Collective intelligence is needed to ensure beneficial Artificial Intelligence

Lahlou, Saadi [1 2] Las Casas, Estevam [3] Bouin, Olivier [4 5 6] Boucekkine, Raouf [7] Rabinovici, Eliezer [8 9 10] Candiotto, Laura [11] Copeland, Jack [12] Cunha, Evandro [13] Di Luca, Massimiliano [14] Frassinelli, Diego [15] Fried, Itzhak [16] Fujita, Andre [17] Fukuda, Toshio [18 19] Growiec, Jakub [20 20] Guedj, Benjamin [21 22] Kasturi, Suranga [23 24] Kellmeyer, Phillip [25 26] Livermore, Michael [27] Mézard, Marc [28] Moodley, Deshen [29 30] Nowotny, Helga [31] Paré, Zaven [32] Plonski, Ary Guilherme [17 33] Rabinowitch, Ithai [34] Ralitera, Talina [35] Rees, Martin [36] Shagrir, Oron [37] De Soarez, Patricia [17] Taylor, Henry [38] Wevers, Melvin [39] Yasutomo, Kawanishi [40 41]

[1] *Paris Institute for Advanced Study,* [2] *London School of Economics,* [3] *Institute for Advanced Study of Belo Horizonte,* [4] *Foundation-Excellence Laboratory RFIEA,* [5] *EASSH (European Alliance for Social Sciences and Humanities),* [6] *NetIAS,* [7] *Rennes School of Business,* [8] *CERN Council (European Organization for Nuclear Research),* [9] *Leon H. and Ada G. Miller Chair of Science,* [10] *Racah Institute of Physics, Hebrew University of Jerusalem,* [11] *Centre for Ethics of the University of Pardubice, Czech Republic,* [12] *University of Canterbury in New Zealand,* [13] *Faculty of Letters of the Federal University of Minas Gerais (UFMG),* [14] *University of Birmingham School of Psychology,* [15] *University of Konstanz,* [16] *University of California, Los Angeles (UCLA),* [17] *University of São Paulo,* [18] *Nagoya University,* [19] *Institute of Electrical and Electronics Engineers,* [20] *Warsaw School of Economics,* [21] *Inria (France),* [22] *Centre for AI at University College London,* [23] *Center for Biomedical Informatics at Regenstrief Institute,* [24] *Indiana University School of Medicine,* [25] *University Medical Center Freiburg (UKF),* [26] *Freiburg Institute for Advanced Studies (FRIAS),* [27] *University of Virginia,* [28] *École Normale Supérieure (ENS) in Paris,* [29] *University of Cape Town,* [30] *South African national Centre for Artificial Intelligence Research,* [31] *ETH Zurich,* [32] *Universidade Federal de Juiz de Fora (IAD/UFJF),* [33] *UBIAS,* [34] *Faculty of Medicine of the Hebrew University of Jerusalem,* [35] *French Alternative Energies and Atomic Energy Commission (CEA),* [36] *Centre for the Study of Existential Risk (CSER) at the University of Cambridge,* [37] *Hebrew University of Jerusalem,* [38] *University of Birmingham, UK,* [39] *University of Amsterdam,* [40] *Nagoya University ,* [41] *RIKEN Guardian Robot Project,*

DOI 10.5281/zenodo.13588574

TO CITE

Lahlou, S., Las Casas, E., Bouin, O., Boucekkine, R., Rabinovici, E., Candiotto, L., Copeland, J., Cunha, E., Di Luca, M., Frassinelli, D., Fried, I., Fujita, A., Fukuda, T., Growiec, J., Guedj, B., Kasturi, S., Kellmeyer, P., Livermore, M., Mézard, M., … Yasutomo, K. (2023). Collective intelligence is needed to ensure beneficial Artificial Intelligence. In *Proceedings of the Paris Institute for Advanced Study* (Vol. 12). https://doi.org/10.5281/zenodo.13588574

PUBLICATION DATE

14/06/2023

ABSTRACT

The fourth Intercontinental Academia (ICA4) "Intelligence and artificial intelligence" (2020-2021) identified four thematic priorities for future AI research: (1) developing an ontology of AI systems to assess their agency in performing cognitive and behavioural tasks; (2) addressing the challenges of human-artificial agent interaction (H2AI); (3) clarifying the values used by artificial agents and determine their legal status in society; and (4) regulating artificial intelligent agents (AIAs) with appropriate certifications to safeguard human agency and well-being. These priorities require research programmes and funding schemes that involve scientists with complementary perspectives and methodological approaches (neuroscience, computer science, mathematics, engineering, humanities and social sciences).

