Technology in Society 76 (2024) 102480

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Colleges and universities are important stakeholders for regulating large language models and other emerging AI

Veljko Dubljević [a], [b]

[a] Science Technology and Society Program, North Carolina State University, 1911 Building, Suite 106 (Campus Box 7107), Raleigh, 27695, United States

[b] Department of Philosophy and Religious Studies, North Carolina State University, 101 Lampe Drive (Withers Hall 453), Raleigh, 27695, United States

ARTICLE INFO

Keywords:
Artificial intelligence (AI)
Ethics
Public policy
Legitimacy
Oversight

ABSTRACT

AI technology has already gone through one "winter," and alarmist thinking may cause yet another one. Calls for a moratorium on AI research increase the salience of the public request for comment on "AI accountability." Prohibitive approaches are an overreaction, especially when leveraged on virtual (non-embodied) AI agents. While there are legitimate concerns regarding expansion of AI models like ChatGPT in society, a better approach would be to forge a partnership between academia and industry, and utilize infrastructure of campuses to authenticate users and oversee new AI research. The public could also be involved with public libraries authenticating users. This staged approach to embedding AI in society would facilitate addressing ethical concerns, and implementing virtual AI agents in society in a responsible and safe manner.

New technology has always inspired feelings of both awe and dread, and rapid proliferation in society is invariably accompanied by both "hype and hope" and "gloom and doom" perspectives [1]. Artificial Intelligence (AI) is similar in this respect, but with a crucial difference: this technology has already gone through one "winter" [2]. Thus, it is important to avoid alarmist thinking about AI and to proceed with regulation guided not by dystopian or utopian flights of fancy, but by careful consideration of policies which would limit harms and promote benefits for all.

This, of course is easier said than done, as the most recent "AI Accountability Policy Request for Comment" in the U.S. attested to [3]. So, how can society manage the risks inherent in the development of AI without stifling innovation and perhaps causing a new AI winter? Recently, the Future of Humanity Institute issued an open letter, signed by many leaders in the technology industry, which asked for a six months long moratorium on new AI research while policy proposals are being crafted and considered [4]. While embodied AI agents, such as autonomous vehicles [5] and carebots [6], need to be more carefully regulated, and may even need to have ethical guidance functions prior to widespread deployment [7], in large extent due to the very real possibility of causing physical harm, it would be misguided to completely stop all AI research, including on language models and virtual AI assistants [8], *before* having a clear and feasible plan of action for developing accountability. To be fair, the Future of Humanity Institute [4] does provide several policy recommendations (including mandating third-party auditing, regulating access to computational power, establishing national AI agencies, establishing strict liability, watermarking, etc.).

However, all of these policy proposals heavily rely on two powerful stakeholders: industry and government(s). While civil society, academia and the public are noted as important stakeholders, they are offered no meaningful role in any of the proposed policies. That means that if such policies are implemented as recommended, this would widen existing power differences, further compromise trustworthiness of the industry and exacerbate mistrust in government. Heavy-handed regulation like the recent attempt at banning ChatGPT in Italy [9] does not truly solve the problems nor address legitimate concerns levelled against widespread use of Large Language Models (LLMs).

What are the ethical issues pertaining to the use of LLMs and other virtual (i.e., non-embodied) AI agents? The prohibitive response of Italy was issued because ChatGPT had "no way to verify the age of users," whereas it "exposes minors to absolutely unsuitable answers compared to their degree of development and awareness" [9]. Additional concerns regarding LLMs in general, and ChatGPT in particular, pertain to educational settings, where plagiarism and cheating by students could become rampant [10].

Therefore, any regulatory framework that doesn't engage educational institutions as important stakeholders will be missing the mark. One constructive way forward would be for the industry to partner with colleges and universities in granting access to LLMs and other virtual AI

E-mail address: veljko_dubljevic@ncsu.edu.

https://doi.org/10.1016/j.techsoc.2024.102480
Received 8 May 2023; Received in revised form 12 January 2024; Accepted 12 February 2024
Available online 18 February 2024

agents. Colleges and universities already have authentication infrastructures in place, notably to provide access to courses and educational resources, which virtually guarantees that all users would be of age and verified, and not by industry (which collects far too much data from the public as it is), but by institutions of higher education. This would also (at least in principle) enable tracking of any student activity involving LLMs which is contrary to student codes of conduct and help reduce plagiarism. In fact, such a policy would facilitate healthy innovation on our campuses and promote the culture of collaboration and teaming with AI agents [11].

