



## OPEN ACCESS

EDITED BY  
Kaleem Siddiqi,  
McGill University, Canada

REVIEWED BY  
Kristof Van Laerhoven,  
University of Siegen, Germany

\*CORRESPONDENCE  
Virginia Dignum  
✉ virginia@cs.umu.se

RECEIVED 22 April 2023  
ACCEPTED 28 April 2023  
PUBLISHED 18 May 2023

CITATION  
Baum K, Bryson J, Dignum F, Dignum V,  
Grobelnik M, Hoos H, Irgens M, Lukowicz P,  
Muller C, Rossi F, Shawe-Taylor J, Theodorou A  
and Vinuesa R (2023) From fear to action: AI  
governance and opportunities for all.  
*Front. Comput. Sci.* 5:1210421.  
doi: 10.3389/fcomp.2023.1210421

COPYRIGHT  
© 2023 Baum, Bryson, Dignum, Dignum,  
Grobelnik, Hoos, Irgens, Lukowicz, Muller,  
Rossi, Shawe-Taylor, Theodorou and Vinuesa.  
This is an open-access article distributed under  
the terms of the [Creative Commons Attribution  
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# From fear to action: AI governance and opportunities for all

Kevin Baum<sup>1</sup>, Joanna Bryson<sup>2</sup>, Frank Dignum<sup>3</sup>, Virginia Dignum<sup>3\*</sup>, Marko Grobelnik<sup>4</sup>, Holger Hoos<sup>5</sup>, Morten Irgens<sup>6</sup>, Paul Lukowicz<sup>7</sup>, Catelijne Muller<sup>8</sup>, Francesca Rossi<sup>9</sup>, John Shawe-Taylor<sup>10</sup>, Andreas Theodorou<sup>11</sup> and Ricardo Vinuesa<sup>12</sup>

<sup>1</sup>Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) and Algoright, Kaiserslautern, Germany, <sup>2</sup>Hertie School, Berlin, Germany, <sup>3</sup>Umeå University, Umeå, Sweden, <sup>4</sup>OECD, Paris, France, <sup>5</sup>Aachen University, Aachen, Germany, <sup>6</sup>CLAIRE-AI.org, Oslo Metropolitan University, Oslo, Norway, <sup>7</sup>Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Kaiserslautern, Germany, <sup>8</sup>ALLAI, Amsterdam, Netherlands, <sup>9</sup>IBM, Yorktown, NY, United States, <sup>10</sup>IRCAI, International Research Institute on AI, Ljubljana, Slovenia, <sup>11</sup>VerAI, Umeå, Sweden, <sup>12</sup>KTH Royal Institute of Technology, Stockholm, Sweden

## KEYWORDS

Artificial Intelligence, governance, responsible AI, Trustworthy AI, large language models, generative AI

OpenAI's GPT-4 (OpenAI, 2023) reignited the public discussions regarding Artificial Intelligence (AI) and its risks. In a recent open letter (Future of Life Institute, 2023) technology leaders (including Elon Musk and Steve Wozniak), prominent academics (including Yoshua Bengio and Stuart Russell) and many others, call for a 6-months "pause" of "giant AI experiments", or more precisely, a pause in the training of AI systems more powerful than GPT-4. The letter has sparked much needed broad public discussion, but has also led to unhelpful debates on matters such as who did and did not sign, the goals and intentions of the founders of the Future of Life institute and speculations about hypothetical artificial general intelligence (AGI) and its capabilities.

We welcome the public discussion the letter has generated, but we also see an urgent need to move beyond those debates; in fact, it is crucial to address current problems with the development and use of AI. As it has been our continued position for years,<sup>1</sup> we advocate greater support for, and engagement with, the ongoing comprehensive debate and regulatory actions concerning all aspects related to the impact, development, use, and governance of advanced AI systems, whether generative or not. "Efforts towards shared safety protocols for advanced AI design and development that are rigorously audited and overseen by independent outside experts", as is proposed in the open letter, should not be left to AI labs, nor is it something that can be done once and subsequently deemed fixed forever. It requires a concerted, continuous and global effort.

Going forward, it is important to realize that progress is already taking place, and to take note of the many efforts that are being made in this regard. Such efforts have been pursued since 2019 by several international organizations, including the European Union, the Council of Europe, the OECD and UNESCO. The EU released its Guidelines for Trustworthy AI in 2019 (European Commission, 2019), which served as a stepping stone toward AI regulation (AI Act) that is now in its final stages, and the Council of Europe is well underway with its negotiations on an AI treaty, which it hopes the US, Canada and Japan will also sign; the OECD adopted principles on Artificial Intelligence in 2019 (OECD, 2019), and UNESCO delivered an agreement on Ethics of AI in 2021 (UNESCO, 2022). Most

1 E.g., several of us are or have been involved in the European High Level Expert Group on AI, UNESCO, OECD, Council of Europe, Confederation of Laboratories for Artificial Intelligence in Europe (CLAIRE) and other initiatives.

recently, the United States of America released the “Blueprint for an AI Bill of Rights” (White House, 2022), setting the scene for US federal regulations. Finally, standardization bodies have already released standards related to the development and use of AI systems, e.g., the IEEE SA’s 7001 on transparency in AI systems (Winfield et al., 2022) and the ISO JTC1/SC42 standards on robustness, supported by the CEN and CENELEC (Winter et al., 2021).

