Citizens Infrastructures as a Way to Govern AI's Power to Shape our Shared Representations

How to Make Sure Auditing Institutions Are Aligned with Values and Societal Expectations?

Chiara Marcoccia

Abstract: *The example of generative AI models (genAI) for the production of images points to a serious gap in AI auditing. While genAI models deployed on digital platforms for information and communication (ICTs) have an increasing and particularly persuasive influence on users' representations of social events, groups and dynamics; users themselves have no effective way of reclaiming power over their shared representations, governing when and how AI can shape their vision of the world. I argue that auditing institutions, which assess the alignment of AI systems with a set of previously defined societal expectations, are not able to keep up with the challenge of aligning with societal values and expectations genAI models and recommender systems that are evolving at an incredible speed and through a non-disclosed frequency of retrainings.*

In this paper the author argues that, to fill this gap, the auditing pipeline needs to leverage users themselves, in the form of an intermediary structure that will enable a dynamic feedback loop to keep up with the evolving impact of recommender systems and genAI models on ICTs, across the individual, the community and the societal level. I propose the form of this structure by drawing from the social sciences and the concept of citizens infrastructures. These are an enduring structure to leverage existing citizen engagement into the socio-technical networks that interact with and are impacted by the deployment of a given AI system. In this paper two examples of such infrastructures are being discussed and an argument is made for their potential to govern the power of AI over our shared representations, before drawing the implications for how the very concept of AI auditing needs to evolve.

Keywords: *AI governance; Algorithmic auditing; Digital ecosystems; Artificial Intelligence; Generative AI; Recommender systems*

Introduction

AI-generated images are one of the most glaring failures of auditing institutions in AI governance. Companies have been cultivating artificial image synthesis since the development of Generative Adversarial Networks (GAN) in 2014 (Elgammal et al. 2017), but since OpenAI's image generator DALL-E, that generates images from text, was opened to the public in 2022 (OpenAI 2022), AI-generated images have flooded all public information and communication platforms (ICP) (Bond 2024). Presently, any internet image search displays some kind of AI-generated images, not all of which are immediately recognizable as such (Gangadharbatla 2022).

Yet, the public response to the large diffusion of AI-generated images is not a positive one[1]. Ratings of comfortableness towards and estimated capability of AI producing artistic images is very low compared to other areas of application of AI technology (Schepman et al. 2020), and the appreciation of AI-generated images is much lower than that of human crafted ones (Bellaiche et al. 2023; Hong et al. 2019; Chamberlain et al. 2017). This points out a misalignment between the industry and the public, but one that doesn't currently have a suitable reflection upon effective governance, as there is no adapted institution to assess, support and address this misalignment.

However, what is at stake behind this representation, are our very own shared representations. Indeed, ICPs where AI-generated images are taking up more and more space actually convey social and cultural cues that influence the construction of our shared beliefs and behaviours (Glickman et al. 2025; Ashkinaze et al. 2024; Pappalardo et al. 2024; Hoang et al. 2023; Sîrbu et al. 2019; Bail et al. 2018). Consequently, AI-generated contents themselves are already becoming a non-neglectable factor in our perception and appreciation

1 A *caveat* to be introduced here, as most data available on the public's reaction to AI-generated images is culturally situated, and, while significant in its proportion for the purpose of calling out on current auditing processes in AI governance, doesn't necessarily represent all attitudes around the globe equally. See for example Wu et al. (2019) about the different perception of AI-generated art among American test subjects as compared to Chinese.

of reality, influencing our opinions (Floridi 2024), our tastes (Wu et al. 2019) and our values (Brinkmann 2023), on a personal, interpersonal and societal level, with the potential to shape our perception of the self, our attitude towards others and our shared representations of the future, of democracy and of social identities (Floridi et al. 2024; Burrell et al. 2023; Ovadya et al. 2023). For such high stakes and wide-ranging consequences, we have no matching institutional levers: we lack infrastructures to connect between the big-tech industry, national and international regulatory instances, and the public. An issue for which I try to suggest a workable solution in this paper.

In fact, in the first part of my paper I show that the auditing institutions that have failed to govern the development of AI-generated images in alignment with societal values and expectations, have done so because of the lack of citizen infrastructures. These can take the shape of digital platforms, community forums, shared skills, information networks, or lived environments (AbdouMaliq 2004) and aim at supporting and enabling citizen-led and enduring public engagement (Gabrys 2021). Their specificity is that they often originate from everyday practices and in the very spaces where citizens interact (Livingstone et al. 2010).

I make the claim that such infrastructures can be particularly well-suited and effective for the purpose of enabling and supporting citizen engagement in auditing institutions involved in AI governance, effectively and durably aligning those institutions with social values and expectations. Indeed, a successful infrastructure connecting citizens and auditing institutions must allow governance issues to be formulated and discussed by citizens *contextually*, in a way that is coherent and organic with the live perception of these issues and with the experience-led evaluation of their consequences.

For this reason, citizens infrastructures appear as an interesting solution: they emerge from and in turn provide structure to the living environment or the everyday activities that are the very reason for citizens' perceptions, valuing and expectations regarding the issue at hand. For AI regulation, this embedding of the infrastructure into the living experience would mean that the infrastructure is part of the citizens' consistent interaction with AI-generated content, which mostly happens through ICPs. This naturally points toward a kind of digital infrastructure, whose characteristics I outline in the last part of my paper.

In fact, digital infrastructures have become more and more promising as technology advances (Barns 2016), but they have an especially interesting potential today to address the issue of aligning social values and expectations

towards AI with auditing institutions. Indeed, the live context in which citizens encounter AI-generated content are ICPs, and these already host community forums, shared skills and information networks, which are the channels through which citizens infrastructures are commonly organized and maintained.

Hence, a digital citizens infrastructure appears to be an interesting option to govern AI's power to shape shared representations in alignment with social values and societal expectations. In fact, such a digital infrastructure would close the gap between consuming and governing AI-generated content, insofar as the space where AI-generated content has an impact on citizens' life and perception — ICPs — would also be the space where citizens can have an impact on the generation and diffusion of artificial content. This brings me to conclude my paper on the idea that it is possible to counter the trend that makes out people to be more passive as AI becomes more pervasive, and stress on the contrary how human-AI interaction can be a chance to enhance governance, interaction and decision-making.

The Misalignment Between AI Deployment, Societal Values and Expectations

Over-Promising and Under-Delivering

When we look at the expectations surrounding AI, we are faced with a deep contrast between high hopes and actual user experiences (Kinney et al. 2024). Indeed, the computing power unlocked by AI technology opens up the potential for many promising applications in the fields of healthcare (Lamberti et al. 2019; Aung et al. 2021; Nguyen et al. 2023), business (Pathak et al. 2010; Chen et al. 2004), government (Pi 2021), education (Tlili et al. 2023; Zawacki-Richter 2019), and justice (Westermann et al. 2024).