There is one additional benefit of partnering the tech industry with colleges and universities. Unlike the tech industry, which has tried and failed to incorporate ethics teams [12], colleges and universities have full-fledged Institutional Review Boards (IRBs), which already serve the function of ethical and regulatory oversight of research. Indeed, the fact that IRBs have sufficient authority and independence [13] has contributed to their continued relevance and widespread social trust, despite shortcomings. Of course, IRBs face a variety of issues, but the core model can be extended and adapted to accommodate research that uses AI tools and techniques [13]. A key feature will be proportionality, as over-extension of ethics oversight would again be detrimental for innovation [13]. Arguably, the failure of industry ethics teams [12] stems from their lack of independence and authority, and the industry-academia partnership in AI has many features that would prevent this from becoming a mere exercise in "ethics washing." Again, this proposal only pertains to LLMs and virtual AI agents, as embodied AI agents raise unique ethical issues (for instance, they have already resulted in human deaths), and so must be designated as "high risk," with appropriate regulatory approaches [14].

One could forcefully object to any reduction of access to LLMs to the public, note that self-regulation or "corporate digital responsibility" are sufficient measures [15], and even criticize this proposal as elitist. After all, the public is a key stakeholder, and although institutions of higher education are an important part of society, the majority of the public wouldn't have immediate access as they have no affiliation with colleges or universities. The response to this (legitimate) critique is that any policy needs to be tested first, and that regulators should rely on colleges and universities in this regard. Future of Humanity Institute called for a 6-month moratorium in which there would be no public access [4]. In contrast to that, the policy proposal explained above envisions that colleges and universities (and all students, staff and faculty, and perhaps even alumni) would have full access while unintended consequences and potential for misuse are assessed by the industry, government and academia. Then, access could be increased to anyone holding a public library account. Similar to the case of colleges and universities, public libraries in most if not all municipalities have infrastructure in place to authenticate users and to provide access to library content, including databases. Thus, this staged approach to embedding AI in society would facilitate addressing ethical concerns, and implementing virtual AI agents and LLMs in society in a responsible and safe manner.

CRediT authorship contribution statement

Veljko Dubljević: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

None.

Data availability

No data was used for the research described in the article.

Acknowledgments

The author is a University Faculty Scholar and Associate Professor at NC State University where he also leads the NeuroComputational Ethics Research Group.

Special thanks to the faculty affiliates of the AI in Society group for helpful comments on an earlier draft of this paper.

References

[1] V. Dubljević, Neuroethics, Justice and Autonomy: Public Reason in the Cognitive Enhancement Debate, Springer, 2019.
[2] B.C. Smith, The Promise of Artificial Intelligence: Reckoning and Judgment, MIT Press, 2019.
[3] United States Department of Commerce, National telecommunications and information administration, AI accountability policy request for comment; https://ntia.gov/issues/artificial-intelligence/request-for-comments..
[4] Future of Life Institute, Policymaking in the pause. https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf, April 12, 2023.
[5] J.-F. Bonnefon, et al., The social dilemma of autonomous vehicles, Science 352 (6293) (2016) 1573–1576.
[6] A. Coin, V. Dubljević, Carebots for Eldercare: Technology, Ethics, and Implications" in Trust in Human-Robot Interactions, Elsevier, 2020, pp. 553–569.
[7] V. Dubljević, Toward implementing the ADC model of moral judgment in autonomous vehicles, Sci. Eng. Ethics 26 (2020) 2461–2472.
[8] W.A. Bauer, V. Dubljević, AI assistants and the paradox of internal automaticity, Neuroethics 13 (2019) 303–310.
[9] S. McCallum, ChatGPT Banned in Italy Over Privacy Concerns, BBC News, April 1, 2023. https://www.bbc.com/news/technology-65139406.
[10] B. McMurtrie, Teaching: will ChatGPT change the way you teach? Chron. High Educ. (January 5, 2023). https://www.chronicle.com/newsletter/teaching/2023-01-05.
[11] M. Pflanzer, et al., Ethics of Human-AI Teaming: Principles and Perspectives, AI and Ethics, 2022, https://doi.org/10.1007/s43681-022-00214-z.
[12] R. Bellan, Microsoft Lays Off an Ethical AI Team as it Doubles Down on Open AI, TechCrunch, March 13, 2023. https://techcrunch.com/2023/03/13/microsoft-lays-off-an-ethical-ai-team-as-it-doubles-down-on-openai.
[13] P. Friesen, et al., Governing AI-driven health research: are IRBs up to the task? Ethics Hum. Res. 43 (2) (2021) 35–42.
[14] V. Dubljević, et al., Toward a rational and ethical sociotechnical system of autonomous vehicles: a novel application of multi-criteria decision analysis, PLoS One (2021), https://doi.org/10.1371/journal.pone.0256224.
[15] L. Lobschat, et al., Corporate digital responsibility, J. Bus. Res. 122 (2021) 875–888, https://doi.org/10.1016/j.jbusres.2019.10.006.