti to go beyond rigid question-and-answer flows, adapting responses "based on individual needs and user experiences."

Julieta Rappan, Director of Digital Channels, emphasized in a Microsoft Customer Story that "generative technology allowed us to centralize government information and provide more personalized, effective experiences for citizens." These developments reinforce a discourse of smart governance, wherein AI becomes not only a backend tool but a front-facing asset capable of empathizing, adapting, and personalizing state interaction at scale.

Taken together, these discursive layers—affective trust, technological proximity, service efficiency, and algorithmic personalization—constitute what we term the GCBA's rhetoric of digital transformation. These narratives configure Boti not merely as a chatbot but as a trusted digital intermediary that mediates between citizens and the city.

Yet, as Caputo (2023) warns, these framings risk depoliticizing participation by transforming citizens into satisfied users and reducing democratic engagement to feedback loops. Boti listens, but does not deliberate. It responds, but does not reflect. In the next section, we contrast this polished narrative with independent assessments of Boti's governance infrastructure, institutional safeguards, and data practices, thus revealing the tensions between the GCBA's vision of trust and the institutional conditions necessary to sustain it.

Findings: Trust, Opacity, and the Politics of Operationalization

The promise of trust and accessibility embedded in Boti's official narrative contrasts sharply with the governance practices surrounding its development. As Pedro Pérez explained: "What we do, what we did and what we'll keep doing is to prioritize resources so that citizens can do everything through the platform that feels easiest and most comfortable. Time is today's scarcest resource, and this optimizes it." This narrative of empathy and convenience, while compelling, often conceals the fragility, opacity, and informality of the institutional infrastructure that sustains Boti.

Beyond its polished promotional narrative, Boti has been subject to only 1 comprehensive institutional review to date: the 2023 audit conducted by the General Audit Office of the City of Buenos Aires (*Auditoría General de la Ciudad*, AGCBA). This report offers critical insights into the structural, legal, and operational foundations of the chatbot and reveals serious governance deficiencies that directly challenge the official discourse of transparency and innovation.

While the audit recognizes Boti's growing role in facilitating citizen access to information and public services, it simultaneously exposes the absence of basic planning, oversight, and institutional safeguards necessary for a digital service of this scale. Among the most alarming findings is the

lack of a formalized process for evaluating Boti's operational budget or determining the eventual ownership of the platform's infrastructure and intellectual property. This means that the city government has deployed and expanded a central public service without establishing who owns its technological base or how future development and funding decisions will be made.

The audit also notes that no service-level agreements (SLAs) were submitted to define the rights, obligations, and performance expectations between the GCBA and the private companies contracted to develop and maintain Boti (primarily Botmaker S.R.L. and ASInf). SLAs are a standard tool in technology governance, designed to ensure transparency in vendor relationships, establish performance metrics (such as system uptime or response times), and define remedies in case of failure. Their absence suggests a troubling lack of contractual formality and legal safeguards, leaving the city vulnerable to technical disruptions or vendor discontinuity with no enforceable mechanisms for accountability or redress.

Compounding this concern, the audit found that the GCBA had not presented any documented plans for continuity, disaster recovery, or vendor migration in the event that the companies currently managing Boti cease operations or change contractual terms. In other words, despite the chatbot's central role in service delivery to millions of residents, there is no contingency protocol in place to ensure that Boti would remain operational if its technological providers became unavailable.

Furthermore, the report revealed that no documentation was provided to confirm the existence of confidentiality agreements or non-disclosure clauses with the developers regarding Boti's source code or internal architecture. This omission not only raises legal and security concerns, but also makes it impossible to assess whether the GCBA has retained the capacity to oversee, audit, or replicate the system independently. In the absence of such agreements, sensitive information about the system's functioning, vulnerabilities, and internal logic may lie fully outside public reach and governmental control.

Ultimately, the audit underscores a generalized lack of clarity regarding which public entities are responsible for key aspects of Boti's infrastructure and governance. The GCBA failed to delineate who within the government is tasked with ensuring data security, monitoring system performance, managing updates, or maintaining audit logs. This institutional ambiguity generates a scenario where no single office or official is clearly accountable for the platform's functioning—undermining basic principles of administrative transparency and public oversight.

These shortcomings are particularly alarming considering Boti's symbolic centrality as the GCBA's "first channel of contact". As Pérez noted, "what for us was a dream—being the first contact channel—became a reality." Yet the paradox remains: the more central Boti becomes, the less transparent and accountable its governance appears.

From the user perspective, Boti lacks standardized mechanisms for feedback, error correction, or resolution tracking. Although the GCBA claims high satisfaction rates, these are presented without clear benchmarks or transparent methodologies for measurement. According to the AGCBA, "the current situation does not guarantee the continuity and availability of Boti in case of provider

disruptions," suggesting that citizen access to services is ultimately contingent on unregulated private infrastructure.

These concerns are further reinforced by a 2024 report published by the Asociación por los Derechos Civiles (ADC) in collaboration with the regional organization Derechos Digitales, which critically examines the design, deployment, and regulatory oversight of Boti (Ferreyra, 2024). The report highlights a series of inconsistencies, omissions, and structural deficiencies that raise serious doubts about the governance model underpinning the chatbot. According to the authors, the GCBA has not clearly defined how Boti processes the personal data it collects, what technologies are involved in that process, or how the associated risks are assessed and mitigated.

One of the most pressing concerns raised in the report is the lack of clarity surrounding the collection and storage of personal and sensitive data. Although Boti is designed to interact conversationally through WhatsApp and other digital channels, it often requests identifying information such as names, national identity numbers (known in Argentina as "DNI"), addresses, and phone numbers in order to process certain public services. However, the GCBA has not provided a publicly accessible explanation of how this data is processed, where it is stored, for how long, or under what legal safeguards.

The report also criticizes the absence of comprehensive documentation on key elements of algorithmic governance. There is no public information about whether Boti's design incorporates mechanisms to prevent algorithmic bias, nor whether its conversational models are subject to regular audits or impact assessments. This is particularly problematic given that the chatbot increasingly mediates access to basic services and potentially shapes how residents perceive and interact with the state.

Moreover, the ADC notes that there is no publicly disclosed evidence of the existence of a designated Data Protection Office, nor any indication that the GCBA has conducted Data Protection Impact Assessments, which are required under many global privacy standards when processing data at scale. This absence further weakens the institutional safeguards meant to ensure accountability in the deployment of AI systems.

Finally, the report highlights troubling discrepancies between the user-facing privacy notice and the technical realities of how Boti operates. While the legal notice available to citizens outlines certain protections in generic terms, it does not align with the opaque technical processes described in internal reports. The disconnect between stated protections and actual practices signals a serious transparency gap given that not only undermines users' informed consent, but also raises concerns about the democratic legitimacy of Boti's operational model.

Further validation of these findings is provided by Resolution No. 2536/22 of the Public Defender's Office of Buenos Aires, a constitutional body tasked with protecting citizens' rights (*Defensoría del Pueblo de la Ciudad de Buenos Aires*, 2022). The resolution emerged in response to a formal complaint submitted by a resident, who reported serious concerns regarding the accessibility of personal data through the GCBA's chatbot. The investigation conducted by the Public Defender's Office offers a rare glimpse into the institutional oversight of Boti.