In fact, the machine learning technology that powers AI systems enables them to extract patterns from huge amount of data (the so called big data) and use them to interpret new data (LeCun et al. 2015), potentially offering a tool to reduce human work-load (Zysman et al. 2024), support decision-making (Turki et al. 2024; Shin et al. 2024), democratize access to knowledge (Mehandry et al. 2025; Conryut et al. 2015) and enhance task efficiency (Autor et al. 2024; Crafts 2021; Ernst et al. 2019).

As a result, people expect AI technology to take up an important societal role, improving institutions and enhancing citizens' quality of life (Kinney et al. 2024: 1). Indeed, a global survey published in the 2025 United Nations' Human Development Report indicates that two thirds of people across countries expect to use AI in education, health and work within one year (UNDP 2025: 4). In fact, in the public sector, AI technology could be an opportunity to fuel innovative public services improving services to citizen queries and enhancing predictive capability for decision-making thanks to its capacity to analyze high-dimensional as well as unstructured data (Eggers et al. 2017).

However, when we look at current AI uses, not only do they not live up to expectations, but they point to a deeper problem in the overall direction of the development of this technology. On the one hand, current AI systems don't offer enough transparency in their reasoning (Burrell 2016) and trustworthiness in their sources (Castelvecchi 2016; Kaur et al. 2022) to support widespread applications in healthcare, business, government, education, and justice (Bender et al. 2021). On the other hand, current AI deployment actually causes at least as many (if not more) harms as benefits for citizens.

First of all, the capacity of AI to reduce human work-load is dependent on increasing hidden (and sometimes exploitative: see Capraro et al. 2024) human labor (Muldoon et al. 2024), such as data labelling and annotation (Acemoğlu et al. 2019). Secondly, evidence shows that current use of AI to democratize knowledge and support decision-making actually also has a negative social impact, especially on marginalized groups (Sartori et al. 2022), insofar as it reproduces harmful biases (Mehrabi et al. 2021), spreads misinformation (Allcott et al. 2019) and leads to unfair decision-taking (Crawford 2021; Angwin et al. 2016; Dastin 2018).

As a consequence, there is a mismatch between society's high expectations for AI's beneficial applications and the actual possibilities offered by current AI systems. Crucially, the problem seems to be more than just a question of technical advancement (UNDP 2025: 48): it seems to be part of a broader misalignment between societal expectations and the current direction of AI development.

In fact, while the techno-solutionist narrative would have biased AI outputs be just a sign of a still underdeveloped technology to be fixed by future refinement (Altman 2024), research shows that it's the very development strategy of AI systems that produces biased outcomes. Indeed, the datasets used to train AI systems such as Large Language Models (LLMs) are built from data coming with overwhelming majority from Western countries (Rahman et al.

2024), and consequently carry a certain culturally situated value system (Atari et al. 2025).

Moreover, training data are taken from the internet, and subsequently reflect the over-representation of English speakers (McIntosh et al. 2024), young people and people from developed countries among internet users (World Bank 2018), but also the over-representation of men among the writers of online forums used as training data, like for instance Reddit (Pew 2018) and Wikipedia (Barera 2020).

For this reason, if we want AI systems to reflect societal values such as inclusiveness, diversity and fairness, we can't rely on technical advancement (Winner 2017), but we have to purposefully include such values in the development and deployment of AI technology (UNDP 2025: 103–104; 119–121). This paper suggests a way how. First, I propose to identify the difficulties for the inclusion of societal values in AI development.

The Alignment Problem

The difficulty to design AI systems to reflect societal values has come to be known as the alignment problem (Russell 2020). This consists on the one hand in a normative and on the other hand in a technical difficulty: how to frame the set of values that AI should be aligned with, and how to successfully incorporate them in AI systems (Gabriel 2020).

The latter is only partially a task for AI engineers to solve, for it requires we identify what the goals of AI are, in order to identify how they can meet our own (Kinney et al. 2024). This is not an anthropomorphization of AI, which would attribute goals to the AI system and question its intrinsic moral values. On the contrary, it's about understanding the technology behind the AI system, what it was developed for and what it can do (McCoy et al. 2023).

Indeed, the two questions are interconnected, and to successfully incorporate societal values into AI systems, we have to identify the values that an AI system should be aligned with, *given the tasks that the system has been developed for* (McCoy et al. 2023). For instance, when addressing the alignment problem in the case of LLMs, if we take into account that the technology on which they are based is designed to predict the next word given a context (Bender et al. 2021), we can identify *what* values are concerned by their computing strategies and *how* they are mobilized in their behaviors (Millière 2023), allowing us to determine the relevant discussions and measures to be undertaken.

This requires an understanding of AI technology that is often confined to techno-economical *niches* that are not accessible to the public and sometimes not even to regulating institutions (Qi et al. 2024; Wachter et al. 2024), which concentrates the power over the social transformations associated with AI in a few hands and away from citizens, fueling the misalignment (UNDP 2025: 137–139).

But solving the alignment problem also requires an assessment of citizens' attitudes and positioning towards specific societal values and their concrete applications in AI development (UNDP 2025: 3–4). Indeed, the alignment problem is also a problem of operationalizing human agency within the development and deployment of AI technology: the values with which AI is aligned must emerge by democratic and deliberate expression of values and expectations in order for AI to reflect societal values (Innerarity 2024).

Hence, to address the misalignment between AI and societal values and expectations we need infrastructures that are apt to bring together societal auditing with technical knowledge about AI development. For this reason we will now focus on the infrastructures currently available for citizens to participate in AI governance, to identify the gap between citizens and auditing institutions and map the way for the development of new infrastructures.

The Problem with Current Auditing Institutions

Defining AI Auditing

Auditing AI means assessing the alignment of given AI systems with a set of previously defined expectations (Birhane et al. 2024). Who defines these expectations and whether they reflect societal values varies among different AI auditing institutions. These include the media, law firms, regulators, consulting agencies and academia (Bandy 2021). The goals for which each of these stakeholders assess AI systems differ, and, consequently, so do the standards they define as expectations for the evaluation of the AI system (Costanza-Chock et al. 2022).

Indeed, current auditing institutions either look to comply with mandatory requirements and minimize corporate liability (internal auditing), or to identify and minimize the harms impacting users (Birhane et al. 2024). Hence the auditing process, which could offer a means for actors developing and regulating AI to align expectations, actually offers no common ground for differ-

ent stakeholders to exchange assessments, only entertaining the distinction in their roles.

The implication is that the policy applications of these audits are far from systematic (Birhane et al. 2024): AI assessments often don't lead to proactive re-design, products recall, voluntary corporate action or broader governmental regulation; firstly, because of their heterogeneous nature (ALI 2021), and, secondly, because of their lack of a structured articulation into the social fabric around AI development (Vecchione et al. 2021).

