



## City Research Online

### City, University of London Institutional Repository

---

**Citation:** Baronchelli, A. (2024). Shaping new norms for AI. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1897), 20230028. doi: 10.1098/rstb.2023.0028

This is the published version of the paper.

This version of the publication may differ from the final published version.

---

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/32432/>

**Link to published version:** <https://doi.org/10.1098/rstb.2023.0028>

**Copyright:** City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

**Reuse:** Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

---

---

---

City Research Online:

<http://openaccess.city.ac.uk/>

[publications@city.ac.uk](mailto:publications@city.ac.uk)

---



**Cite this article:** Baronchelli A. 2024

Shaping new norms for AI. *Phil. Trans. R. Soc. B* **379**: 20230028.

<https://doi.org/10.1098/rstb.2023.0028>

Received: 19 July 2023

Accepted: 18 September 2023

One contribution of 15 to a theme issue ‘Social norm change: drivers and consequences’.

**Subject Areas:**

behaviour, evolution

**Keywords:**

social norms, artificial intelligence, regulation, complex systems, ChatGPT

**Author for correspondence:**

Andrea Baronchelli

e-mail: [a.baronchelli.work@gmail.com](mailto:a.baronchelli.work@gmail.com)

# Shaping new norms for AI

Andrea Baronchelli<sup>1,2</sup>

<sup>1</sup>City, University of London, Northampton Square, London EC1V 0HB, UK

<sup>2</sup>The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK

AB, 0000-0002-0255-0829

As artificial intelligence (AI) becomes increasingly integrated into our lives, the need for new norms is urgent. However, AI evolves at a much faster pace than the characteristic time of norm formation, posing an unprecedented challenge to our societies. This paper examines possible criticalities of the processes of norm formation surrounding AI. It focuses on how new norms can be established, rather than on what these norms should be. It distinguishes different scenarios based on the centralization or decentralization of the norm formation process, analysing the cases where new norms are shaped by formal authorities or informal institutions, or emerge spontaneously in a bottom-up fashion. On the latter point, the paper reports a conversation with ChatGPT in which the LLM discusses some of the emerging norms it has observed. Far from seeking exhaustiveness, this article aims to offer readers interpretive tools to frame society’s response to the growing pervasiveness of AI. An outlook on how AI could influence the formation of future social norms emphasizes the importance for open societies to anchor their formal deliberation process in an open, inclusive and transparent public discourse.

This article is part of the theme issue ‘Social norm change: drivers and consequences’.

*What is the social norm for using it? What are the legal norms?*  
Jensen Huang, CEO of Nvidia, January 2023 [1].

*When it comes to human interaction with LLMs like myself, there are evolving social norms that have started to emerge.*  
ChatGPT. Conversation with the author, May 2023.

## 1. Introduction

It is likely that 2023 will be remembered as the year of artificial intelligence (AI). ChatGPT [2] was the fastest internet service to reach 100 million users until now (May 2023) [3] and the technology of large language models (LLMs) at its core is a fundamental element of sister apps for images such as Dall-e2, Midjourney and many others. One of the most fascinating aspects of LLMs is that they exhibit unpredicted emergent features. While the media excitedly reported how AI art generators have developed their own taste [4] or chatbots are able to pass school-level exams in a growing number of disciplines [5], only in 2023 was it released that, for the past two years, GPT models had consistently improved its performance in tests designed to measure theory of mind in children [6].

For anyone familiar with complexity science, observing emergent properties in a complex system made of billions of artificial neurons is perhaps not surprising, but the growth in human, or even superhuman, -like capabilities has attracted huge attention from the media and the public, sparking a hectic debate between the technology apocalyptic and integrated [7]. While it is clear that AI could bring us spectacular benefits, from better medical diagnosing to drug discovering, the risks have so far catalysed most of the public attention. Perils associated with narrow AI include increasing opportunities for manipulation of people, enhancing and dehumanizing weapons, and rendering human labour increasingly obsolescent [8]. On the other hand, self-improving ‘artificial general intelligence’ (AGI) could pose an existential threat to humanity itself.