First, the investigation revealed that the legal notice informing users of their rights and the terms of data processing was either absent or inaccessible when accessing Boti through the government's official website. This lack of visibility violates Article 18 of Buenos Aires' Law No. 1845, which governs the protection of personal data within the city and mandates that such information be clearly communicated to users at all times. According to the law, users must be informed of the identity of the data controller, the purposes for which data is collected, the legal basis for processing, the potential recipients of the data, and the mechanisms for exercising rights such as access, correction, or deletion.

Second, the complaint highlighted that sensitive health information—specifically COVID-19 test results—could be accessed through Boti with only a DNI and a mobile phone number. Although the GCBA later responded that such access is restricted to the same device and number used during test registration, the Public Defender emphasized that no clear protocol had been published to explain how the system verifies user identity or protects against unauthorized access. This is particularly problematic, as health data is considered sensitive under both local and international data protection standards, requiring heightened security and explicit consent mechanisms.

Third, the resolution noted the absence of publicly available information regarding the responsible party for the database used by Boti, as well as the lack of clarity regarding how user consent is obtained and managed. Users are not informed of who manages their data, how to contact this entity, or how to exercise their right to rectify or delete personal information—a failure that directly contradicts the principles of legality, transparency, and informational self-determination enshrined in data protection law.

The investigation found that there was no public record confirming that the databases used by Boti—such as those linking mobile phone numbers to personal identities—had been registered with the city's official Data Protection Registry. This lack of registration is more than a technicality: it suggests a failure to comply with one of the basic administrative requirements for lawful data processing in Buenos Aires, as established by Law No. 1845 and its complementary regulations.

The resolution concludes that the GCBA must reformulate its legal notice using plain language, register all relevant databases, and ensure that sensitive data—especially health information—is adequately protected, access-restricted, and clearly governed. It also underscores the need for public accountability around how personal data is collected, processed, and shared across government systems.

Discussion: Performing Trust, Obscuring Politics

The case of Boti illustrates a striking tension between the GCBA's public narrative of trust and inclusion and the material conditions underpinning its implementation. Presented as an empathetic and reliable virtual assistant, Boti is framed by the GCBA as a transformative tool that fosters proximity, reduces bureaucratic friction, and personalizes citizen–state interaction. This carefully curated image is not incidental: it is central to the chatbot's institutional legitimacy. Drawing from our analytical framework, we interpret this as a process of technological domestication (Silverstone and Haddon, 1996; Liste and Sørensen, 2015), whereby Boti is framed as a friendly, familiar

presence in everyday digital routines, emotionally integrated into the lives of Buenos Aires residents.

The emotional design is particularly important in the context of public administration, where trust is essential for maintaining democratic legitimacy: as Aoki (2020) argues, citizens are unlikely to use AI systems if they do not trust them. The GCBA addresses this challenge by designing Boti as an affable and approachable virtual assistant. Boti's personality—honest, decisive, didactic, and empathetic—is intended to create a sense of emotional connection with users, making government interactions feel more personal and less bureaucratic.

However, the domestication of Boti is not merely a technical or aesthetic process; it is deeply ideological. Drawing on Caputo's (2023) analysis, we can see how Boti reflects broader trends in neoliberal governance, where technological solutions are framed as apolitical tools for efficiency and progress. The GCBA's narrative about Boti is performative, shaping how citizens perceive and interact with the chatbot while obscuring the structural inequalities and power dynamics that underpin its implementation. For instance, by emphasizing Boti's ability to detect "unmet demands" and provide real-time responses, the GCBA positions the chatbot as a neutral facilitator of citizen needs.

Yet, as Caputo argues, this framing is also depoliticizing. By positioning Boti as a neutral facilitator of citizen needs—detecting unmet demands, responding instantly, automating bureaucratic processes—the administration enacts a vision of participation that is data-driven but politically hollow. Citizens are configured as inputs to be processed, not as deliberative actors. Boti listens, but does not engage in political exchange; it responds, but does not reflect.

The GCBA's emphasis on efficiency and affability further underscores this depoliticization. Efficiency is a cornerstone of the GCBA's vision for Boti, with the chatbot touted as a solution to the inefficiencies of traditional government services. By automating repetitive tasks, centralizing multiple services into a single platform, and providing timely responses to citizen inquiries, Boti aims to streamline government-citizen interactions and improve the overall efficiency of public administration (Androutsopoulou et al., 2019; van Noordt and Misuraca, 2019).

The logic of convenience, however, raises concerns about technological sovereignty and democratic oversight. WhatsApp's proprietary infrastructure, coupled with the lack of transparency over how data is managed, limits public control over Boti's operations. As Van Dijck et al. (2018) and Leszczynski (2020) warn, platform-based governance often blurs lines between democratic accountability and corporate logic. The GCBA's discourse of proximity masks the reality of infrastructural dependence and opaque delegation. Platforms like WhatsApp occupy a privileged position in the digital economy, providing the infrastructure for user interactions while collecting vast amounts of data. In the case of Boti, this dynamic is particularly concerning: a tension emerges between public service and private profit, raising questions about who ultimately benefits from the chatbot's implementation.

Moreover, the GCBA often equates "digital transformation" with increased citizen inclusion and incorporation of *avant-garde* technologies like GPT-4o. While WhatsApp boasts broad adoption in Buenos Aires—as noted before—the assumption of universal access erases structural inequalities in device ownership, data connectivity, and digital literacy. As the Ada Lovelace Institute (2021) argues, inclusion cannot be reduced to access alone. Real democratic participation requires understandability, accountability, and contestability. In Boti's case, algorithmic opacity—combined with the absence of public documentation on how decisions are made, data is processed, or services are prioritized—restricts citizens' capacity to critically engage with the system. As Eubanks (2018) notes, tools like Boti risk deepening exclusion by privileging already-connected populations and marginalizing those left offline.

Boti's design reflects what Liste and Sørensen (2015) (drawing on Woolgar's (1991) framework) call user configuration: the implicit shaping of the ideal citizen. In this case, the ideal citizen is digitally fluent, emotionally responsive, efficient, and satisfied with pre-scripted, non-negotiable forms of interaction. Boti offers no mechanism for deliberation, dissent, or co-design. Participation is reduced to a feedback loop, where citizens report, and the system adapts—without ever opening up the logic of that system to public contestation. This user configuration reflects broader trends in digital governance, where citizens are increasingly framed as "users" of government services rather than active participants in democratic processes (Caputo, 2023).

The implications for AI-driven governance in the Global South are profound. In non-central countries, where institutional fragility often coexists with technological enthusiasm, discursive framings of trust can obscure power asymmetries and infrastructural dependencies. As our findings show, the GCBA's strategic use of affective narratives not only legitimates Boti's expansion, but also masks the precariousness of its backend governance. This highlights the need to evaluate AI initiatives not merely by adoption metrics or user satisfaction, but by their capacity to institutionalize trust through clear rights, protections, and participatory avenues.

It is also worth mentioning the alarming lack of transparency and evaluative rigor surrounding Boti. Several questions remain unanswered in official documentation: What defines a "conversation" in this context? Does it count as one conversation if a user exchanges multiple messages in one interaction, or are each of those messages treated as separate conversations? Is a high number of interactions evidence of effectiveness, or could it instead suggest that users are struggling to obtain the answers they seek, resulting in prolonged or repeated queries? Why are no satisfaction metrics, resolution rates, or follow-up indicators shared publicly?