The fourth Intercontinental Academia (ICA4) "Intelligence and artificial intelligence" convened 40 top scientists to explore the current and future challenges of artificial intelligence[1]. Over two 10-days in person meetings, one-week online meeting and several specialized working groups that met between June 2020 and November 2021, the

Lahlou, S., Las Casas, E., Bouin, O., Boucekkine, R., Rabinovici, E., Candiotto, L., Copeland, J., Cunha, E., Di Luca, M., Frassinelli, D., Fried, I., Fujita, A., Fukuda, T., Growiec, J., Guedj, B., Kasturi, S., Kellmeyer, P., Livermore, M., Mézard, M., … Yasutomo, K. (2023). Collective intelligence is needed to ensure beneficial Artificial Intelligence. In *Proceedings of the Paris Institute for Advanced Study* (Vol. 12). https://doi.org/10.5281/zenodo.13588574
2023/13 - Intercontinental Academia 4 - Article No.32. Freely available at https://paris.pias.science/article/collective-intelligence-is-needed-to-ensure-beneficial-artificial-intelligence - ISSN 2826-2832/

pluridisciplinary group discussed the state of the art in natural and artificial intelligence and proposed a future agenda for future AI research and its societal impacts. ICA4 was organised by the international network of University-Based Institutes for Advanced Study (UBIAS), and hosted by the Paris Institute for Advanced Study and the Institute for Transdisciplinary Advanced Study of Belo Horizonte. The ICA4 scientific advances are pursued through various formats, including participation in high-level initiatives such as the Artificial Intelligence Summit (from April 2023 onwards) and several scientific residencies for experienced fellows during the 2023-2024 academic year.

The accelerating pace of innovation in artificial Intelligence (AI) opens many fascinating scientific and practical opportunities. Current AI technologies already show remarkable empirical performance in a growing number of tasks. Expert systems based on Large Language Models, driverless vehicles, AI-assisted justice, communication with and support by Artificial Intelligent Agents (AIAs) in cognitive impairment and dementia bring benefits; they also raise serious societal and ethical issues [2]. The breadth and adaptability of AI applications are rapidly increasing as these capabilities are combined and become pervasive.

Before we reach a tipping point through large-scale deployment of transformative AI technologies, with the risk of humans losing control of their own affairs, some crucial questions need to be addressed and resolved. Over the past two years, an interdisciplinary group of forty scholars from leading institutions around the world gathered to reflect on the priorities in research and policies to enable AI technologies to serve the common good and benefit humankind globally [3].

Thematic priorities for AI research

The ICA4 identified four thematic priorities for future AI research: (1) developing an ontology of AI systems to assess their agency in performing cognitive and behavioural tasks; (2) addressing the challenges of human-artificial agent interaction (H2AI); (3) clarifying the values used by artificial agents and determine their legal status in society; and (4) regulating artificial intelligent agents (AIAs) with appropriate certifications to safeguard human agency and well-being. These priorities require research programmes and funding schemes that involve scientists with complementary perspectives and methodological approaches (neuroscience, computer science, mathematics, engineering, humanities and social sciences).

Lahlou, S., Las Casas, E., Bouin, O., Boucekkine, R., Rabinovici, E., Candiotto, L., Copeland, J., Cunha, E., Di Luca, M., Frassinelli, D., Fried, I., Fujita, A., Fukuda, T., Growiec, J., Guedj, B., Kasturi, S., Kellmeyer, P., Livermore, M., Mézard, M., … Yasutomo, K. (2023). Collective intelligence is needed to ensure beneficial Artificial Intelligence. In *Proceedings of the Paris Institute for Advanced Study* (Vol. 12). https://doi.org/10.5281/zenodo.13588574
2023/13 - Intercontinental Academia 4 - Article No.32. Freely available at https://paris.pias.science/article/collective-intelligence-is-needed-to-ensure-beneficial-artificial-intelligence - ISSN 2826-2832/

At present, the term AI is a fashionable label loosely applied to very different processes and systems, from data mining to unsupervised learning. The first priority is to **develop an ontology of AI systems**. In information science, an ontology is a system for naming and representing entities, categories, properties and relations relevant to a given discourse. Formal ontologies facilitate reasoning and collaboration by clearly defining the concepts used to model an object of study or engineering project. Current AI taxonomy efforts typically distinguish systems by algorithmic properties: learning mode, risk level, functions, or output. These categories are valuable, but should be extended to include socially relevant dimensions such as level of autonomy and domain of action. An AI ontology should define the functions, purposes and values embedded in AI systems. It should integrate further research to better understand human 'intelligence' and how it relates to human values and affects.