Despite the prevailing uncertainty, there is consensus on at least two points. First, AI is going to have a significant impact on our life. Second, society is not prepared to deal with the technology [9]. New rules are needed to help the transition towards a world where humans and machine coexist to the benefit of the former, if not of both parties. The call for action comes from some of the leading figures of the AI revolution. In an official note, OpenAI, the company behind chatGPT, recognize an existential risk associated with AGI and advocate a strong public oversight over the governance of the ‘most powerful systems’ [10]. Similarly, Alphabet CEO Sundar Pichai wrote that AI is ‘too important not to regulate’ [11]. Earlier in the year, Nvidia CEO Jensen Huang said that AI will create tools ‘that require legal regulation and social norms that have yet to be worked out’ [1], highlighting how social norms can be a solution to solve large-scale problems [12,13].

How to shape these new norms? In the rest of the paper, I will explore some aspects of the onset of new norms for AI from a complex systems perspective. Following recent results from the complex systems approach to the emergence of linguistic norms [14], I will distinguish whether the new norm is imposed by a formal authority (e.g. in the form of a regulation or law), by an informal authority (e.g. through social influence) or rather emerges spontaneously in a bottom-up fashion. In all cases, I will not discuss what norms would be desirable or not, but rather focus on some potentially critical aspects of the norm formation process. Before proceeding, two caveats are in order. The first is that I will use the term ‘norm’ in a broad sense, encompassing both regulations and conventions, although of course this is valid only as a first-order approximation [15] and neglects the distinction between personal, descriptive, injunctive and other types of social norms. The second caveat is that, regarding AI’s novelty, the situation we are facing falls somewhere between norm change, when a new norm replaces an old one, and norm emergence, the case in which a new norm is established in a context where there was none. Although these cases share several similarities, both cases resulting in the establishment of a new norm, they also present theoretical differences involving the mechanisms of spontaneous emergence and tipping point, respectively, that I will not delve into [16].

## 2. Formal institutions

Formal institutions, such as governments, are created with the aim to govern human behaviour, and are endowed with the means to produce regulations—or laws—and enforce them.<sup>1</sup> Top-down formal regulation of AI is what the industry has been vocal about for the past months, amidst claims of existential threats to humanity and legal uncertainty on liabilities caused by AI [17,18]. There is agreement that regulation is needed for a safe adoption of AI in established sectors such as health and finance, to guarantee that the use of AI does not conflict with human rights, existing laws and other ethic considerations. Auditing, i.e. conducting independent and systematic evaluations of an entity’s actions or properties and communicating the results to stakeholders, is a promising pathway in this sense [19]. At the moment of writing, the European Union is at the forefront of the regulation race [20]. On the other hand, it is less clear how to eliminate the *existential threat* that AI poses to humanity, an

aspect that has gained most of the public attention. There are at least three main issues making the regulation of super-intelligent, or also current, AI hard to implement in an effective way.

### (i) Unknown unknowns

The first problem is that it is difficult to imagine what AI will be able to do, even just in the near future. Technology is often faster than regulation, as cryptocurrencies have shown us for over a decade now, and AI seems poised to be one of the fastest evolving technologies humanity has ever seen. History is full of examples of bad regulation stemming from hurry. For example, in the United Kingdom, the Locomotive Act 1865, popularly known as the Red Flag Act, mandated speed limits for self-propelled vehicles—i.e. cars—of 4 mph (6 km h<sup>-1</sup>) on country roads and 2 mph (3 km h<sup>-1</sup>) in cities [21]. Vehicles required a three-person crew consisting of a driver, a stoker, and a flagman who, carrying a red flag, would walk 60 yards (55 m) ahead of the vehicle. This ensured the vehicle’s speed aligned with the flagman’s pace, enabling them to warn horse carriages and facilitate the vehicle’s stoppage until they safely passed [21]. As much as this sounds funny, such an act remained in place for 31 years, until it was repealed in 1896. In 1884, the US state of Vermont proposed a similar act, which also lasted until 1896. In the case of AI, the risk that regulation grossly misses the target, being either too strict or too loose, is high.