Conclusion

Boti, the official chatbot of the Government of the City of Buenos Aires, offers a compelling case to examine how artificial intelligence is reshaping the interface between citizens and the state. On the surface, Boti represents a success story: millions of interactions, seamless integration into WhatsApp, and a narrative of empathy, accessibility, and user-friendliness. Through the lens of institutional discourse, trust becomes a central promise—crafted through affective design, emotional tone, and a strategic deployment on familiar platforms. This aesthetic of proximity is

not incidental; it is part of a carefully assembled narrative infrastructure meant to inspire confidence in AI-enabled public services.

Yet, as this paper has shown, such trust is largely performed, not institutionalized. Drawing on theoretical frameworks of technological domestication (Silverstone and Haddon, 1996), user configuration (Liste and Sørensen, 2015), and discursive interpellation (Caputo, 2023), we argued that Boti constructs the image of a digitally fluent, emotionally attuned citizen while minimizing the conditions for democratic deliberation and accountability. What is configured, ultimately, is not only a tool, but a user-subject that is satisfied with efficiency and proximity, but removed from processes of contestation and co-governance.

This disjuncture is sharpened by empirical evidence. Our analysis of audit reports, public resolutions, and civil society investigations reveals significant institutional blind spots: weak contractual safeguards, lack of clarity around data processing and consent, absent user feedback loops, and opacity in the governance of Boti's underlying infrastructure. The trust Boti invites is, therefore, deeply asymmetric: it relies on emotional design and symbolic closeness, while foundational mechanisms of scrutiny and transparency remain fragile or absent.

The chatbot encapsulates a broader paradox of AI adoption in the public sector: trust is performed at the front-end, but unsettled at the back-end. Boti promises inclusion, efficiency, and transparency, yet it is built upon infrastructures that are opaque, externally managed, and weakly regulated. Its success, according to officials, lies in being where people already are: on platforms like WhatsApp. But that very strategy of proximity entails a governance by convenience, where public services are layered atop commercial platforms whose logics remain outside public control. The result is a model of digital governance that appears user-centric while silently restructuring the boundaries of state responsibility and democratic engagement.

In consequence, Boti's alleged success as a trusted interface tells only part of the story. When trust is (seemingly) produced through aesthetics but not accompanied by structural safeguards, it becomes a surface-level solution that may sustain engagement but erode accountability. Understanding this disjuncture is essential for critically assessing AI technologies in the public sector, not only in Buenos Aires, but across the platformized landscapes of contemporary governance.

Methodologically, the paper is limited by restricted access to public officials and internal documentation. However, this limitation is itself symptomatic of the opacity and institutional reluctance to subject AI initiatives to public scrutiny. The reliance on second-hand interviews and public audits was not a choice of convenience but a reflection of the constraints—and the critical opportunities—of doing qualitative research on state-led technological innovation in Latin America.

Looking forward, this case invites broader comparative inquiry. What does Boti tell us about the adoption of AI tools in cities across the Global South? How can we ensure that technological domestication does not come at the cost of democratic deliberation? And most urgently: how do

we build AI systems in the public sector that are not only functional and friendly, but also contestable, transparent, and accountable?

Endnotes

1. Given the discursive and structural nature of the research questions, no quantitative analysis was conducted, and no reliable public data exists regarding Boti's long-term impact on democratic participation. The only cited metric is the number of conversations Boti handles each month—a figure that the GCBA frequently uses to demonstrate "success." However, this metric is both ambiguous and potentially misleading.

2. This temporal window was deliberately chosen to extend and complement the analysis conducted by Caputo (2023), whose examination of Boti's public narrative focused primarily on the 2019–2022 period. By concentrating on more recent developments, we aim to assess the continuity, evolution, and reframing of GCBA's institutional discourse in light of new technological integrations such as generative AI.

References

Ada Lovelace Institute. (2021). *Participatory data stewardship: A framework for involving people in the use of data* [White paper]. https://www.adalovelaceinstitute.org/wp-content/uploads/2021/11/ADA_Participatory-Data-Stewardship.pdf

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.*

Auditoría General de la Ciudad de Buenos Aires. (2023). *Informe final de auditoría, Proyecto Nº 10.22.04, CHATBOT Boti. Período 2021.*

Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., ... & Wallach, H. (2021, July). Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 368–378).

Benaich, N., & Hogarth, I. (2020). *State of AI report.* London, UK.

Berker, T., Hartmann, M., & Punie, Y. (2005). *Domestication of media and technology.* McGraw-Hill Education (UK).

Bifulco, L. (2013). Citizen participation, agency and voice. *European Journal of Social Theory, 16*(2), 174–187.

Caputo, M. (2016). ¿La nueva era de los chatbots? Apuntes acerca de las determinaciones ideológicas y discursivas del caso "Boti" en la Ciudad de Buenos Aires. *Cuadernos de H Ideas, 17*(17), e077.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the "good society": The US, EU, and UK approach. *Science and Engineering Ethics, 24*, 505–528.

Dafoe, A. (2018). *AI governance: A research agenda.* Governance of AI Program, Future of Humanity Institute, University of Oxford.

Dwivedi, Y. K., Sharma, A., Rana, N. P., Giannakis, M., Goel, P., & Dutot, V. (2023). Evolution of artificial intelligence research in *Technological Forecasting and Social Change*: Research topics, trends, and future directions. *Technological Forecasting and Social Change, 192,* 122579.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

Funes, J. M. (2024). El caso "Boti" y la plataformización de la ciudad de Buenos Aires. *Revista Argentina de Comunicación, 12*(15), 56–81.

Gordon, E., & Guarna, T. (2022). *Solving for trust: Innovations in smart urban governance* [White paper]. John S. and James L. Knight Foundation.

Liste, L., & Sørensen, K. H. (2015). Consumer, client or citizen? How Norwegian local governments domesticate website technology and configure their users. *Information, Communication & Society, 18*(7), 733–746.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180089.

Secretaría de Innovación y Transformación Digital. (2022). *Boti: El chatbot de la Ciudad.* Gobierno de la Ciudad de Buenos Aires. https://buenosaires.gob.ar/sites/default/files/2023-02/Caso%20Boti.pdf

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68).

Silverstone, R., & Haddon, L. (1996). Design and the domestication of information and communication technologies: Technical change and everyday life. In R. Mansell & R. Silverstone (Eds.), *Communication by design: The politics of information and communication technologies* (pp. 44–74). Oxford University Press.

Sørensen, K. H. (2006). Domestication: The enactment of technology. In T. Berker, M. Hartmann, Y. Punie, & K. J. Ward (Eds.), *Domestication revisited* (pp. 40–61). Open University Press.

Van Noordt, C., & Misuraca, G. (2019). New wine in old bottles: Chatbots in government—Exploring the transformative impact of chatbots in public service delivery. In *Electronic Participation: 11th IFIP WG 8.5 International Conference, ePart 2019, San Benedetto Del Tronto, Italy, September 2–4, 2019, Proceedings 11* (pp. 49–59). Springer International Publishing.

Weigl, L., Roth, T., Amard, A., & Zavolokina, L. (2024). When public values and user-centricity in e-government collide: A systematic review. *Government Information Quarterly, 41*(3), 101956.