At the same time, the much needed novel approaches to ensure safe, trustable and beneficial AI applications will partly come from AI research and development. More than ever we need to discuss how AI research is brought about. Who funds scientific research development, and who owns the products of that work? Is there a need for publicly owned and controlled large language models (LLM) like GPT-4? How do we address the concentration of so much power derived from new technologies with only a selected few private companies? Monopolization is a negative pattern of inadequate governance, as we have seen repeated for more than a century, leading to more unfairness, participation gaps and power imbalance.

Research efforts are needed on the auditing of novel AI techniques, such as LLMs, which poses a number of challenges (Mökander et al., 2023), including the management of very complex AI pipelines, assigning responsibilities and accountability to a wide variety of actors across the whole process. Moreover, ensuring coherent, concise and applicable regulation is of utmost importance for scientific advancements in AI to be turned into real-life applications with realistic and practical implementation without harming the planet and society.

At the same time, efforts are needed to address the challenge of designing regulation that is future-proof, also when faced with the rapid pace of research developments. As with any political process of a complex subject matter, this must be done in discussions with subject experts from academia, industry and civil society. Responsible development and use of AI is not a zero-sum game: either by regulation, or by industry efforts. Rather, a concerted effort bringing together, and keeping together, all stakeholders is needed.

As much as addressing concerns about the use of generative AI, it is important to discuss and incentivize positive use of transformative technology. This includes improving on current models toward accountable, responsible and transparent practices in development and deployment. Focusing on speculation about imagined futures in which current AI developments may render humans obsolete or lead to catastrophic loss of control of our civilization detracts from the very serious problems current AI technologies and their applications are likely to cause. Rational debate on the long(er)-term perspective including hypothetical future AGI capabilities—grounded in facts and hard science (including social sciences)—should indeed be part of the overall social and political discourse on the benefits and potential risks of AI. However, we should be careful not to create an aura of “mysticism” around AI, nor a perception of inevitability. Techno-determinism is a myth that undermines participation in governance, and commitment to government and enforcement (Jelinek et al., 2021).

A temporary halt in “giant AI experiments” that are more powerful than GPT-4 will not *per se* result in a human-centered, ethical, and socially beneficial development and use of existing and future AI-systems. Only through increased efforts on and support of research, in a coordinated spirit of collaboration, can the perceived, recognized, and as yet unknown risks of advanced AI technologies be effectively addressed. These risks include bias, inequality, exclusion, erosion of privacy, deterioration of human dignity, autonomy and agency, environmental impact, corruption, disruption of security and cooperation, and insufficient product control to ensure agreed safeguards. Precisely because we consider these issues of utmost importance, we think that any remediating effort at this moment can best focus on limiting and/or guiding the use of (already) impactful AI.

For example, the danger of amplifying propaganda and disinformation through large language models (LLMs) should be taken very seriously, as should the major impact they are likely to have on our jobs and skills. At the same time, we need to address the environmental impact of those models (Vinuesa et al., 2020) and their lack of transparency and explainability. Addressing these concerns, as well as the need for a deeper understanding of the consequences and risks of how LLMs process human prompts and access information, are aspects that require urgent and continued multidisciplinary research and development efforts.

It is also interesting to note that even if further development of these large models is stopped, their growing socio-economic effect will likely remain (Eloundou et al., 2023). Moreover, all AI technologies need to be considered, not just the *generative or large*, or the ones that are “*more powerful than GPT-4*” including those that are currently exploiting versions of LLM, such as AutoGPT and BabyAGI. Just because the linguistic character of LLMs is particularly anthropomorphic and therefore appeals to the public’s imagination and fear, as well as to the potential overestimation of the capabilities of these models, this does not necessarily imply that these models are inherently or overall more dangerous than others. In order to guarantee that regulatory efforts can be operationalized, it is crucial to focus on transparency and explainability in AI and consider other approaches that can incorporate human-understandable formulation of knowledge and operational requirements.

A real risk of particular concern is the use of LLMs for programming—an activity where correctness matters precisely, because errors can be propagated quickly and widely, with far-reaching consequences. A combination of “deep reasoning” and good, documented, iterative development practice is required to achieve that correctness. “Almost correct” solutions can be more detrimental than obviously incorrect ones. Therefore, the use of LLMs for programming tasks should be approached with an abundance of caution.

The biggest risk is not that “AGI implies strong AI getting out of control”, but that relatively weak AI systems (or other digital technologies) cause economic and societal damage as a result of being developed or used in irresponsible ways. These might include impingement of citizens’ rights, impact on fairness, concentration of wealth and power, the undermining of democratic processes and the rule of law, mass manipulation through new and effective forms

of targeted information, as well as safety and security concerns. It is high time to move beyond the fear of AI, and rather to deploy well-documented, transparent AI methods and systems as part of our focus on the real regulatory and governance questions at hand, as well as to focus societal investment in this direction.