### (ii) Limited regulatory control

Analogies are often drawn between the need to regulate AI and the existing approaches to limit the risk of nuclear war or aeroplane accidents [10]. There is, however, a substantial difference between the AI industry and many of the examples we are familiar with. AI is relatively easier to implement and deploy, leaving virtually no traces. It is mainly a private endeavour, in contrast to government-run sectors like nuclear power. Additionally, thanks to the internet, AI is not bound by geography. Finally, while large-scale LLMs may require substantial resources, more focused alternatives can be way less demanding. Hence, while formal institutions such as governments can perhaps impose regulations concerning the narrow, ‘official’, use of AI, for example on the workplace, schools and hospitals, it is hard to see how the rule of law might limit malicious actors willing to use AI to commit crimes. For example, while regulation can help force online platforms to meet good security standards, there is little doubt that there will be efforts to use AI to organize misinformation campaigns and spread deep fakes. One could object that this is always the case, and that today an aeroplane can be used to bring destruction. While valid, this objection violates the assumption that AI poses an existential threat to humanity and neglects that it is the first smartphone-compatible technology to do so. Thus, for example, invitations to halt the development of self-improving artificial general intelligence [8] are difficult to implement, even where there is a will to do so, owing to the opacity of the research taking place behind closed doors.

### (iii) Quis custodiet ipsos custodes?

Who will guard the guards themselves? The old dilemma of every governance system seems poised to be critical in the case of AI [19]. While the technology is fragmented and open source versions of AI exist, the bottlenecks of talent and

energy costs have concentrated cutting edge research in the hands of a few private companies, creating power imbalances. Thus, the statement in favour of public oversight of AI that I mentioned above, posted by OpenAI on 23 May 2023, was followed 2 days later by the threat that OpenAI might leave Europe if the European Union (EU) imposed too much regulation [22]. Similarly, the above-mentioned piece by Google's CEO Sundar Pichai does not just state that AI is 'too important not to regulate', but adds that it is also 'too important not to regulate *well*' (italics my own). Who should be the judge for this 'well' is not clarified, and the suspicion that Google itself may want to retain a voice in the legislative debate is strong. Such a position is far from absurd, given the complexity of the subject, but it highlights obvious conflicts of interest. In this perspective, a possibly illuminating example of the difficult dialogue between regulation and digital technology is the announcement, dated 27 May 2023, that Twitter—the social media platform—will leave the EU voluntary Code of Practice on fighting misinformation [23]: regulation has been maintained on a voluntary basis not to scare oligopolists, with the risk that the latter adhere to it just until when they do not.

### 3. Informal institutions

Another main driver of norm change is represented by informal institutions [14]. These organizations can not enforce the adoption of a new norm, or can do so only within their non-exclusive remit, yet are influential in proposing new codes of behaviour. For example, in most of today's Western countries, religious leaders can prescribe codes of conduct, but have no legal permission to police behaviour (we do not enter here on the psychological aspects that may of course play a major role in such cases). Similarly, a university may produce a code of conduct that includes sanctions for violators, but it cannot prevent students from joining another university with a different set of rules. With their local regulations, typically informal institutions either make up for the lack of formal regulation, as it is currently the case for AI, or detail it further in order to make it fitter for their own environment. Two interesting sectors for an academic audience, among many, where informal institutions have been faced with the urgent need of new norms for AI are scientific publishing and education.

Journal editors, researchers and publishers started debating about LLMs in the published literature early on: Should ChatGPT be listed as author? Should authors disclose its use? A number of influential publishers, such as Springer, *Science* and *JAMA*, concluded right away that ChatGPT can not be listed as an author [24–26]. Motivations range from the practical-level 'attribution of authorship carries with it accountability for the work, and AI tools cannot take such responsibility' (*Nature* journals, [27]) to more philosophical considerations that 'the [scientific] product must come from—and be expressed by—the wonderful computer in our heads' (*Science*, [26]). In a blatant act of anglophone-centric deliberation, which ignores the liberating potential of LLMs as text editors for non-native English speakers, proposed policies regarding the use of LLMs range from banning any text written by AI to asking authors to disclose its usage in the paper, when no existing rule mandates disclosure of human proof-editing of a published article [28]. Of course, the main limitation of such policies is that compliance is hard to verify.

Education is another heated battleground on how to regulate AI. Several institutions in countries ranging from France to the USA and, most recently, to India and Italy, have moved to outlaw ChatGPT completely, raising concerns that they may limit student access to a formidable learning tool [29,30]. But the cat is out of the box, and AI assistants are doomed to remain a huge temptation for students facing essay-based homework [31]. Countries where oral examinations are the norm, such as France or Italy, find themselves in a stronger position. Given that oral exams represent a natural antidote to the LLM doping for students, the question of whether they will face wider adoption worldwide is natural. The fact that different universities may in principle adopt different approaches could represent an ideal experimental set-up to evolutionarily determine what the best approach is, yet contagion effects are strong, and not all universities are equally influential. For now, it is hard to imagine that outright bans are a definitive solution. While conservative approaches have been preferred in the beginning of the revolution, more nuanced approaches are likely to emerge soon.