Indeed, firstly, socio-political repercussions of AI audit are impeded by the lack of clearly defined auditing standards (Costanza-Chock et al. 2022), for it's those standards that drive not only the assessment of AI systems, but also their development. In fact, standards are 'consensus-based and agreed-upon ways of doing things' (Dignum 2022), that auditing helps to define and that in AI development translate into the minimum specification on how to carry out a process in alignment with socially-acknowledged expectations. This means that the standards that should be defined by auditing institutions are the same standards that are needed to guide AI development (Raji et al. 2020).

Secondly, transformative actions don't come out of AI audit because the auditing process lacks a direct involvement of the stakeholders that are most likely to be harmed by AI systems (Costanza-Chock et al. 2022). Indeed, institutions that carry out AI auditing lack a structured articulation into the social ramifications of AI development and deployment: the channel between tech-developing companies and users is severed (UNDP 2025: 137–139), preventing auditing standards to be informed by practical expectations and societal values, and limiting the active and democratic determination of these standards, which is an essential requirement for meaningful and impactful auditing (Innerarity 2024).

One example of these stalling effects of auditing institutions in their current form is the massive deployment of image-generating AI systems (Gangadharbatla 2022). Because impacted parties are not directly present in the elaboration of the auditing process, auditing institutions lack the focus and the pervasiveness to identify and address necessary changes in the regulation and the development of image-generating technology and translate them into auditing standards (Luccioni et al. 2023). Indeed, issues regarding privacy, copyrights and discrimination are central to a fair and trustworthy development of image-generating AI technology (Kaur et al. 2022), but auditing institutions are not sufficiently integrated into the societal adoption of and response to this

technology to meaningfully and effectively mediate the dialogue between the different stakeholders (Costanza-Chock et al. 2024).

For this reason, governing AI in alignment with societal values and expectations requires a new kind of auditing institution, apt to build meaningful and impactful infrastructures between auditing institutions and the different stakeholders affected by AI development. To define such a structure, we will first identify the role that it needs to have in AI governance.

Why Add Infrastructures?

The need for AI audit is widely advocated for in the name of AI (or algorithmic) accountability (Raji et al. 2020; ALI 2021; Bandy 2021; Costanza-Chock et al. 2022; Raji et al. 2022; Birhane et al. 2024). Indeed, the goal of AI audit is to provide an account of the way that a given AI system exercises its power in society (Bandy 2021: 74:3), meaning to give an account of the AI system's performance as well as of its social impact.

For this reason Wieringa (2020) talks about a 'networked account' for what is essentially a 'socio-technical system' (Wieringa 2020): AI systems are deployed within social networks, as for instance social media (Knott et al. 2021) and online retail platforms (Lee et al. 2014), and have network effects, like the spreading of information and the success of products (Pappalardo et al. 2024); therefore, the identification of the network effects of AI deployment and the understanding of their relation to specific technical features of the given AI system (Fabbri et al. 2022; Donnelly et al. 2021) is essential for the design of action-informative AI audit (Birhane 2024).

This is an added requirement to those successfully covered by already developed benchmarking systems (Liu et al. 2024) because it takes into account the unprecedented characteristic of AI systems: namely the fact that they evolve through their deployment (Martínez et al. 2023). Indeed, AI systems update their behaviors in response to their interaction with users: thus recommenders (Isinkaye et al. 2015) update their recommendations to align with users' preferences (Piao et al. 2023; Mansoury et al. 2020; Jiang et al. 2019), and Large Language Models (LLMs) integrate the new data generated by the interaction with users into their following outputs (Hataya et al. 2023). It's the process commonly known as feedback loop (Jiang et al. 2019; Sun et al. 2019; Pedreschi et al 2024).

Furthermore, this process is augmented by the fact that AI systems are re-trained over time, to integrate new sets of data (Shumailov et al. 2024; Pe-

dreschi et al. 2024). As a consequence, benchmarks that are defined on the data fed into the AI system at time t_1 won't necessarily match the characteristics of the data added at time t_2.

For instance, let's look at the benchmarks defined to account for the trustworthiness of a LLM system (Liu et al. 2024; Huang et al. 2024), namely reliability (the information is accurate and the system does not hallucinate), safety (the outputs are not violent, guarantee privacy and do not cause danger for the user), fairness (the reasoning is not biased towards certain communities and the outputs are not discriminatory), resistance to misuse (the system cannot be used to cause harm and protects minors), explainability (the reasoning is transparent and the outputs interpretable), alignment with social norms (it does not go against universally acknowledged social values) and robustness (typos, grammatical errors and repetitions in the prompts do not cause inaccurate outputs).

The specific data that need to be labeled as problematic in accordance with these benchmarks are defined from the dataset on which the LLM was built, and they will change with the usage that users make of the LLM. Bender et al. (2021) already noted that the list of words related to sex that was chosen to filter out offensive, violent and pornographic data from the training data of ChatGPT-2 actually reduced the influence of many forums built by or for the LGBTQ+ community, because it contained slurs reclaimed by certain marginalized groups (Bender et al. 2021).

This is where infrastructures come into play to align auditing institutions with societal values and expectations. If benchmarks, and especially the concrete application of benchmarks, evolve with use, then users need to be integrated in the process of continually keeping those benchmarks up-to-date. To be more specific, it's the very structure of the auditing system that has to be updated to match its role in ensuring accountability, given the specificities of AI technology (Pasquale 2019). Indeed, AI technology itself makes traditional auditing difficult: technology advances with considerable speed (Wu et al. 2025), models are retrained on new data (Zhu et al. 2024) and each system evolves with usage, by learning from the feedback of its users (Glickman et al. 2025; Mansoury et al. 2020), but also by tracking and analysing their behaviors (Boeker et al. 2022). For these reasons, audits can no longer be confined to their role of evaluating the compliance with previously established standards and benchmarks, as they did with softwares that didn't learn from interaction (Vecchione et al. 2021; Birhane et al. 2024). On the contrary, they need to take on an active

role in informing standardization and benchmarking institutions as well as the mitigation practices from system developers (Birhane et al. 2024).

It's to meet this new requirement that infrastructures are needed to bridge the gap between auditing institutions and expectations arising through actual AI usage. In the next part of this paper, I will present the characteristics of these infrastructures, and argue in favor of developing citizen infrastructures to support AI auditing.

Proposing a Solution: Citizens Infrastructures to Support AI Audit

A Proposition for Governance

The aim of this paper is to propose a workable proposition to build the missing link between citizens and auditing infrastructures that is essential to leverage the potential of AI audits to fuel policy, research and industry initiatives. The idea is to leverage existing citizen engagement (Gabrys 2021) into the socio-technical networks (Wieringa 2020) that interact with and are impacted by the deployment of a given AI system (Pedreschi et al. 2024, Pappalardo et al. 2024).

It is not about distributing data-collection tasks (Gabrys 2021: 88) and exploiting consumers for technology development, but rather about leveraging what is currently the most viable and transformative way of incorporating, on the one hand, democratic engagement into the auditing processes, and, on the other, auditing institutions into the social networks impacted by AI (Gabrys 2021: 89).