AI systems and tools should be seen as products, not something you can really negotiate with, trust or love. As any other product, they require safeguards of product safety, liability, accountability and control. As a technology with key impact on all sectors of industry and all areas of our lives, AI needs to be approached with particular care and foresight, with the public interest in mind.

A focus on a “mystical” perception of AI distorts the public view of future risks. Informed public oversight is necessary to legitimize and indeed improve these efforts. Making generative AI open to the wider research community enables the proper assessment of their associated risks, and contributes to scientific advancements. It will also benefit from society’s collective intelligence to identify challenges and operation shortcomings in a much more efficient way than any internal quality-assurance process.

What is needed is not a pause in development, but governance. This includes increased regulatory enforcement of accountability, responsibility, and transparency throughout the entire AI lifecycle, the development of appropriate institutions, as well as further collaborative multi-stakeholder initiatives to assess, measure, and mitigate the risks of models, techniques, and applications.

The processes of regulating and governing AI are not new. Many current initiatives far outstrip the needs mentioned in the Open Letter. This legislative, regulatory, and diplomatic work may or may not *slow* development, but more importantly it will provide protective oversight and democratic legitimacy, as well as a level playing field for all those trying to develop and use AI responsibly.

Ensuring that AI is employed for the benefit of people and planet will require broad participation. It is of utmost importance to establish a constructive, collaborative, and scientific approach across disciplines that ensures that technological development goes together with a deep knowledge of humanities and social sciences (Dignum, 2019). We stand with other international organizations, including, e.g., the AAAI (Rossi et al., 2023) and CLAIRE (see [claire-ai.org/vision](http://claire-ai.org/vision)), on a plea for collaboration with a broad spectrum of stakeholders, academia, industry, government, civil society, which will improve our understanding and build a rich system of collaborations to ensure the responsible development and use of AI technologies. Nevertheless, any call for more

governance and interdisciplinary work must be accompanied by major investments on responsible development and use of AI. Without such investments, we risk that even more than now, ownership and power are placed in large tech companies outside democratic scrutiny, leading to a global loss of prosperity and independence of citizens, societies and public administrations.

The real issue with generative AI systems is not whether they are close to AGI, or that AGI may do great damage, but that current systems and those we can expect in the near future can easily lure people into believing that they understand and trust them more than they should, into overestimating their capabilities, underestimating their weaknesses and limitations, and as a result, into using them in problematic and potentially harmful ways.

## Author contributions

Authors are listed in alphabetical order. The writing effort was coordinated by VD and the original idea was by RV. All authors contributed equally to the article.

## Funding

The European AI networks of Excellence HumaneAI-net and VISION have contributed to this work.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Dignum, V. (2019). AI is multidisciplinary. *AI Matt.* 5, 18–21. doi: 10.1145/3375637.3375644
- Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*.
- European Commission (2019). *Ethics Guidelines for Trustworthy AI*. European Commission, Directorate-General for Communications Networks, Content and Technology.
- Future of Life Institute (2023). *Pause Giant AI Experiments: An Open Letter*. Available online at: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> (accessed May 4, 2023).
- Jelinek, T., Wallach, W., and Kerimi, D. (2021). Policy brief: the creation of a G20 coordinating committee for the governance of artificial intelligence. *AI Ethics* 1, 141–150. doi: 10.1007/s43681-020-00019-y
- Mökander, J., Schuett, J., Kirk, H. R., and Floridi, L. (2023). Auditing large language models: a three-layered approach. *arXiv*. doi: 10.2139/ssrn.4361607
- OECD (2019). *Recommendation of the Council on Artificial Intelligence, Organisation for Economic Co-operation and Development*.
- OpenAI (2023). *GPT-4 Technical Report*.
- Rossi, F., Smith, S., Selman, B., Gil, Y., Kambhampati, S., Dietterich, T., et al. (2023). *Working Together on our Future With AI*. Association for the Advancement of Artificial Intelligence (AAAI). Available online at: <https://aaai.org/working-together-on-our-future-with-ai/> (accessed May 4, 2023).
- UNESCO (2022). *Recommendation on the Ethics of Artificial Intelligence*. SHS/BIO/PI/2021/1. Available online at: <https://unesdoc.unesco.org/ark:/48223/pf0000381137> (accessed May 4, 2023).

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., et al. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nat. Commun.* 11, 233. doi: 10.1038/s41467-019-14108-y

White House (2022). *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. White House Office of Science and Technology Policy. Available online at: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/> (accessed May 4, 2023).

Winfield, A. F., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., et al. (2022). IEEE P7001: a proposed standard on transparency. *Front. Robot. AI* 8, 665729. doi: 10.3389/frobt.2021.665729

Winter, P. M., Eder, S., Weissenböck, J., Schwald, C., Doms, T., Vogt, T., et al. Trusted artificial intelligence: Towards certification of machine learning applications. *arXiv preprint arXiv:2103.16910* (2021). doi: 10.48550/arXiv.2103.16910