These examples reveal some critical aspects of regulation by informal institutions. Firstly, informal institutions need to act immediately. As soon as a tool such ChatGPT is released, students will use it and higher education institutions are forced to take a position. Secondly, and partly as a consequence of the previous point, informal institutions most often aim to produce regulation tailored on the *status quo* of the technology, with no intent to provide a framework able to accommodate future developments. While this may not be a problem for the single institution, which can update its regulation at any point, the consequences can be undesirable for the stakeholders. It is easy to imagine how a student will be told to follow very different, if not contradictory, rules concerning AI during their career. Thirdly, complex contagion effects between institutions operating in the same sector, such as publishing or education, may be strong. As institutions watch one another, standards risk originating from either the first mover or the most prestigious, by some conventional metric, institutions, rather than being the result of an evolutionary process selecting the most desirable norms. In other words, and sticking to the education example, only a tiny fraction of the huge potential universities have in terms of expertise and strategic thinking would be used, as most institutions might just copycat the deliberations made elsewhere.

### 4. Spontaneous norms

The third main process of norm creation—and change—involves spontaneous emergence. Universally accepted norms are the unintended consequence of individuals' efforts to coordinate locally with one another [32–36]. Similar to other emergent phenomena observed in complex systems, global coordination in this context results from self-organization within a network of locally interacting individuals. This spontaneous process interacts with the activity of formal and informal institutions in complex ways that range from complementing existing regulations to countering them in ways that can result in conflict and legislative change. More importantly, spontaneous norms tend to occupy a much wider space than regulations. From shaking hands to wearing ties on formal occasions, to language and notion of fairness,

most of our expectations about the behaviour of others are based on unwritten norms.

Given the anticipated ubiquity of AI, it is reasonable to expect that norms will emerge on how to relate to it in different contexts. In 2 years time, will the reader of this piece be annoyed by my non-perfect English given that I could have so easily resorted to ChatGPT to polish my text? Or conversely, would a machine editing of my words be perceived as suspicious? Similarly, will the use of LLMs for brainstorming be condemned or encouraged?

Modelling [34], laboratory experiments [36] and data [14] suggest that the spontaneous emergence of norms in small-world social networks such as the ones we populate online occurs in two phases. Initially, and possibly for a long time, several alternatives compete. The dominating norm has only limited advantage for the runners up and there is a fast turnover in the top ranks. Then, owing to the fluctuations intrinsic to the decentralized conversation, one convention breaks the symmetry and the system enters a ‘winner takes all’ phase, which very rapidly leads to the establishment of a single shared norm [34]. Importantly, spontaneous consensus does not necessarily select the ‘best’ norm, and differences in utility among candidate norms, unless very significant, may play only a marginal role in their fate. In the context of AI, this means that we may expect new norms to emerge suddenly, apparently without early signals. Furthermore, we can not rely on the fact that they will be optimal. For instance, the current informal norm of declaring the use of ChatGPT as a text editor for scientific publications could potentially be consolidated across journals, leading to a discriminatory impact on non-native English speakers.

The paradox of norm change is that, once a norm is in place, it may be hard to overturn since it is in everyone’s best interest to comply to it. Yet bottom-up processes can also lead to norm change. In particular, the so-called critical-mass theory states that when a minority of the population formed by individuals committed to overturn the existing norm reaches a critical group size—commonly referred to as a ‘critical mass’—the social system crosses a tipping point [37–39]. Once the tipping point is reached, the actions of a minority group trigger a cascade of behaviour change that rapidly increases the acceptance of a minority view. The range of cases in which critical-mass theory may work is somewhat fuzzy, and the framework has been used to account for changes concerning spontaneous conventions as well as formal regulations. Experimental results, obtained in the framework of the Naming Game model [33,34], suggested a critical mass of 25% of the population [40], while modelling indicated a lower threshold [39]. In the case of AI, if the current trend of (informal, so far) regulation continues, one may expect spontaneous attempts to promote new norms. The case of students demanding better regulation for the use of AI in higher education seems an easy first guess. Analogously, non-native English speakers could try and overturn regulations that prevent them from using AI as a tool to level the playing field.