In fact, citizens infrastructures are citizen-led network practices that take place in everyday spaces and activities (AbdouMaliq 2004), such as interaction with AI powered systems on Very Large Platforms (VLOPs), but that have a rare enduring quality among traditional participatory infrastructures (Gabrys 2021: 88). For this reason, they make a particularly good candidate for bridging the divide between auditing institutions and the socially impacting and time-transformative nature of the AI systems they aim at assessing.

The concept of citizens infrastructures comes from data collection practices in climate studies (Gabrys 2021). It describes when communities that are impacted by and have an impact on a climate-related phenomenon, rather than passively providing data to institutions in the form of surveys for instance, take an active and network-organised role in the data collection, organising feedback, participating in the measurements, monitoring the territory.

Examples of these sustainable practices and of their role in making AI audits actually impactful for the industry and informative for regulation practices can be found in the Open Data Science Initiative (Lawrence 2014), which consists in a users-organized space for open-access development of new analysis methodologies available 'as widely and rapidly as possible with as few conditions on their use as possible' (Lawrence 2014), addressed to the data science community, as well as to commercial, scientific and medical institutions, with the aim of achieving 'a balance between data sharing for societal benefit and the right of an individual to own their data' (Lawrence 2014).

This initiative leverages spaces that are already invested by its users: GitHub, Jupyter Notebook and Reddit; and directs initiatives towards defining AI accountability in the domain of data usage. If relayed by auditing institutions, such a space could become not only an informative but also a directive structure for the design of audits that are fit to support policy and industry action (Capraro et al. 2024). Auditing AI is about assessing the alignment of AI systems with a set of previously defined expectations, however, to serve their purpose, these need to be defined, firstly, relatively to what AI is used for (which means through consultation of the public), and, secondly, relatively to what AI has been developed for (which means with some technical knowledge of how AI is designed and how it works), which indicatives like Open Data Science allow.

Notably, this prevents the characteristic weaknesses of traditional attempts to involve citizens in AI governance, which often don't encounter sufficient participation or continuous engagement (Lahdili et al. 2024; Sieber et al. 2024). Indeed, citizen infrastructures emerge from already rooted practices (Livingstone et al. 2010), which gives them a unique strength within the landscape of available means of citizen engagement with AI development to support the alignment with societal values and expectations (Wilson 2022).

This rootedness in existing practices gives citizen infrastructures a continuity and legitimacy that top-down participatory initiatives often lack. Rather than requiring the creation of new frameworks for engagement, these infrastructures build on collective habits of knowledge production, discussion, and critique that are already active within technical communities and user networks. As such, they are not only more likely to sustain participation over time, but also better positioned to translate between technical discourse and social concerns. This embedded quality can enable them to support the development of auditing standards that are both technically informed and socially respon-

sive, reinforcing the role of auditing as a site of alignment between AI systems and evolving public expectations.

Moreover, a citizen infrastructure modelled after the Open Data Science Initiative could ensure continuity between mitigation actions and technology developments, effectively tackling the challenge of keeping up with the evolution of AI systems, not only following industry developments (Shin et al. 2023), but also as a consequence of the feedback loop between users and AI (Pedreschi et al. 2024), and following possible re-trainings of already available generative AI and recommender systems (Shumailov et al. 2024).

As such, structures of this kind could act as successful infrastructures bridging between citizens and auditing institutions, however, there is one more reason why citizens infrastructures are an attractive solution to the problem of aligning auditing institutions with societal values and expectations. By guaranteeing accountability, AI audits support our governance over AI's power in society (Bandy 2021: 74:3), and integrating citizens infrastructures to the process can elevate this governance.

A Way to Claim Agency over the Power of AI to Shape Our Shared Representations, Values and Behaviors

AI auditing assesses the power of AI technology and of its deployment over society (UNDP 2025: 136–147), which governance strives to align with values and expectations that citizens actively decide upon (UNDP 1997). This is what powers the idea of creating an infrastructure through which AI auditing can be leveraged for governance. On the one hand, citizens infrastructures can potentially support the integration of auditing institutions into the circular evolution of AI with users, but, on the other, they can also add to auditing processes the space for active deliberation that is necessary to make auditing a means of exercising governance (Kuziemski et al. 2020).

Indeed, interaction with AI systems has an impact on the evolution over time of human behaviors (Pappalardo et al. 2024), as can be seen in the way that recommender systems influence the tone of social interactions on digital platforms (Ovadya et al. 2023), and chatbots can improve cooperation within a human team (Shirado and Christakis 2020). But, in the longer term, interaction with AI also has an influence on the evolution of cultural products (Fogarty et al. 2023) and values (Brinkmann 2023) as well as on the evolution of social norms (UNDP 2025: 61), ultimately contributing to defining human values, behaviors and representations.

For these reasons, it is essential that we, as citizens, are able to exercise governance over the power of AI to shape our very own behaviors, representations and values. And, in fact, the integration of citizens infrastructures to institutions auditing AI accountability can successfully be a step in the direction of such a governance (Stilgoe 2024), by providing a space for citizens to contribute to the assessment of AI, the mitigation of its risks and the fostering of its potential, but also by bridging the gap between regulating institutions and AI users (Dave 2019; McQuillan 2018).

Crucially, citizens infrastructures could bridge the gap between usage and development of AI without going through the established lobbies to reach the AI industry (Livingstone et al. 2010). Indeed, if auditing institutions can incorporate citizen-led governance initiatives such as the Open Data Science Initiative, then the auditing process can become a channel through which citizens, by informing the standards set and assessed by auditing institutions, can have a real space in the AI industry, and develop an alternative kind of AI auditing, in contrast with profit-driven certifications (Kuziemski et al. 2020).

Indeed, the Open Data Science initiative is a kind of citizens infrastructure that emerges from the spaces and activities that are characteristic of digital platforms in general: however, one other option is to incorporate as citizens institutions, spaces and activities that emerge from the integration of AI itself into digital platforms, leveraging not only user networks but also user interaction with AI and the way it resonates within these networks. An example of this is something that has recently started to be explored in the literature (see Hayashi et al. 2024 for discussion of the potentials of decentralized platforms; but also Dotan et al. 2023 for discussion of their weaknesses) under the name of decentralized platforms. In this type of platform, each AI system evolves within the community of interest that it interacts with, and, if this structure had to be leveraged for auditing, the AI system could be audited within that same community, for the purpose that it was built for and for the usage that it evolves with. Because of this, auditing standards would be closely tailored and constantly evolving, matching the pace of user-system feedback loops. Moreover, on such a platform, AI systems could also be audited across different communities, insofar as, on decentralized platforms, whenever a user from one community wants to connect with something in another community (item, person, opinion, information) and it is up to the AI system to find a so-called bridge for that connection (Ovadya et al. 2023; Stray et al. 2023). This way, both the expectations of the users and the actual evolution of the network become variables that can be taken into account, enriching the auditing process with standards

for the network effects of user-AI interaction (Fabbri et al. 2022; Donnelly et al. 2021), rather than remaining limited to the individual effects of a given AI system on a given user.