Woolgar, S. (1991). Configuring the user: The case of usability trials. In J. Law (Ed.), *A sociology of monsters: Essays on power, technology and domination* (pp. 57–102). Routledge.

Yeung, K., & Lodge, M. (Eds.). (2019). *Algorithmic regulation.* Oxford University Press.

Etienne Malecki. Postdigital Art & Privacy. In Search of a Sensible Experience of Technology.

Erasmus School of History, Culture and Communication, Erasmus University Rotterdam.
etienne.malecki@proton.me

Abstract

Aware of increasing digital surveillance and datafication, some artists are developing innovative aesthetic practices that critically engage with the politics of technology and privacy. This article examines how a group of European multimedia artists creatively question and reshape digital tools through their work. Based on a thematic analysis of in-depth interviews, it shows how they explore technological opacity, encourage embodied and participatory experiences, and subvert dominant digital norms. The study focuses on how these artists negotiate, reconceptualize and make tangible such privacy issues through creative processes and play. Artists' playfulness often challenges surveillance norms or digital control, making "play" a potential conceptual hinge between postdigital aesthetics, privacy, and critical practice. Consequently, by focusing on artists' reflexive and critical engagement with digital media, the article positions postdigital art as a form of situated or contextual resistance, offering alternative forms of knowledge, perception and creation in an increasingly opaque and surveilled digital landscape.

Keywords: Postdigital Art, Surveillance, Privacy, Aesthetics,

Introduction

In recent years, artists working at the intersection of digital technologies and media practices have increasingly developed strategies for making accessible data collection infrastructures, algorithmic biases, and intrusive surveillance. This article explores how artists associated with what is increasingly referred to as postdigital art address issues of digital surveillance and transparency as well as technological dynamics. Their creative practice is characterised by being socially engaged, critically and reflexively exploring the relationship between humans and technologies (Vlavo, 2017). While its roots can be traced to earlier forms of media art, hacktivism, and tactical aesthetics, postdigital art is distinct in its attention to the entanglement of physical and digital materialities, and in its orientation toward embodied participation, hybrid environments, and open critique of technological progress narratives (Paul, 2020; Berry & Dieter, 2015). Instead of producing digital art as an autonomous aesthetic, these artists work across media to explore the political and sensory dimensions of our relationship to the digital. In this sense, this research shows that postdigital art

shares affinities with relational aesthetics and participatory art in its emphasis on interaction, embodiment, and co-creation (Bishop, 2012; Bourriaud, 1998). This research demonstrate that such aesthetics resonates with contextual theories of privacy, which argue that data sharing issues, or hyperconnectivity per example, must be understood in relation to the social norms, expectations, and power dynamics that govern specific contexts (Nissembaum, 2004; Richards, 2021). By creating works that challenge default digital behaviors and invite situated reflection, these artists offer new ways of navigating the relational boundaries of digital interactions.

On the methodology side, this article draws on interviews with twelve European artists whose work explicitly engages with digital privacy, surveillance, and data. Instead of presenting a generalized account of digital privacy, the paper focuses on how these artists experience, frame, and intervene in privacy concerns through their aesthetic and conceptual choices. In doing so, postdigital aesthetic connects the audience to broader debates in surveillance issues and privacy reflections. In addition, the research was guided by the *CreaTures* framework (Vervoort, et al. 2024), an EU-funded research project (2020–2024) that investigates how creative practices can contribute to ecological and societal transformation. Developed by a multidisciplinary team across Europe, the *CreaTures* framework *(Creative Practices for Transformational Futures)* provides tools and methods for evaluating the impact of art and design practices in fostering social, political, and environmental change. At the heart of the framework is the notion that transformative change is not only political or technological, but also cultural and experiential. The project emphasizes the unique role of creative practitioners in imagining, prefiguring, and enacting alternatives to dominant systems, a perspective that closely aligns with postdigital art.

Indeed, the framework outlines nine dimensions of practice across imagination, embodiment, care, collectivity, reframing, and sense-making, among others. These dimensions served as interpretive lenses during the analysis, helping to contextualize how artists described their process in relation to issues such as surveillance, datafication, and hyperconnectivity. Instead of applying the *CreaTures* framework as a rigid checklist, it was used as a flexible guide to interpret the interview data. This approach helped identify themes such as embodiment, participatory art, hybridity, or imagination as ways to understand how creative practices act as forms of cultural and political resistance. Three overarching dimensions of practice emerged:

1. Exploring the possibilities of combining scientific research with new imaginaries and hybrid environments.

2. Changing the audience's relationship with technology by creating a more human, participative and playful experience.

3. Challenging current narratives on technologies by opening and subverting the "black box".

Likewise, it is important to note that throughout these dimensions, playfulness emerged as a central and often underestimated element, both as a means of engaging audiences and as a critical tool for navigating the complexities of digital tools. Across interviews, artists frequently described their use of play, humor, and metaphor as essential to engaging audiences in complex themes such as privacy, autonomy, and algorithmic control. While often overlooked in tech-critical discourse, play has a deep history in both media studies and art theory. As Dale Leorke (2018) shows in *Location-Based Gaming*, play in public space often operates as a form of informal resistance, inviting people to reimagine systems and rules. In the context of post-digital art, playfulness functions as a design principle, a method of interpretation and a relational strategy between the public and issues of privacy. It allows artists to transform digital complexity into creative environments, to embed critique within interaction, and to foster what philosopher Miguel Sicart (2014) calls "playful subversion." Importantly, play here is an embodied means of resistance, one that leverages surprise, friction, and co-creation to surface new possibilities.

More than a single theory of post-digital art, the article offers an in-depth reflection on how artists are generating new ways of seeing, feeling and reflecting the dynamics and infrastructures that are shaping society's digital transformation.

Postdigital Art in Context

While digital media art has a long history, extending from Futurism and Constructivism to the experimental work of Nam June Paik and tactical media in the 1990s, *postdigital art* signals a shift in how artists relate to technology. More than simply using digital tools for exploring digital aesthetic landscape, postdigital artists reflexively engage with the socio-technical infrastructures that shape our lives. This create works that not only use technology but critically reveal and reconfigure it. In fact, contemporary postdigital artworks and born-digital arts such as immersions, simulations and augmented realities represent new challenges for established cultural institutions as well as for the public, as the individual's experience is transformed (Giannini and Bowen, 2019). Using immersive, interactive, sensitive, connective, and tactile technologies, postdigital art aim to create a more intimate and personal experience for the individual bodies and the audience (Langdon, 2014). Some interpret this phenomenon as contributing to the "humanization of digital

technologies" (Edmundson, 2015). This opens the door to new ways of curating and especially dealing with topics that previously could not be represented by other mediums (Zuanni, 2021). As Christiane Paul (2020) notes, postdigital practices often foreground digital materiality itself, exposing algorithms, network protocols, and sensor environments as sites of meaning, struggle, and imagination: "the *embeddedness* of the digital in the objects, images, and structures we encounter daily and the way we understand ourselves about them". This paper adopts the following working definition:

> *Postdigital art* is a socially engaged and reflexive practice that explores the material, political, and affective dimensions of human-technology relationships through hybrid, often participatory, forms.

This definition not only builds on the work of Paul (2020) but also reflects the self-understanding of the artists interviewed in this study, many of whom resist categorization and instead define themselves through process, experimentation, and critical engagement.