### (a) Spontaneously emerging norms on artificial intelligence, as seen by ChatGPT

Gathering data on the spontaneous processes currently shaping the norms around AI is difficult. But there is a privileged point of view to observe how our behaviour has evolved in the first months of public LLM chats, namely the one of ChatGPT. Therefore, I report here two extracts of a

conversation I had with the freely accessible chatbot on 22 May 2023. This may appear as a lazy choice, but I think it has merit. If this article can only try to guess potential scenarios for the ongoing yet still-in-its-infancy process of norm formation, the text below may be of interest for the future reader, as a testimony of what the machine was ‘thinking’ in these early days. It is evident that some of the points exposed by ChatGPT correspond to the desiderata of the programmers. Yet it is also possible to accept that some of the answers were generated based on the learning process of the AI during the interactions with its users.

After the LLM clarified that it’s definition of norm corresponds to the one used throughout this piece, i.e. that a norm ‘is a widely accepted standard or expectation of behaviour in a society or group’, it replied to my question on what emerging norms it has noted during the first months of its activity as reported below.

While social norms may vary depending on cultural and individual factors, here are some general trends that can be observed:

1. **Politeness and Respect:** Many people tend to interact with LLMs in a polite and respectful manner, treating them as intelligent entities. This includes using appropriate language, expressing gratitude, and avoiding derogatory or offensive remarks.
2. **Acknowledging Limitations:** Social norms often involve recognizing that LLMs have limitations. Users understand that LLMs provide responses based on patterns and data but may not always possess contextual understanding or have access to real-time information. It is important for users to critically evaluate and verify information obtained from LLMs.
3. **Contextual Understanding:** Users have started to recognize that LLMs may not fully grasp the nuances and complexities of certain topics or cultural contexts. Consequently, users are encouraged to provide additional context or clarification to ensure accurate and relevant responses.
4. **Responsible Use:** As LLMs become more prevalent, there is a growing awareness of the importance of responsible use. This includes not relying solely on LLMs for critical decisions, being cautious about potential biases in the training data, and understanding the ethical implications of using LLMs.
5. **Human-Machine Collaboration:** Social norms are evolving to promote collaboration between humans and LLMs. Users understand that LLMs can be valuable tools for information retrieval, idea generation, and learning. The norm is to view LLMs as partners rather than replacements, leveraging their strengths while recognizing the irreplaceable role of human creativity, empathy, and critical thinking.
6. **Privacy and Data Protection:** Norms around privacy and data protection apply to interactions with LLMs as well. Users expect that their personal information and interactions with LLMs are handled securely and responsibly, adhering to established privacy norms and regulations.

These social norms continue to evolve as the technology and usage of LLMs progress. As society adapts to the presence of LLMs and similar AI systems, it is important to have ongoing discussions to shape these norms and ensure that they align with our values and aspirations for human-AI interaction.

I then asked about the norms it has learned, and that it sticks to, when interacting with humans. The answer, redacted for brevity, was:

As an AI language model, I adhere to certain norms when interacting with humans to ensure a positive and respectful experience.

1. **Respectful and Polite Language:** I aim to use respectful and polite language when engaging with users.
2. **Transparency about AI Nature:** I make an effort to clarify that I am an AI language model and not a human.

3. Encouraging Critical Thinking: I promote critical thinking and encourage users to independently verify information obtained from me.
4. Acknowledging Subjectivity and Bias: I acknowledge that, despite efforts to be unbiased, I may have inherent biases in my responses owing to the training data I was exposed to.
5. User Empowerment: I aim to empower users by providing them with information, insights, and suggestions. I encourage users to make their own decisions based on their unique circumstances and values.
6. Privacy and Data Protection: I prioritize user privacy and confidentiality. I do not store personal data unless explicitly provided by the user for the purpose of the conversation.

It is important to note that these norms are constantly evolving, and I strive to improve and adapt based on user feedback and societal expectations.

Put together, these two lists highlight how the process of norm evolution with such a new kind of objects as LLMs will in fact be a process of coevolution. As a linguistic entity, the coming AI will need to learn and adapt to our social norms, and it is likely that we will need to do the same—with newly emerged norms—when interacting with it. How a machine could nudge us into behaving in a certain way, by making it believe that the expected behaviour is just ‘normal’, is a point of concern.