This is particularly important because the power of AI over our shared representations comes mostly from network effects (Alcott et al. 2019). Indeed, although our individual interaction with AI-generated content or with a recommender system may transmit ideas, knowledge or bias to us (Ashery et al. 2025; Colther et al. 2024; Peralta et al. 2021; Mansoury et al. 2020; Sîrbu et al. 2019; Sun et al. 2019; Zhao et al. 2019), what makes AI so increasingly powerful is the fact that that interaction and its effects have so to say ripple effects onto all of the networks that we are a part of (Pansanella et al. 2022a; 2022b). First, because of the network and sharing based nature of the digital platforms where human-AI interactions take place (Wang et al. 2019). Second, because of what Coté et al. (2025) call the recursive nature of AI technology, in which every interaction is always compared to and reinvested in another one. Sometimes explicitly, like in the case of comparative filtering (Sun et al. 2019), but more largely because of the traceability of human-AI interaction (Pedreschi et al. 2024), which allows AI to keep track of all interactions and to turn them into data, but also to perform the next interaction as a kind of projection (LeCun 2019). Indeed, an AI assistant makes an estimation of what the reaction of the user will be to its suggestions or notifications, that it generates accordingly, and it registers the actual reaction of the user as a confirmation or a contradiction of its estimation. It then uses this information to generate the next estimation, making it act as a kind of prediction (Coté et al. 2025). As a consequence, what individual interaction with AI tends to do, is to isolate the user into an autoreferential bubble (*confer* the literature on filter bubbles, echo chambers, etc. See: Del Yang et al. 2023; Srba et al. 2023; Tomlein et al. 2021; Cinus et al. 2022; Aridor et al. 2020; Del Vicario et al. 2016; Pariser 2011), where every interaction appears self-referential and as closely matched as possible to the respective user. While, in fact, the effects of that bubble are always network effects, as shows a very interesting study on music streaming with AI-powered recommenders (Porcaro et al. 2024): each individual is listening to more diverse music thanks to AI recommendations, but the overall audience is all listening to the same tracks, which shows that even when it's all about an individual user, it is always about everyone else too (see also Doshi et al. 2024). For this reason, it is so important to operationalize networks as an intentional infrastructure to govern the power of AI, countering this isolating trend.

In fact, what I propose is to update the auditing system to keep up with the originality of AI technology. To do so, I argue that we should leverage expertise from the social sciences, to identify and integrate into the auditing process networks that could work as citizens infrastructures, solving the problem of the speed and complexity of AI evolution, to make sure auditing institutions are aligned with societal expectations. Indeed, the development of citizens infrastructures to support auditing institutions would not only transform citizen engagement in AI governance, but could also transform the structure of the auditing systems around AI and reinforce auditing institutions themselves, redefining their role and curating the channel that connects them to other stakeholders in society, industry and in regulation.

Citizens infrastructures, by embedding deliberation into the very process of AI auditing, would make it possible for societal values to emerge through continual collective reflection. Rather than auditing AI against static standards, this model allows values to be negotiated and redefined over time, in response to technological developments and shifting social contexts. Auditing, then, becomes a recursive and situated process—one that not only aims to hold AI accountable, but also enables the articulation and iterative reconfiguration of normative frameworks in dialogue with the systems that affect them. We may look back here on the case of AI-generated images, and of the necessity to govern their influence over socially shared representations. An active and iterative confrontation between the stakeholders impacted by this issue is the key here to achieve governance over the power of AI to shape shared representations.

Indeed, by grounding oversight in use and experience, citizens infrastructures respond to the fragmentation between AI development, deployment, and evaluation. For this reason, it would be a strong framework to support further research about how such infrastructures could contribute to establishing a circular model of accountability in which the feedback of users informs the criteria that auditing institutions apply, and in turn shapes how systems are designed and governed. Achieving this continuous flow is of crucial importance, for it would reduce the dependency on intermediaries such as industry lobbies, and instead support the institutionalization of citizen input within audit bodies themselves. In doing so, citizens infrastructures would not merely complement existing mechanisms, but strengthen them, transforming auditing institutions into agents of governance embedded in the social environments they are meant to serve.

Conclusion

In this paper I relayed the criticism that current AI auditing institutions do not or do not often produce an impact on AI industry and regulators, and suggested a solution to leverage AI audit for algorithmic governance: the integration of citizens infrastructures.

I argued in favor of this solution firstly by describing the misalignment between the current AI deployment system and societal values and expectations. I then proceeded to point out how the current auditing system is unable to respond to this misalignment, and what the reasons are for this *impasse*. Finally, I argued in favor of a solution, advocating for citizen infrastructures as action-oriented, enduring and democratic structures, with the potential to transform the role of auditing institutions and their integration into the social ecosystems affected by AI development and deployment.

To conclude, the integration of citizens infrastructures into AI auditing not only offers a response to the limitations of current institutions, but also redefines the epistemic and political conditions under which algorithmic governance takes place. By anchoring auditing processes within the communities affected by algorithmic systems, this model grounds accountability in situated knowledge and collective deliberation. It enables a redistribution of authority, bringing into the governance process a broader range of actors—public interest organizations, local collectives, and individual users—whose perspectives are often marginalized in existing audit frameworks.

Further research is needed to operationalize this model into AI governance: among the foreseeable future steps, there is a need to define the scalability of citizens infrastructures, and to identify and address limitations and failure modes.

What is at stake here is not simply the improvement of technical procedures, but the reconfiguration of auditing as a democratic practice. The proposal to embed citizens infrastructures into audit institutions entails a shift away from compliance-driven, external assessments, towards forms of oversight that are participatory, reflexive and entangled with the lived realities of AI deployment. It opens the possibility for a mode of governance in which standards are co-produced by those whom they actually affect —one in which the evaluation of AI systems evolves with the values, concerns and expectations of the societies they shape and inhabit.

References

AbdouMaliq, S. (2004): "People as Infrastructure: Intersecting Fragments in Johannesburg", in: Public Culture 16 (3), 407–429.

Acemoğlu, D., Restrepo, P. (2019): "Automation and New Tasks: How Technology Displaces and Reinstates Labor", in: Journal of Economic Perspectives 33(2), 3–30.

Allcott, H., Gentzkow, M., Yu, C. (2019): "Trends in the diffusion of misinformation on social media", in: Res. Polit., 6 (2).

Altman, S. (2024): "The Intelligence Age". Available at: https://ia.samaltman.com/. Accessed 28 April 2025.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016): "Machine bias", in: ProPublica, May 23, 2016.

Aridor, G., Goncalves, D., Sikdar, S. (2020): "Deconstructing the filter bubble: user decision-making and recommender systems", in: Proceedings of the 14th ACM Conference on Recommender Systems, pp. 82–91.

Ashery, A.F., Aiello, L.M., Baronchelli, A. (2025): "Emergent social conventions and collective bias in LLM populations", in: Science Advances 11, eadu9368, doi: 10.1126/sciadv.adu9368.