In addition, art has always played a crucial role in this cultural politics. The field of surveillance art, particularly, includes practices that make surveillance visible, challenge asymmetries of control, or creatively reframe data collection as a participatory or subversive act. Artists such as Hasan Elahi, Trevor Paglen, and the collective *!Mediengruppe Bitnik* have developed projects that highlight the aesthetics and affects of surveillance. Scholars like Clare Birchall (2011) have also drawn attention to the concept of "tactical opacity" in art, a way of resisting datafication not through transparency, but through ambiguity, refusal, or misdirection.

The artists in this study align with the tradition of critical media and surveillance art, which seeks to expose the mechanisms of control embedded in digital systems (Monahan, 2006). However, their work departs from earlier forms of critique that rely primarily on representing surveillance, these artists embed critique within the interactive, material, and immersive dimensions of their work. By crafting participatory installations, interactive workshops, and playful interfaces, these artists stage encounters that make users feel surveillance as embodied constraint, friction, or behavioral manipulation. Such works challenge the logics of seamless UX design, and instead foreground discomfort, ambiguity, and agency as tools of subversion (Paul, 2020; Birchall, 2011).

This shift from representation to immersion is particularly relevant in an era where surveillance is increasingly experiential, participatory, and internalized (Lyon, 2018). Beyond making surveillance visible, these artists create experiences helping participants rehearse alternative forms of agency and relationality within digital tools. Furthermore, Paul (2020) suggests three ways in which this

new aesthetic can be seen as revealing or reflecting the intersections between digital technologies and physical materiality:

1. Using integrated networked technologies, reflecting the human and non-human environment around them.

2. Revealing their own coded materiality as part of their form, becoming themselves a residue of digital processes.

3. Reflecting the way machines and digital processes perceive us and our world.

In these terms, the research suggests that artists act as mediators or facilitators between what is widespread and internalized as the degree of surveillance and privacy in our society and by each individual, and the openness to reflection on this situation through immersion, play and participatory art. This is what the research calls the phenomenon of reflexivity.

Surveillance, Privacy, and Creative Practice

Contemporary concerns around digital privacy are frequently framed through the lens of surveillance capitalism (Zuboff, 2019), in which user data is extracted and monetized by opaque platforms and infrastructures. While Zuboff's work has helped popularize a critique of data commodification, it is just one perspective within a broader and more nuanced field of surveillance studies. Scholars such as David Lyon (2001, 2018) and Elise Morrison (2016) emphasize the cultural and spatial dimensions of surveillance, including how it is represented, normalized, and contested in everyday life.

Plus, academic research has also shown that privacy is fundamentally more akin to power than something to hide. That it is, in fact, above all a contextual and relational process, deeply dependent on how, where, and by whom information is accessed or disclosed (Nissenbaum, 2004; Richards, 2021). When digital tools ignore these contextual boundaries, blurring private and public spheres across platforms and interactions, they threaten individual autonomy.

Historically, the recognition of privacy as a right led to a complex interplay of power, technology, liberty, agency, identity, surveillance, and autonomy between the state and individuals. The focus always was on finding a balance between power and privacy in a society continuously transformed by technologies (Keulen and Kroeze, 2018). The collection of data, design of infrastructures, and creation of connective interfaces are shaped by powerful actors, including governments and technology corporations (Johnson & Acemoglu, 2023). These platforms often prioritize profit, optimization, and behavioral prediction over transparency, accountability, or user agency. The

result is described, among scholars as well as artists, as a black box in which human experience is rendered into data flows, collected, commodified, and manipulated for strategic ends.

A review of academic literature on privacy in the digital age often converges around three major concerns: the use and manipulation of human information (personal and big data), the expansion of intrusive surveillance techniques, and the social and psychological consequences of hyperconnectivity. Artists engage with these concerns as their practices address the very dynamics that undermine contextual privacy. Through speculative design, participatory and immersive installations, interactive workshops, and playful experimentation, they engage audiences in rethinking their relationships to data, surveillance, and digital agency. Thus, they render the black box visible, felt, tested, and negotiated in artistic context. In doing so, they contribute to a growing cultural effort to reclaim agency and reimagine how privacy and power are shaped and influenced each other in digital environments. The following sections explore how this critical creativity unfolds in practice, focusing on three interwoven processes: exploring new imaginaries and environments, creating a more human-centred experience, and challenging the current status quo.

Methodology

This study employed in-depth, semi-structured interviews to explore how postdigital artists engage with issues of privacy, surveillance, and digital agency through their creative practice. In-depth interviews were chosen because they are especially suited to understanding complex, experiential, and reflexive processes, in this case, how artists conceptualize and materialize digital resistance through aesthetic strategies, design decisions, and participatory environments.

A purposive sampling strategy was used to identify twelve artists who met two core criteria: (1) they work primarily with digital media and have created at least one artwork that explicitly addresses themes of privacy, surveillance, or hyperconnectivity; and (2) they have exhibited or participated in at least one residency in Europe focused on the societal impacts of digital technology. While only one respondent explicitly used the term *postdigital* to describe their practice, all artists demonstrated a critical and reflexive engagement with digital tools consistent with the working definition adopted in this study. Interviews were conducted in 2024, either online or in person, and generated over ten hours of audio-recorded material. Interview questions were loosely structured around four areas: (1) the artist's relationship with digital media; (2) the conceptual development of recent works; (3) the role of participation, play, and embodiment; and (4) the political and ethical concerns motivating their practice. This format allowed artists to reflect on

both their conceptual intentions and material methods, while also leaving room for unexpected insights and divergent framings.

The data was analyzed using a combination of direct content analysis and thematic coding inspired by the CreaTures framework (Vervoort et al., 2024), which provides a set of dimensions for evaluating how creative practices contribute to societal transformation. As mentioned, this interdisciplinary tool proved useful in identifying how artistic practices move beyond critique to foster new imaginaries, relationships, and forms of engagement with technology. Initial coding was open-ended, allowing themes to emerge inductively from the transcripts. Over time, a more structured code tree was developed, revealing three recurring and interconnected processes in the artists' creative practice:

1. Exploring technological tools through coding, research, and interdisciplinary collaboration.
2. Designing embodied and participatory experiences that foreground play, friction, and human agency.
3. Challenging the technological status quo by subverting dominant narratives and creating alternatives.

These categories became the foundation for the analytical sections that follow. Importantly, they were not imposed in advance but emerged through iterative engagement with the data, a process loosely aligned with grounded theory methods (Charmaz, 2006). This inductive approach helped ensure that the theoretical lens remained responsive to the artists' own vocabularies, priorities, and forms of critique.

Finally, while the term *postdigital* was not universally adopted by participants, their resistance to fixed labels reflects the experimental and hybrid nature of their work. This methodological openness was crucial in allowing the study to trace shared strategies and concerns without flattening their diversity. However, several limitations remain. First, the study is geographically bound to Europe and shaped by its specific legal and cultural frameworks. Second, while the CreaTures tool helped foreground social transformation, the study did not include direct audience evaluation or long-term impact analysis, important areas for future research. Despite these limitations, the methodological approach enabled a rich exploration of how artists themselves conceptualize and enact privacy, play, and critique through aesthetic means. The next section presents the findings in detail, structured around the three central dimensions of practice identified above.