## 5. Conclusion and outlook

This paper considered the interplay between AI technology and our societal response by focusing on the three layers of norm formation represented by regulation by formal institutions, informal institutions and spontaneous norms [16]. With no claim of exhaustiveness, it focused on identifying possible criticalities affecting the mechanisms of norm formation. For regulation promulgated by formal institutions, it considered the difficulties introduced by the speed of AI developments, the unavoidably limited effect of regulation on ill-intentioned actors using an easy to replicate technology, and the fact that official governance may be subject to conflict of interests given the high economic stakes at play. In the case of informal authorities, it discussed the possible risks stemming from the need to react immediately to any AI novelty, the consequent short-sightedness of regulations and the impact of the latter on stakeholders, as well as contagion effects across institutions that could lead to suboptimal sector standards. Finally, for the case of bottom-up processes it stressed how, on the one hand, spontaneous norms will likely evolve in populations whose agents are both humans and AI bots, and, on the other hand, classic mechanisms of norm change, such as critical-mass dynamics, could bring about abrupt changes in the normative landscape concerning the place of AI in our societies. Interestingly, according to ChatGPT, norms are emerging that make users treat LLMs as intelligent, yet not omniscient, beings that may not always possess contextual understanding and view them ‘as partners rather than replacements’.

Before concluding, it is worth considering a further implication of the fact that—unlike previous technologies—AI is proficient in human language. Hence, it can in theory, and will likely, not only take part in the coevolution of norms concerning direct AI–human interactions, but more broadly participate in other general processes of norm change. Consider for example the theory of critical mass.

A problematic aspect of the theory is how the minority of committed users can reach the size required for the behavioural cascade to start [41,42]. Soon, AI bots on the internet could contribute to the initial phase of a movement for social change, including but not restricted to change concerning norms broadly related to the role of AI in society. In this respect, it is worth mentioning that the periphery of a social network plays a crucial role in social movements [43], implying that AI bots would not need to be behind influential accounts, but rather contribute behind the scenes to the activity taking place at the outskirts of the conversation. Ubiquitous AI bots could also make social media users more diffident, i.e. less prone to social influence and less willing to adopt a social norm based on exposure to peers complying to it. Counterintuitively, models suggest that such a state of diminished social influence could drastically lower the size of the critical mass required to trigger norm change [44], effectively destabilizing the normative *status quo*.

Being able to persuasively use language, AI could also exasperate the polarization observed in online social media [45]. By individualized analyses of user feeds, AI bots could for example cater to each participant in the network with tailored messages that suit their existing biases. When considering the process of norm formation, this is especially relevant because different echo-chambers can rapidly develop different norms, or also arrive at the same norm via different narratives [16]. The combination of large-scale deployment and targeted messaging is surely one of the most relevant aspect to monitor for linguistically fluent LLMs, and it is hard to provide educated guesses at this stage.

Finally, a remark about the shaping of formal regulations in democratic societies is in order. The speed of AI development is a major issue, but rushing to issue AI regulations in the absence of a well-informed and broadly participated debate could drastically erode the trust of citizens towards institutions. Furthermore, since choices regarding technology are path-dependent and difficult to reverse, badly designed regulation could have long-lasting repercussions. Therefore, in order to preserve public trust, governments should strive to engage citizens and promote a healthy public debate, seeking to guarantee full transparency on the nature of the involvement of AI corporations into the regulatory process. In this respect, special attention should be paid to monitoring the media and public discourse on AI, especially on social media platforms. Proactive measures should be taken to combat misinformation and the subsequent polarization that may pose a threat to the democratic process and, in this context, even the preservation of our open societies.

**Data accessibility.** This article has no additional data.

**Declaration of AI use.** I have used AI-assisted technologies in creating this article. Technology was used as explicitly discussed in the paper, i.e. it was interviewed.

**Authors' contributions.** A.B.: conceptualization, writing—original draft.

**Conflict of interest declaration.** I declare I have no competing interests.

**Funding.** The author acknowledges support from the IRIS Infodemic Coalition (UK government, grant no. SCH-00001-3391).

## Endnote

<sup>1</sup>Different definitions are possible for formal, as well as informal, institutions. Here, and in the remainder of the paper, I stick to a coarse distinction that has proven to be valid empirically [14].