Ashkinaze, J., Mendelsohn, J., Qiwei, L., Budak, C., and Gilbert, E. (2024): "How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence from a Large, Dynamic Experiment", in: arXiv preprint. arXiv:2401.13481.

Aung, Y.Y.M., Wong, D.C.S., Ting D.S.W. (2021): "The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare", in: Br. Med. Bull., 139 (2021), pp. 4–15.

Autor, D., Chin, C., Salomons, A., Seegmiller, B. (2024): "New Frontiers: The Origins and Content of New Work, 1940–2018", in: The quarterly journal of economics: qjae008.

Bail, C., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Fallin Hunzaker, M.B., Lee, J., Mann, M., Merhout, F., Volfovsky A. et al. (2018): "Exposure to opposing views on social media can increase political polarization", in: Proceedings of the National Academy of Sciences 115.37, pp. 9216–9221.

Bandy, J. (2021): "Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits", in: Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 74.

Barera, M. (2020): "Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia". Accessible at http://hdl.handle.net/10106/29572.

Barns, S., Cosgrave, E., Acuto, M., Mcneill, D. (2016): "Digital Infrastructures and Urban Governance", in: Urban Policy and Research, 35(1), 20–31.

Bellaiche, L., Shahi, R., Turpin, M.H. et al. (2023): "Humans versus AI: whether and why we prefer human-created compared to AI-created artwork", in: Cogn. Research 8, 42.

Bender, E.M., Friedman, B., (2018): "Data statements for natural language processing: Toward mitigating system bias and enabling better science", in: Transactions of the Association for Computational Linguistics 6, 587–604.

Birhane, A., Steer, R., Ojewale, V., Vecchione, B., Raji, I. D. (2024): "AI Auditing: The broken bus on the road to AI accountability", in: arXiv:2401.14462v1. Accessed 14/05/25. https://doi.org/10.48550/arXiv.2401.14462.

Boeker, M., Urman, A. (2022): "An empirical investigation of personalization factors on TikTok", in: Proceedings of the ACM Web Conference 2022, pp. 2298–2309.

Bond, S. (2004): "AI-generated spam is starting to fill social media. Here's why", in: *NPR Untangling disinformation*. Accessed 27/02/2025. Availavle at: https://www.npr.org/2024/05/14/1251072726/ai-spam-images-facebook-linkedin-threads-meta.

Brinkmann, L., Baumann, F., Bonnefon, JF. et al. (2023): "Machine culture" in: *Nat Hum Behav* 7, 1855–1868.

Burrell, J. (2016): "How the machine 'thinks': Understanding opacity in machine learning algorithms", in: *Big Data & Society*.

Burtell, M., Woodside, T. (2023): "Artificial influence: An analysis of AI-driven persuasion", in: arXiv preprint. arXiv:2303.08721.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F. et al. (2024): "The Impact of Generative Artificial Intelligence on Socioeconomic Inequalities and Policy Making", in: PNAS Nexus 3(6).

Castelvecchi, D. (2016): "Can we open the black box of AI?", in: *Nature* Oct 6 538 (7623): 20–23.

Chamberlain, R., Mullin, C., Scheerlinck, B., Wagemans, J. (2017): "Putting the art in artificial: Aesthetic responses to computer-generated art", in: *Psychol. Aesthet. Creat. Arts.*

Chen, P.-Y., Wu, S.-Y., Yoon, J. (2004): "The impact of online recommendations and consumer feedback on sales", in: International Conference on Information Systems (ICIS).

Cinus, F., Minici, M., Monti, C., Bonchi, F. (2022): "The effect of people recommenders on echo chambers and polarization", in: Proceedings of the In-

ternational AAAI Conference on Web and Social Media, vol. 16, Association for the Advancement of Artificial Intelligence, Washington, DC USA, pp. 90–101.

Colther, C., Doussoulin, J.P. (2024): "Artificial intelligence: Driving force in the evolution of human knowledge", in: Journal of Innovation & Knowledge, 9 (4) 100625, ISSN 2444–569X, https://doi.org/10.1016/j.jik.2024.100625.

Crawford, K. (2021): *The Atlas of AI*. Yale University Press.

Dastin, J. (2018): "Amazon scraps secret AI recruiting tool that showed bias against women", in: *Reuters.* Available at https://www.reuters.com.

Dave, K. (2019): "Systemic algorithmic harms", in: Data & society, May 31st 2019.

Del Vicario, M. Vivaldo, G. Bessi, A. Zollo, F. Scala, A. Cardelli, G. Quattrocchi, W. (2016): "Echo Chambers: Emotional Contagion and Group Polarization on Facebook", in: Scientific Reports, Open Access, Volume 61, December 2016, Article number 37825, https://doi.org/10.1038/srep37825.

Donnelly, R. Kanodia, A. Morozov I. (2021): "The long tail effect of personalized rankings". Available at SSRN 3649342.

Doshi, A.R., Hauser, O.P. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances 10, eadn5290. https://doi.org/10.1126/sciadv.adn5290.

Dotan, M., Yaish, A., Yin, H-C., Tsytkin, E., Zohar, A. (2023): "The Vulnerable Nature of Decentralized Governance", in: DeFi 25–31, https://doi.org/10.1145/3605768.3623539.

Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M. (2017): "CAN: Creative adversarial networks, generating 'Art' by learning about styles and deviating from stylenorms", in: arXiv preprint. arXiv:1706.07068.

Ernst, E., Merola, R., Samaan, D. (2019): "Economics of Artificial Intelligence: Implications for the Future of Work", in: IZA Journal of Labor Policy 9(1).

Fabbri, F., Croci, M.L., Bonchi, F., Castillo, C. (2022): "Exposure inequality in people recommender systems: the long-term effects", in: Proceedings of the International AAAI Conferenceon Web and Social Media, vol. 16, Association for the Advancement of Artificial Intelligence, Washington, DC USA, pp. 194–204.

Floridi, L. (2024): "Hypersuasion – On AI's Persuasive Power and How to Deal with It", in: Philos. Technol. 37, 64.

Gabriel, I. (2020): "Artificial Intelligence, Values, and Alignment", in: Minds and Machines 30(3), 411–437.

Gabrys, J. (2021): "Citizen Infrastructures and Public Policy: Activating the Democratic Potential of Infrastructures", in: Cohen, K. and Doubleday, R. (eds): Future Directions for Citizen Science and Public Policy, Cambridge: Centre for Science and Policy.

Gangadharbatla, H. (2022): "The Role of AI Attribution Knowledge in the Evaluation of Artwork", in: Empirical Studies of the Arts, 40(2), 125–142.

Glickman, M., Sharot, T. (2025): "How human–AI feedback loops alter human perceptual, emotional and social judgements", in: Nature Human Behavior 9, 345–359. https://doi.org/10.1038/s41562-024-02077-2.

Hagerty, A. Rubinov, I. (2019) Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv preprint. arXiv:1907.07892.