Exploring Technological Tools: Opening the Black Box

A central thread across all interviews was a commitment to *opening up* the hidden structures and logics of digital tools. tech industries design new software to gain access to more data and increase user activity, which in turn enables them to make a profit by selling this information to other companies or placing targeted ads on the platform (Hartzog, 2018; Richards, 2021). Driven by purely economic interests, the design of technologies not only puts users on the back foot but forces them to resign themselves to the opacity of what tech industries call "progress". For many artists, this situation meant engaging not only with conceptual critiques of surveillance and control, but with the technical materiality of code, software, and infrastructure. Their work reflects a sustained effort to make the "black box of technology more malleable and imaginable. Artists described their creative process as both a form of research and a creative reconfiguration of those tools. This process is deeply interdisciplinary, often combining informatics, critical theory, and participatory design. As one artist put it:

> "By avoiding licensed programs, I started using either open-source or just learning how to code, learning the technique rather than the tool. It is not something that you don't control, you can't shape or customize anymore."

For these artists, learning to code is not simply about technical skill; it is a way of reclaiming agency in a system that is often designed to obscure its own operations. Their engagement with open-source tools, self-taught programming, and collaborative experimentation reflects what Morrison (2016) calls a strategy of *critical re-mediation*: using technology against its own tendencies.

Additionally, several respondents emphasized the importance of collaboration and collective learning in this exploration. Respondents' enthusiasm of interdisciplinary approach is explained by their aiming to demystify the complexity of digital tools, which often demand a multidisciplinary knowledge. Interdisciplinary projects, studio discussions, and informal exchanges were described as key to demystifying complex systems. One artist explained:

> "It is also about collective organization, creating a space together. It gives rise to discussions with people from my studio or my collective."

This emphasis on shared learning reflects not just a practical need but an aesthetic and political orientation, one that resists the individualized, privatized experience of mainstream digital tools. It echoes earlier traditions of tactical media and open tech activism, but with a more speculative and imaginative dimension. Indeed, the act of imagining new digital environments was seen as

equally important as many artists described their use of speculative design and future scenarios as ways to provoke critical reflection. As one participant put it:

> "Imagining pessimistic futures, making it tangible or helping people imagine a future where things could go bad, that is how you can get them thinking about what is wrong with the society right now."

This combination of rigorous inquiry and playful exploration allows artists to explore alternatives to dominant techno-optimistic narratives. Importantly, their work is grounded in present conditions: in privacy regulation, algorithmic bias, platform dependency, and design asymmetries. In this context, playfulness also emerged as a significant exploratory tool. Artists described the fun of experimentation not as a superficial byproduct, but as a method for testing limits, generating surprise, and making complexity accessible. One respondent described their approach as:

> "There is a lot of playfulness for sure, as in playing, failing with the tools."

This resonates with Sicart's (2014) notion of play as subversion: a way of interacting with tools that reveals their contingencies and vulnerabilities. For post-digital artists, play enables a freer engagement with tools. Hence, by treating technological exploration as both rigorous research and artistic exploration, artists unsettle the assumption that digital tools are fixed or that their black box is inevitable. Unlike the rigour of the research process, imagining new technological avenues allows artists greater freedom when exploring technologies. As one respondent expressed:

> "Sometimes this box does not offer me enough freedom where I am happy to move to the artistic sense where I let go of things."

Their creative and critical practices do not simply expose the black box, they imagine what lies beyond it. These aspect serves as a crucial starting point for the analysis, as the position of most of the artists interviewed has developed around a relationship of curiosity, play and research around technologies. First findings show that multidisciplinary research, demystifying tools - such as learning how to code-, playfulness, and imagining new futures and environments are key insights into the artists' attitudes toward their exploration of technology.

Designing More Human-Centred Experiences

If exploring technology meant demystifying tools, imagining new environments, thus playing with the black box, the second key practice among respondents was the design of embodied, hybrid, and participatory experiences that invite users to feel and reflect on the human-technology interactions. In doing so, these artists do not just critique surveillance, hyperconnectivity, or

datafication; they create situations in which audiences can encounter and rehearse other relationships with technology. Rather than reinforcing the screen-based norms of interaction that dominate digital interaction, artists in this study consistently sought to center the body in their installations, workshops, and immersive environments. One artist described the intention behind her design in these terms:

> "It is not through a screen. It is not through your phone. It is not through text. It is not through notification. So how can we put the body in different experiences so that they can absorb, understand, or interact with information in a way that is different?

This emphasis on embodiment aligns with postdigital aesthetics that resist seamless, invisible, or frictionless tech design (Paul, 2020). Instead, these artists insert friction and imperfection into their works to foreground choice, constraint, and reflection. For most artists, offering friction within an embodied experience would even deepen the reflexive aesthetic of their art. The whole idea behind this attitude is not to control everything, but to leave the door open to unpredictability, play and randomness, as well as to increase the user's agency in their used of digital tools.

When viewed through the lens of privacy, these creative practices foster a more nuanced and engaged dialogue around surveillance and datafication. Postdigital artists confront the widespread and often dismissive attitude encapsulated in the phrase *"I have nothing to hide"*, a position that frequently leads to privacy fatigue or the belief that privacy is already lost and therefore irrelevant (Solove, 2010; Choi & Jung, 2018). Rather than accepting this resignation, their work reopens the debate by creating experiences that make surveillance personal, perceptible, and negotiable. In doing so, they resist the apathy of *"so why should we care?"* and instead frame privacy as a matter of power, context, and human agency, issues that remain deeply relevant in an age of digital abstraction and algorithmic control.

In addition, artists acknowledged the fact that design influences human behaviour, and thus, used this approach through various strategies to affect the audience relationship with technology. Therefore, this design philosophy directly challenges dominant HCI and UX paradigms that treat smoothness and efficiency as optimal. As Christian Paul would put it, postdigital art changes user's experience by contributing to reflect "the human and non-human environment around them." (Paul, 2020). More critically, drawing on what Morrison (2016) calls critical discomfort, these artists make space for hesitation, and interruption; conditions that allow audience, and thus users, to become more aware of how digital tools shape their behavior and decisions. Several artists

described these embodied experiences as a way of reclaiming agency, by giving audiences opportunities to co-create, respond, and experiment. One respondent explained:

> "How do I want agency and autonomy and how do I want it in my routine? And if there is no friction at all, then there is no way of reflecting on how it is situated in my routine."

This was especially apparent in participatory formats such as workshops, AR experiences, and interactive installations. Giving them back their power also means making them aware of the choices made without their knowledge in the privacy and default settings, as well as how, for example, cookies. This participatory impulse often takes material form in curated spaces that blend physical and digital interaction. Several artists reported designing installations where visitors are required to make decisions, perform tasks, or follow alternate rules. As one artist put it:

> "They are ways of sharing my research with the public and also inviting them into my research, my practice as well as into the discussion."

In a privacy perspective, artists create experiences that subtly mimic or expose the logic of surveillance infrastructures and behavioral design. These setups encourage what Birchall (2011) might call "tactical opacity": a form of user resistance not through transparency, but through awareness, refusal, or playful subversion. Interestingly, the works also foreground care and trust. Playful context acts as a safe environment for participants, insofar as the artists are motivated to share and create a participative and caring experience of technology. Play creates an experimental environment, a kind of safe laboratory for both artists and participants. One respondent noted:

> "Play influences the audience to feel more open to experiment, to try things that they wouldn't otherwise do."