This suggests that we should refer to specific and well-defined Artificial Intelligent Agents (AIAs), with explicit functions and specifications. For example, one should distinguish between the ability to generate formally correct texts, which current large language models approach, and the ability to construct a representation of the world and to interact with it. The current loose use of the term "AI" risks creating a society in which non-human entities are attributed "intelligence" and human-like personality and agency, without first clarifying the limits of their domain of responsibility. This already applies to education, health, finance, justice, employment, control, security and even to private matters and affects.

Many efforts are underway to make AI more ethical or beneficial and to align it with humanistic values. These efforts need to address the different stages of design, development and deployment. At the level of the final product and its effective performance, certification and legislation must address the intended and unintended uses of AI, just as for any other technology. An operational ontology is therefore needed to establish standards and certifications for the AI systems that will be released to the public.

The second priority is to **address the challenges of Human to Artificial Agent Interaction (H2AI)**, in particular in terms of collaboration, delegation and responsibility. H2AI goes beyond traditional Human-Computer Interaction (HCI). AIAs do not only respond to human input, but also have increased autonomy to make decisions with little to no human supervision. This leads to a blurring of the lines of responsibility. Perhaps in no other field is this blurring more critical than in medicine and clinical

Lahlou, S., Las Casas, E., Bouin, O., Boucekkine, R., Rabinovici, E., Candiotto, L., Copeland, J., Cunha, E., Di Luca, M., Frassinelli, D., Fried, I., Fujita, A., Fukuda, T., Growiec, J., Guedj, B., Kasturi, S., Kellmeyer, P., Livermore, M., Mézard, M., … Yasutomo, K. (2023). Collective intelligence is needed to ensure beneficial Artificial Intelligence. In *Proceedings of the Paris Institute for Advanced Study* (Vol. 12). https://doi.org/10.5281/zenodo.13588574
2023/13 - Intercontinental Academia 4 - Article No.32. Freely available at https://paris.pias.science/article/collective-intelligence-is-needed-to-ensure-beneficial-artificial-intelligence - ISSN 2826-2832/© 2025 The authors

neuroscience, where implanted neuroprosthetic devices and brain-computer interfaces already operate in hybrid platforms. In the future, AIAs developed for clinical needs and to augment human potential may affect human autonomy in ways that are not fully predictable.

Questions also arise about the accountability of AIAs and their ability to communicate and share meaning with their users and/or targets. This is all the more important as AIAs do not yet "perceive" or "understand", nor can they be judged as responsible intentional agents, but some already recognise and label patterns and individuals, and even make decisions. We call for these boundaries of responsibility to be exposed in the H2AI specifications, with clearer mechanisms and criteria. This requires a better understanding of the involvement of AIAs - their functioning, performance and limitations - and how these systems are integrated with humans in hybrid decision-making structures and institutions. For example, when AI algorithms are used by financial institutions to evaluate, rank, and approve loan applications, we advocate for stating on the application forms: "This application will be assessed by an algorithm with no direct human supervision".

The third priority addresses the challenge of integrating AI systems into specific social, ethical and cultural contexts. This is the issue of **AI "citizenship"** or digital humanism [4]. As new agents are introduced into our society, the question of how we integrate them becomes crucial. How should accountability be distributed among designers, owners and users of AIAs? What behaviours are acceptable and desirable, what and whose values are behind the choices and decisions that determine how we want to live well together? The involvement of the humanities and social sciences is thus indispensable for value alignment in AIAs. This typically requires a better understanding of how people themselves deal with goals, purpose and values in decision-making. Recognition of the social and cultural diversity of societies, evolution of norms over time, and the diversity of contexts in which AIAs will operate must be translated into technically feasible, legally protected, pragmatic and humanly acceptable ways. This raises difficult questions because of gradation and context-sensitivity of functions and dispositions. Addressing these and related issues and finding solutions that work in practice is essential for the successful embedding of AIAs in societies.