Hataya, R., Bao, H., Arai, H. (2023): "Will large-scale generative models corrupt future datasets?", in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20555–20565.

Hayashi, S., Caron, B. A., Heinsfeld, A. S. et al. (2024): "brainlife.io: a decentralized and open-source cloud platform to support neuroscience research", in: Nature Methods 21, 809–813, doi: https://doi.org/10.1038/s41592-024-02237-2.

Hong, J.-W., Curran M. N. (2019): "Artificial Intelligence, Artists, and Art: Attitudes Toward Artwork Produced by Humans vs. Artificial Intelligence", in: ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15(2):58, 1 – 16.

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q. et al. (2024): "TRUSTLLM: trustworthiness in large language models", in: Proceedings of the 41st International Conference on Machine Learning (ICML'24), Vol. 235. JMLR.org, Article 813, 20166–20270.

Innerarity, D. (2024): "Defensa y crítica de la gobernanza algorítmica", in: Revista CIDOB d'Afers Internacionals, 138, pp. 11–25.

Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A. (2023): "Recommendation systems: principles, methods and evaluation", in: Egypt. Inform. J., 16 (3), pp. 261–273.

Jiang, R., Chiappa, S., Lattimore, T., György, A., Kohli, P. (2019): "Degenerate feedback loops in recommender systems", in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19, Association for Computing Machinery, New York, NY, USA, pp. 383–390.

Kimutai, L. (2024): "Is AI killing Pinterest's magic?", in: *Medium*. Accessed 27/02/2025. Available at: https://medium.com/@kimutail/is-ai-killing- pinterests-magic-2f84c395fd98

Knott, A., Hannah, K., Pedreschi, D., Chakraborti, T., Hattotuwa, S., Trotman, A., Baeza-Yates, R., Roy, R., Eyers, D., Morini, V., Pansanella, V. (2021): "Responsible AI for social media guidance: a proposed collaborative method for studying the effects of social media recommender systems on users", in: Technical report. Global Partnership on AI (GPAI).

Kuziemski, M., Misuraca, G. (2020): "AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings", in: Telecommunications Policy, 44(6).

Lahdili, N., Onder, M., Nyadera, I. (2024): "Artificial Intelligence and Citizen Participation in Governance: Opportunities and Threats", in: Amme Idaresi Dergisi. 57. 202–229.

Lamberti, M.J., Wilkinson, M., Donzanti, B.A., Wohlhieter, G.E., Parikh, S., Wilkins, R.G., Getz, K. (2019): "A study on the application and use of artificial intelligence to support drug development", in: Clin. Therapeut., 41 (2019), pp. 1414–1426.

Lawrence, N. (2014): "Open Data Science", in: inverseprobability.com: Neil Lawrence's Homepage, accessed on: 15/05/25, available at: https://inverseprobability.com/2014/07/01/open-data-science.

LeCun, Y., Bengio, Y., Hinton, G. (2015): "Deep learning", in: Nature, 521 (7553), pp. 436–444.

Lee, D., Hosanagar, K. (2014): "Impact of recommender systems on sales volume and diversity", in: International Conference on Interaction Sciences.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M.F., Li, H. (2024): "Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment", in: arxiv.org/abs/2308.05374.

Livingstone, S., Lievrouw, L. A. (2010): "How to infrastructure", in: Star, S., Bowker, G. (Eds.): *Handbook of New Media: Social Shaping and Social Consequences of ICT*, pp. 230–245. SAGE Publications Ltd.

Martínez, G. Watson, L. Reviriego, P. Hernández, J. A., Juarez, M. Sarkar, R. (2023): "Towards understanding the interplay of generative artificial intelligence and the Internet", in: arXiv:2306.06130.

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., Burke, R. (2020): "Feedback loop and bias amplification in recommender systems", in: Proceedings of the 29th ACM International Conference on Information

& Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, pp. 2145–2148.

Mazzone M., Elgammal, A. (2019): "Art, Creativity, and the Potential of Artificial Intelligence", in: *Arts* 8(1):26.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., Griffiths, T. L. (2023): "Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve", in: arxiv.org/abs/2309.13638v1.

McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., and Halgamuge, M. N. (2024): "A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination", in: IEEE Transactions on Artificial Intelligence 5(6): 2739–2751.

McQuillan D. (2018): "Rethinking AI through the politics of 1968", in: Open Knowledge, October 13th 2018.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021): "A survey on bias and fairness in machine learning", in: ACM Comput. Surv., 54 (6) (2021), pp. 1–35.

Muldoon, J., Graham, M., Cant, C. (2024): Feeding the Machine: The Hidden Human Labour Powering AI. Canongate Books.

Nguyen, M.H., Tran, N.D., Le, N.Q.K. (2023): "Big data and artificial intelligence in drug discovery for gastric cancer: current applications and future perspectives", in: Curr. Med. Chem., 31.

OpenAI (2022): DALL·E 2. Accessed 27/02/2025. https://openai.com/index/dall-e-2/.

Ovadya, A., Thorburn, L. (2023): "Bridging Systems. Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance", in: Knight First Amend. Inst., 23–11.

Pansanella, V., Rossetti, G., Milli, L. (2022a): "Modeling algorithmic bias: simplicial complexes and evolving network topologies", in: Appl. Netw. Sci., 7 (1), p. 57.

Pansanella, V., Rossetti, G., Milli, L. (2022b): "From mean-field to complex topologies: network effects on the algorithmic bias model", in: R.M. Benito, C. Cherifi, H. Cherifi, E. Moro, L.M. Rocha, M. Sales-Pardo (Eds.), Complex Networks & Their Applications X, Springer, Cham, pp. 329–340.

Pappalardo, L., Ferragina, E., Citraro, S., Cornacchia, G., Nanni, M., Rossetti, G., Gezici, G., Giannotti, F., Lalli, M., Gambetta, D., Mauro, G., Morini, V., Pansanella, V., Pedreschi, D. (2024): "A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions", in: Arxiv 2407.01630. https://arxiv.org/abs/2407.01630.

Pariser, E. (2011): The Filter Bubble: What the Internet is Hiding from You, New York: Penguin Press. 294 pp.

Pathak, B., Garfinkel, R., Gopal, R.D., Venkatesan, R. Yin, F. (2010): "Empirical analysis of the impact of recommender systems on sales", in: J. Manag. Inf. Syst., 27 (2) (2010), pp. 159–188.

Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A-L., Dignum, F., Dignum, V ., Eliassi-Rad, T., Giannotti, F., Kertész, J., Knott, A., Ioannidis, Y ., Lukowicz, P., Passarella, A., Pentland, A.S., Shawe-Taylor, J., Vespignani, A. (2025): "Human-AI coevolution", in: Artificial Intelligence, V olume 339, 104244, ISSN 0004–3702, https://doi.org/10.1016/j.artint.2024.104244.

Peralta, A.F., Kertész, J., Iñiguez, G. (2021): "Opinion formation on social networks with algorithmic bias: dynamics and bias imbalance", in: J. Phys. Complex., 2 (4).