## References

- Schmidt B. 2023 *Chatgpt 'arms race' adds \$4.6 billion to Nvidia founder's fortune*. Bloomberg. See <https://www.bloomberg.com/news/articles/2023-01-27/chatgpt-arms-race-adds-4-6-billion-to-nvidia-founder-s-fortune> (accessed 14 June 2023).
- OpenAI. 2022 *Introducing ChatGPT*. See <https://openai.com/blog/chatgpt>.
- Hu K. 2023 *Chatgpt sets record for fastest-growing user base—analyst note*. Reuters. See <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> (accessed 14 June 2023).
- Anonymous. Art made by artificial intelligence is developing a style of its own. *Economist*, 24 May 2023. See <https://www.economist.com/culture/2023/05/24/art-made-by-artificial-intelligence-is-developing-a-style-of-its-own>.
- Murphy Kelly S. 2023 *Chatgpt passes exams*. *CNN Business*, 26 January 2023. See <https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html>.
- Kosinski M. 2023 *Theory of mind may have spontaneously emerged in large language models*. *arXiv*, 2302.02083v3. (doi:10.48550/arXiv.2302.02083)
- Eco U. 2000 *Apocalypse postponed: essays*. Bloomington, IN: Indiana University Press.
- Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D. 2023 *Threats by artificial intelligence to human health and human existence*. *BMJ Glob. Health* **8**, e010435. (doi:10.1136/bmjgh-2022-010435)
- Hinton G *et al.* 2023 *Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war*. In *Statement on AI risk*. San Francisco, CA: Center for AI Safety. See <https://www.safe.ai/statement-on-ai-risk>.
- Altman S, Brockman G, Sutskever I. 2023 *Governance of superintelligence*. *OpenAI*. See <https://openai.com/blog/governance-of-superintelligence>.
- Pichai S. 2023 *Google CEO: building AI responsibly is the only race that really matters*. See <https://www.ft.com/content/8be1a975-e5e0-417d-af51-78af17ef4b79>.
- Nyborg K *et al.* 2016 *Social norms as solutions*. *Science* **354**, 42–43. (doi:10.1126/science.aaf8317)
- Andrighetto G, Vriens E. 2022 *A research agenda for the study of social norm change*. *Phil. Trans. R. Soc. A* **380**, 20200411. (doi:10.1098/rsta.2020.0411)
- Amato R, Lacasa L, Díaz-Guilera A, Baronchelli A. 2018 *The dynamics of norm change in the cultural evolution of language*. *Proc. Natl Acad. Sci. USA* **115**, 8260–8265. (doi:10.1073/pnas.1721059115)
- Przepiorka W, Szekely A, Andrighetto G, Diekmann A, Tummolini L. 2022 *How norms emerge from conventions (and change)*. *Socius* **8**, 237802312211245. (doi:10.1177/23780231221124556)
- Baronchelli A. 2018 *The emergence of consensus: a primer*. *R. Soc. Open Sci.* **5**, 172189. (doi:10.1098/rsos.172189)
- Čerka P, Grigienė J, Širbikytė G. 2015 *Liability for damages caused by artificial intelligence*. *Comput. Law Secur. Rev.* **31**, 376–389. (doi:10.1016/j.clsr.2015.03.008)
- Glaubitz A, Raymond N. 2021 *How should liability be attributed for harms caused by biases in artificial intelligence*. PhD thesis, BA thesis, Yale University, New Haven, CT. See <https://politicalscience.yale.edu>.
- Mökander J, Schuett J, Kirk HR, Floridi L. 2023 *AI Ethics* **2023**. (doi:10.1007/s43681-023-00289-2)
- Browne R. 2023 *EU lawmakers pass landmark artificial intelligence regulation*. CNBC. See <https://www.cnbc.com/2023/06/14/eu-lawmakers-pass-landmark-artificial-intelligence-regulation.html> (accessed 15 June 2023).
- Agnew J. 2020 *Steam engines on UK roads, 1862–1865: banning orders, agricultural locomotives and the 'red flag' act*. *Int. J. Hist. Eng. Tech.* **90**, 53–74. (doi:10.1080/17581206.2020.1797447)
- Perrigo B. 2023 *OpenAI could quit Europe over new AI rules, CEO Sam Altman warns*. *Time*, 25 May 2023. See <https://time.com/6282325/sam-altman-openai-eu/>.
- Lomas N. 2023 *Elon Musk takes Twitter out of the EU's disinformation code of practice*. See <https://tcrn.ch/43tQ8ml>.