This balance between critical engagement and emotional openness is one of the most distinctive features of the artists' practice. By designing for touch, friction, and shared experience, they make the politics of digital black box felt, and not only understood. The scientific and creative enthusiasm in their creative practice is also one of the search engines for many of the respondents to immerse themselves in new subjects. For example, one respondent was invited to take part in an exhibition on the Olympic Games 2024:

> "I didn't have a project on that at all. I had to do a new project, a project around the new algorithmic video surveillance".

For many of the artists interviewed, postdigital practice provides a space to engage with pressing issues of digital governance, including privacy, surveillance, datafication, and hyperconnectivity.

Their participatory works do not merely represent these issues; they perform alternative digital relations, characterized by friction, unpredictability, randomness, and a renewed sense of humanity. These aesthetic choices challenge the smoothness and opacity of mainstream digital design, and instead foreground vulnerability, trust, and contextual nuance. This orientation resonates with contextual theories of privacy, which emphasize that information sharing is not universally acceptable, but deeply dependent on social settings and relational boundaries (Strahilevitz, 2005; Richards, 2021). As Richards puts it, "our decision to share information in one context doesn't mean that we should share it in all contexts." By crafting intimate, tactile, or disruptive experiences, artists offer audiences new cultural reference points, or sensible landmarks, for engaging with digital tools. In doing so, they may help shift how people perceive the meaning and consequences of privacy in the networked age.

The next section describes major issues that this creative practice aims to address i.e. the social and common experience of technology by emphasising the users' autonomy, agency, and awareness, vis-a-vis the big tech and digital governance. And finally, to rebalance the current privacy, technology and power dynamics in favour of democratic process.

Challenging the Technological Status Quo

While exploring technologies and designing embodied experiences were central to the artists' practices, this research shows that their aim often extended further: to challenge the current techno-optimists' narratives and ideologies underpinning the dominant evolution of digital technology. Across the interviews, artists expressed a desire not only to reclaim agency, but to destabilize default norms, and propose alternatives to extractive, manipulative digital environments. On one hand, they contribute to creating a more open technological environment by creating open-source tools and by opening conversations on the black box and its dark patterns. On the other hand, artists are subverting the evolution of technology by creating alternative languages, exposing new rules, and raising awareness about the current status quo.

Many respondents framed their work in opposition to what one called the "relentless pursuit of efficiency" and data-driven design paradigms embedded in platforms. They criticized the economic logics of surveillance capitalism (Zuboff, 2019), not simply as abstract concerns but as material realities encoded into everyday tools and interfaces. One artist explained:

> "We don't know who owns it. We don't know the impact. Everything is magnified by the distance."

This distance, from code, from governance, from big tech, was a recurring motif. Several artists described their creative work as an attempt to shorten that distance and reveal the stakes of default settings, opaque algorithms, and manipulative nudging techniques. Rather than merely critique these tools, artists often engaged in tactical subversion. Some designed artworks that mimicked or distorted surveillance logics; others rewrote user agreements or created poetic interfaces that defied optimization. These strategies reflect what Leorke (2018) and Morrison (2016) describe as aesthetic resistance through misdirection and rule-bending. Indeed, for respondents, rules are put in place to help change the participant's experience, and the gaming environment facilitate the integration of complex subjects. Play as accessibility and as rules are strongly linked. As one respondent explained:

> "This is a way of guiding a person through a complicated topic and letting them experience it. Then [the audience] can reflect on their own choices of behaviour that were, of course, influenced by me."

Therefore, the notion of playfulness remained central to this effort. Artists used it not just as an access point, but as a political design choice, to transform experiences that rely on compliance into spaces for experimentation. Here, rules and play become tools of mutual reflection, rather than unilateral control. A bridge can be made with the literature on contextual privacy (Strahilevitz 2005): play creates an extraordinary experience for experimenting and thus redefining contextual relationships between the audience and issues of privacy. By playing with the rules, artists ensure that they create a reflexive environment, moving away from opaque digital curtains. Another parallel can be made between play and Torin Monahan notion of "defamiliarization", which explain that tactics are used "to draw critical attention to everyday surveillance that has become mundane". Thus, play would seem to be an important lever for post-digital's artists: it allows embodiment of current surveillance and datafication issues, as another interviewee explained:

> "Games turn information into a pedagogical process that enables embodied knowledge."

By using metaphor, simulation, and open-ended interaction, playful experiences invite participants to question the status quo of technological development. In these contexts, play is not simply entertainment, it becomes a strategy for destabilizing norms, allowing audiences to step into unfamiliar roles, rules, and relational dynamics. Importantly, these experiences are often designed to feel intimate, experimental, or even subversive. They rely on a tacit social contract: participants must trust that what unfolds within the installation remains protected within that space. In this

way, play becomes not just a design choice, but a framing device that temporarily redefines privacy, enabling participants to explore vulnerability and agency in a safe, bounded context.

Furthermore, by revealing how systems shape choice, and how they could be otherwise, artists unsettle the default and propose alternatives. However, creating accessible, playful and open-source digital tools is essential for artists in their explorations to push their boundaries and understand them. But it is no easy task, and respondents are often, if not always, confronted with the thick and opaque digital curtains. Many suggested that opening access to digital tools should take on the form of political regulation. However, one respondent expressed his doubt in these words:

> "I don't have a lot of belief that regulations will be our answer to defining those boundaries for the use of technology. I think regulations will help but regulations can also just be swayed by money or personal interest for power."

Confronting to this situation, most of the respondent are subverting and regaining empowerment by stopping letting themselves be manipulated and dictated to by tech industries' interests. To do so, respondents pointed that technologies influence not only our behaviour but also our language. As one respondent said:

> "Suddenly our language itself is sort of shaped by the tools we use, because otherwise, the AI can't understand it."

Language is therefore not a neutral medium; it is a terrain where power is negotiated. By designing alternative scripts, gestures, and symbolic systems, they attempt to remake the grammar of human-technology interaction itself. Another respondent expressed the same feeling of being surpassed by large language models (LLM):

> "Big question mark about AI. We get emotionally dependent on AI. We are talking with sort of mirrors of ourselves".

For all these respondents, as things stand, technological advances tend to develop a design that makes us forget that the actual digital mirror in front of our eyes is nothing more than a tinted window serving as a tool for economic profit and surveillance that threatens our privacy and democracy. Our self-image, and even self-esteem, are increasingly dependent on and are made through this mirror, which may favour certain visions and values (magnification) and diminish others (narrowing). The situation is even worst, as one respondent added:

> "We are intimately susceptible to its updates."

For the artists, working with alternative languages means asking what happens if we change the design and parameters of this mirror. In sum, they have placed their hope in creative action, in designing alternatives that are open, shareable and based on caring rather than capture. Their subversions are not about overturning platforms in a single act but about altering the relationship between human and technology and redistributing autonomy in digital environments that seem increasingly deterministic. In this way, post-digital art becomes not just a discourse on technology, but a field of intervention, a space where agency is reclaimed, tools are opened up and futures are democratic.