Piao, J., Liu, J., Zhang, F. et al. (2023): "Human–AI adaptive dynamics drives the emergence of information cocoon", in: Nature Machine Intelligence 5, 1214–1224. https://doi.org/10.1038/s42256-023-00731-4.

Pew (2018): Internet/Broadband Fact Sheet. https://www.pewinternet.org/fact-sheet/internet-broadband/.

Pi, Y. (2021): "Machine learning in Governments: Benefits, Challenges and Future Directions", in: JeDEM – EJournal of EDemocracy and Open Government, 13(1), 203–219.

Porcaro, L., Gómez, E., and Castillo, C. (2024): "Assessing the Impact of Music Recommendation Diversity on Listeners: A Longitudinal Study", in: ACM Trans. Recomm. Syst. 2, 1, Article 3 (March 2024), 47 pages. https://doi.org/10.1145/3608487.

Rahman, R., Owen, D., You, J. (2024): "Tracking Large-Scale AI Models", in: EpochAI. Availavle at: https://epoch.ai/blog/tracking-large-scale-ai-models.

Russell, S. (2020): Human Compatible: Artificial Intelligence and the Problem of Control, Penguin Publishing Group.

Sartori, L., Theodorou, A. (2022): "A sociotechnical perspective for the future of AI: narratives, inequalities, and human control", in: Ethics Inf. Technol., 24 (1), p. 4.

Schepman, A., Rodway, P. (2020): "Initial validation of the general attitudes towards Artificial Intelligence Scale", in: Computers in Human Behavior Reports 2020(1).

Shin, M., Kim, J., van Opheusden, B., Griffiths, T. L. (2023): "Superhuman Artificial Intelligence Can Improve Human Decision-Making by Increasing Novelty", in: Proceedings of the National Academy of Sciences 120(12): e2214840120.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y. (2024): "AI models collapse when trained on recursively generated data", in: Nature, 631 (8022), pp. 755–759.

Sieber, R., Brandusescu, A., Sangiambut, S., Adu-Daako, A. (2024): "What is civic participation in artificial intelligence?", in: Environment and Planning B, 0(0).

Sîrbu, A., Pedreschi, D., Giannotti, F. and Kertész, J. (2019): "Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model", in: PloS one, 14(3), p.e0213246.

Srba, I. Moro, R. Tomlein, M. Pecher, B. Simko, J. Stefancova, E. Kompan, M. Hrckova, A. Podrouzek, J. Gavornik, A. et al. (2023): "Auditing YouTube's recommendation algorithm for misinformation filter bubbles", in: ACM Trans. Recommend. Syst., 1 (1), pp. 1–33.

Stilgoe, J. (2024): "AI has a democracy problem. Citizens' assemblies can help", in: *Science* 385, eadr6713.

Stray, J., Iyer, R., Puig Larrauri, H. (2023): "The Algorithmic Management of Polarization and Violence on Social Media", in: Knight First Amendment Institute. KnightColumbia.Org, forthcoming, Available at: http://dx.doi.org/10.2139/ssrn.4429558.

Sun, W., Khenissi, S., Nasraoui, O., Shafto, P. (2019): "Debiasing the human-recommender system feedback loop in collaborative filtering", in: Companion Proceedings of the 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, pp. 645–651.

Tlili, A., Shehata, B., Adarkwah, M.A., Bozkurt, A., Hickey, D. T, Huang, R., Agyemang, B. (2023): "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education", in: Smart Learning Environments, 10.

Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., Bielikova, M. (2021): "An audit of misinformation filter bubbles on YouTube: bubble bursting and recent behaviours changes", in: Proceedings of the 15th ACM Conference on Recommender Systems, pp. 1–11.

Turki, A. T., Engelke, M., Sobas, M. (2024): "Advances in Decision Support for Diagnosis and Early Management of Acute Leukaemia." The Lancet Digital Health 6(5): e300–e301.

UNDP (United Nations Development Program) (1997): Governance for Sustainable Human Development, New York: UNDP Policy Document.

UNDP (United Nations Development Programme) (2025): Human Development Report 2025: A matter of choice: People and possibilities in the age of AI, New York: UNDP Report.

Vecchione, B., Levy, K., Barocas, S. (2021): "Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies", in: Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, NY, USA. ACM, New York, NY, USA.

Wagner, C., Strohmaier, M., Olteanu, A. et al. (2021): "Measuring algorithmically infused societies", in: Nature 595, 197–204. https://doi.org/10.1038/s41586 -021-03666-1.

Wang, W., Liu, Q.-H., Liang, J., Hu, Y., Zhou, T. (2019): "Coevolution spreading in complex networks", in: Physics Reports, Volume 820, Pages 1–51, ISSN 0370–1573, https://doi.org/10.1016/j.physrep.2019.07.001.

Westermann, H., Savelka, J. (2024): "Analyzing Images of Legal Documents: Toward Multi-Modal LLMs for Access to Justice", in: *Arxiv* 2412.15260. Accessed on 15/05/25. https://doi.org/10.48550/arXiv.2412.15260.

Wieringa, M. (2020): "What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability", in: FAT* 2020 – Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, Inc, pp. 1–18.

Wilson, C. (2022): "Public engagement and AI: A values analysis of national strategies", in: Government Information Quarterly, 39(1).

Winner, L. (2017): Do Artifacts Have Politics? Computer Ethics. London: Routledge.

World Bank (2018): "Indiviuals Using the Internet". https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2017amp;locations=USamp;start=2015.

Wu, Y., Yi Mou, Zhipeng Li, Kun Xu (2019): "Investigating American and Chinese Subjects' explicit and implicit perceptions of AI-Generated artistic work" in: Computers in Human Behavior 2019(104).

Wu, J., You, H., Du, J. (2025): "AI generations: from AI 1.0 to AI 4.0", in: arXiv preprint. arXiv:2502.11312.

Yang, C., Xu, X., Nunes, P., Siqueira, S.W.M. (2023): "Bubbles bursting: investigating and measuring the personalisation of social media searches", in: Telemat. Inform., 82, Article 101999.

Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F., (2019): "Systematic review of research on artificial intelligence applications in higher education – where are the educators?", in: International Journal of Educational Technology in Higher Education, 16.

Zhao, J., Wang, T., Yatskar, M., Cotterell, Ordonez, V., and Chang, K-W. (2019): "Gender Bias in Contextualized Word Embeddings", in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 629–634. https://doi.org/10.18653/v1/N19-1064.

Zhu L, Rong Y, McGee L. A., Rwigema J. M., Patel S. H. (2024): "Testing and Validation of a Custom Retrained Large Language Model for the Supportive Care of HN Patients with External Knowledge Base", in: Cancers (Basel). 16(13):2311. https://doi.org/10.3390/cancers16132311. PMID: 39001375; PMCID: PMC11240646.

Zysman, J., Nitzberg, M. (2024): "Generative AI and the Future of Work: Augmentation or Automation?". Available at SSRN 4811728.