- Vincent J. 2023 *Chatgpt can't be credited as an author, says world's largest academic publisher*. *Verge*, 26 January 2023. See <https://www.theverge.com/2023/1/26/23570967/chatgpt-author-scientific-papers-springer-nature-ban>.
- Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. 2023 *Nonhuman 'authors' and implications for the integrity of scientific publication and medical knowledge*. *JAMA* **329**, 637–639. (doi:10.1001/jama.2023.1344)
- Holden Thorp H. 2023 *Chatgpt is fun, but not an author*. *Science* **379**, 313–313. (doi:10.1126/science.adg7879)
- Stokel-Walker C. 2023 *Chatgpt listed as author on research papers: many scientists disapprove*. *Nature* **613**, 620–621. (doi:10.1038/d41586-023-00107-z)
- Brainard J. 2023 *As scientists explore AI-written text, journals hammer out policies*. *Science* **379**, 740–741. (doi:10.1126/science.adh2762)
- Viljoen J. 2023 *AI in higher education—a tool for better learning?* *University World News*, 24 February 2023. See <https://www.universityworldnews.com/post.php?story=20230222130606612>.
- Baidoo-Anu D, Owusu Ansah L. 2023 *Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning*. *J. AI* **7**, 52–62. (doi:10.2139/ssrn.4337484)
- Cano YM, Venuti F, Martinez RH. 2023 *ChatGPT and AI text generators: should academia adapt or resist?* Cambridge, MA: Harvard Business Publishing Education. See <https://hbsp.harvard.edu/inspiring-minds/chatgpt-and-ai-text-generators-should-academia-adapt-or-resist> (accessed 14 June 2023).
- Wittgenstein L. 1958 *Philosophical investigations*. Oxford, UK: Blackwell.
- Steels L. 1995 *A self-organizing spatial vocabulary*. *Artif. Life* **2**, 319–332. (doi:10.1162/artl.1995.2.3.319)
- Baronchelli A, Felici M, Loreto V, Caglioti E, Steels L. 2006 *Sharp transition towards shared vocabularies in multi-agent systems*. *J. Stat. Mech. Theory Exp.* **2006**, P06014. (doi:10.1088/1742-5468/2006/06/P06014)
- Sugden R. 1989 *Spontaneous order*. *J. Econ. Perspect.* **3**, 85–97. (doi:10.1257/jep.3.4.85)
- Centola D, Baronchelli A. 2015 *The spontaneous emergence of conventions: an experimental study of cultural evolution*. *Proc. Natl Acad. Sci. USA* **112**, 1989–1994. (doi:10.1073/pnas.1418838112)
- Schelling TC. 2006 *Micromotives and macrobehavior*. New York, NY: WW Norton & Company.
- Granovetter M. 1978 *Threshold models of collective behavior*. *Am. J. Sociol.* **83**, 1420–1443. (doi:10.1086/226707)
- Xie J, Sreenivasan S, Korniss G, Zhang W, Lim C, Szymanski BK. 2011 *Social consensus through the influence of committed minorities*. *Phys. Rev. E* **84**, 011130. (doi:10.1103/PhysRevE.84.011130)
- Centola D, Becker J, Brackbill D, Baronchelli A. 2018 *Experimental evidence for tipping points in social convention*. *Science* **360**, 1116–1119. (doi:10.1126/science.aas8827)
- Marwell G, Oliver P, Elster J, Hernes G. 1993 *The critical mass in collective action*. Cambridge, UK: Cambridge University Press.
- Centola DM. 2013 *Homophily, networks, and critical mass: solving the start-up problem in large group collective action*. *Ration. Soc.* **25**, 3–40. (doi:10.1177/1043463112473734)
- Barberá P, Wang N, Bonneau R, Jost JT, Nagler J, Tucker J, González-Bailón S. 2015 *The critical periphery in the growth of social protests*. *PLoS ONE* **10**, e0143611. (doi:10.1371/journal.pone.0143611)
- Iacopini I, Petri G, Baronchelli A, Barrat A. 2022 *Group interactions modulate critical mass dynamics in social convention*. *Commun. Phys.* **5**, 64. (doi:10.1038/s42005-022-00845-y)
- Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M. 2021 *The echo chamber effect on social media*. *Proc. Natl Acad. Sci. USA* **118**, e2023301118. (doi:10.1073/pnas.2023301118)