In contrast to the techno-optimistic narratives promoted by those who control the direction of technological development, postdigital artists adopt a critically engaged stance that links aesthetic decisions to social impact. Their work resists passive consumption and instead foregrounds the political dimensions of code, language, and design. Through practices such as speculative design, creative coding, and the invention of alternative languages, these artists develop forms of expression that render the social consequences of technology both visible and graspable.

This resonates with Clare Birchall's (2015) concept of the *aesthetics of the secret*, which reframes secrecy not as a problem to be solved, but as a productive space for political and aesthetic engagement. Rather than striving for total transparency, these artists, like those Birchall discusses, often embrace opacity, ambiguity, and play as forms of resistance, creating experiential encounters that challenge the logic of surveillance without reproducing its visual or epistemic control. In doing so, they help shift the conversation on privacy away from exposure alone and toward the creation of alternative relations to visibility, vulnerability, and digital power. Hence, their conceptual choices are deliberate interventions aimed at exposing how digital infrastructures shape experience, behavior, and power relations within those dynamics.

Conclusion: Postdigital Art as Situated Resistance

The research revealed that postdigital art goes beyond a merely political or activist stance on privacy issues and represents a valuable ally for the design of a more democratic and human digital environment. It has explored how postdigital artists engage with the politics of privacy, surveillance, and digital tools through creative practice. Drawing on interviews with ten multimedia artists based in Europe, the research has highlighted three interwoven dimensions of their work: exploring technological tools, designing more human-centred, embodied and participatory experiences, and challenging the technological status quo through many levers. The artists interviewed approach it

as a felt, contextual, and relational concern, rather than treating digital privacy as an abstract legal or technical issue. Through coding, speculative design, open-source practices, and playful installations, they intervene in tools that typically obscure user agency and reinforce behavioral conformity. Their creative strategies, especially the use of play, friction, participatory art and embodiment, resist the seamlessness of platform design and instead foreground complexity, ambiguity, and negotiation. One might ask what would the human-technology relationship look like if access was open and less profit-driven? If it didn't present a design asymmetrically thought out to ensure profit and perpetuate the status quo about the actual trajectory but rather increasing human sensitivity towards their environment and themselves? It is in addressing these questions and exposing them to the public that this creative practice could well be a form of postdigital artivism.

However, this practice does not offer a singular solution to surveillance capitalism or digital disempowerment. It proposes a different way of being with and thinking through technology, one that is rooted in scientific rigour, creativity, and play. In this sense, postdigital art constitutes a form of situated resistance: a way of reopening closed tools, revealing their politics, and experimenting with more democratic and humane alternatives. Therefore, these findings suggest that artists are not merely responding to technological progress, they are actively shaping public discourse, aesthetic norms, and political imaginaries. As such, postdigital art should be recognized not just as cultural production, but as a meaningful intervention into the broader landscape of digital governance.

Future Research Directions

This study focused on artists' perspectives, practices, and design intentions. Further research could extend this work in several directions. First, by examining how audiences receive and interpret postdigital artworks. Do participants leave installations or workshops with a deeper understanding of surveillance and privacy? Do these experiences lead to behavioral or attitudinal shifts? Second, a more technical study could analyze how open tools, languages, and interfaces are developed and shared across artistic communities. This would offer insight into the material infrastructures of creative resistance. Third, expanding the geographic scope beyond Europe could reveal how different cultural, legal, and technological contexts shape artistic responses to privacy and surveillance issues. Comparative research might uncover common tactics, as well as unique local strategies for engaging with the "black box" of digital patterns.

In all cases, this research underscores the value of approaching privacy not just through law or policy, but through aesthetic, design, sensory, and participatory inquiry. Postdigital artists help make visible what is hidden, negotiable what seems fixed, and creative what often feels predetermined. Their practices remind us that resistance to technological dominance is not only possible, it can be imaginative, embodied, and shared.

References

Berry, D. M., & Dieter, M. (2015). Thinking postdigital aesthetics: Art, computation, and design. In D. M. Berry & M. Dieter (Eds.), *Postdigital aesthetics: Art, computation and design* (pp. 1–11). Palgrave Macmillan.

Birchall, C. (2014). Aesthetics of the secret. *New Formations, 83*, 25–46.

Bishop, C. (2012). *Artificial hells: Participatory art and the politics of spectatorship*. Verso.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Choi, H., Park, J., & Jung, Y. (2018). The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior, 81*, 42–51.

Cohen, J. E. (2012). *Configuring the networked self: Law, code, and the play of everyday practice*. Yale University Press.

Edmunson, A. (2015). Curating in the postdigital age. *M/C Journal, 18*(4).

Giannini, T., & Bowen, J. (2019). Art and activism at museums in a post-digital world. In *Electronic Visualisation and the Arts (EVA 2019)*.

Hartzog, W. (2018). *Privacy's blueprint: The battle to control the design of new technologies*. Harvard University Press.

Johnson, S., & Acemoglu, D. (2023). *Power and progress: Our thousand-year struggle over technology and prosperity*. Hachette.

Keulen, S., & Kroeze, R. (2018). Privacy from a historical perspective. In *The handbook of privacy studies: An interdisciplinary introduction* (pp. 21–56).

Langdon, M. (2014). *The work of art in a digital age: Art, technology and globalisation*. Springer.

Leorke, D. (2018). *Location-based gaming: Play in public space*. Palgrave Macmillan.

Lyon, D. (2001). *Surveillance society: Monitoring everyday life*. Open University Press.

Lyon, D. (2018). *The culture of surveillance: Watching as a way of life*. Polity.

Monahan, T. (2018). Ways of being seen: Surveillance art and the interpellation of viewing subjects. *Cultural Studies, 32*(4), 560–581.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*, 119–157.

Nyst, C., & Falchetta, T. (2017). The right to privacy in the digital age. *Journal of Human Rights Practice, 9*(1), 104–118.

Paul, C. (2002). Renderings of digital art. *Leonardo, 35*(5), 471–484.

Paul, C. (2020). Digital art now: Histories of (im)materialities. *International Journal for Digital Art History, 5*, Article 2.

Quan-Haase, A., & Wellman, B. (2005). Local virtuality in an organization: Implications for community of practice. In *Communities and Technologies 2005: Proceedings of the Second Communities and Technologies Conference, Milano 2005* (pp. 215–238). Springer.

Richards, N. (2021). *Why privacy matters*. Oxford University Press.

Sicart, M. (2014). *Play matters*. MIT Press.

Solove, D. J. (2010). *Understanding privacy*. Harvard University Press.

Strahilevitz, L. J. (2005). A social networks theory of privacy. *University of Chicago Law Review, 72*, 919–988.

Vervoort, J., Smeenk, T., Zamuruieva, I., Reichelt, L. L., van Veldhoven, M., Rutting, L., ... & Mangnus, A. C. (2024). 9 dimensions for evaluating how art and creative practice stimulate societal transformations. *Ecology and Society, 29*(1), 29.

Morrison, E. (2016). *Discipline and desire: Surveillance technologies in performance*. University of Michigan Press.

Vlavo, F. A. (2017). *Performing digital activism: New aesthetics and discourses of resistance*. Routledge.

Zuanni, C. (2021). Theorizing born digital objects: Museums and contemporary materialities. *Museum & Society, 19*(2), 184–198.

Zuboff, S. (2019). Surveillance capitalism and the challenge of collective action. *New Labor Forum, 28*(1), 10–29.