The fourth challenge identified is the **regulation of AIAs for individual and societal well-being and flourishing**. Systems of regulation exist at a wide range of scales - from the metabolic homeostasis of organisms to the global regulation of financial markets.

Lahlou, S., Las Casas, E., Bouin, O., Boucekkine, R., Rabinovici, E., Candiotto, L., Copeland, J., Cunha, E., Di Luca, M., Frassinelli, D., Fried, I., Fujita, A., Fukuda, T., Growiec, J., Guedj, B., Kasturi, S., Kellmeyer, P., Livermore, M., Mézard, M., … Yasutomo, K. (2023). Collective intelligence is needed to ensure beneficial Artificial Intelligence. In *Proceedings of the Paris Institute for Advanced Study* (Vol. 12). https://doi.org/10.5281/zenodo.13588574
2023/13 - Intercontinental Academia 4 - Article No.32. Freely available at https://paris.pias.science/article/collective-intelligence-is-needed-to-ensure-beneficial-artificial-intelligence - ISSN 2826-2832/

Any regulatory system requires mechanisms for implementation, monitoring, evaluation and feedback. Regulating AI will require technical and legal systems capable of continuously monitoring the performance of AI as it is deployed in real-world contexts, identifying unintended or malicious behaviour, and coordinating control measures to prevent harm to humans and other undesirable outcomes. Such regulatory systems will require adaptive governance technologies that can keep pace with the increasing sophistication of AIAs. As AIAs will also require maintenance and repair to perform as expected, processes for certifying the compliance of AIAs will need to be established. People are certified by diplomas and professional qualifications; they are subject to law: what will be the equivalent for AIAs, which we now allow to advise, decide and sometimes act?

“Intelligence” as a transdisciplinary object of study

"Intelligence" is a multidimensional concept. There is belief that 'intelligence' of AI is different from human intelligence. And yet this difference invites the study of 'intelligence' as a transdisciplinary object. Fields as diverse as neuroscience, (computational) linguistics, philosophy, psychology and economics (among many others) are essential to better understand how communication and cooperation between these different forms of 'intelligence' can best be achieved. As discussed above, addressing these issues requires interdisciplinary scientific research.

The experience of ICA4 has shown that such interdisciplinary and multicultural collaboration is possible and fruitful. It is now time to mobilize the scientific community as a whole, including social sciences and humanities whose knowledge and potential contributions have so far been under-utilised. Scientific institutions, research networks and learned societies are urgently called upon to develop new methods and concepts beyond existing disciplinary boundaries, on a scale that will allow a humanistic integration of AI into our societies. The 80+ Institutes for Advanced Study worldwide will actively contribute to the global collective intelligence that is needed to shape the appropriate dynamic framework to ensure that AI will be and remain beneficial. The society as a whole –government, business, citizens-- needs the decisive contribution of science, and it needs it now.

Lahlou, S., Las Casas, E., Bouin, O., Boucekkine, R., Rabinovici, E., Candiotto, L., Copeland, J., Cunha, E., Di Luca, M., Frassinelli, D., Fried, I., Fujita, A., Fukuda, T., Growiec, J., Guedj, B., Kasturi, S., Kellmeyer, P., Livermore, M., Mézard, M., … Yasutomo, K. (2023). Collective intelligence is needed to ensure beneficial Artificial Intelligence. In *Proceedings of the Paris Institute for Advanced Study* (Vol. 12). https://doi.org/10.5281/zenodo.13588574
2023/13 - Intercontinental Academia 4 - Article No.32. Freely available at https://paris.pias.science/article/collective-intelligence-is-needed-to-ensure-beneficial-artificial-intelligence - ISSN 2826-2832/

Footnotes

1 : The participants in ICA4 were as follows. **Chairs and organizers**: Saadi Lahlou (Paris Institute for Advanced Study and London School of Economics); Estevam Barbosa de Las Casas (Institute for Advanced Study of Belo Horizonte); Olivier Bouin (Laboratoire d'Excellence Réseau Français des Instituts d'études avancées); Raouf Boucekkine (Rennes School of Business and Centre for Unframed Thinking); Eliezer Rabinovici (European Organization for Nuclear Research and Hebrew University of Jerusalem). **Steering committee**: Ary Guilherme Plonski (University of São Paulo and University Based Institutes for Advanced Study); Sue Gilligan (Birmingham Institute of Advanced Studies). **Mentors**: Philippe Aghion (Collège de France and London School of Economics); Robert Aumann (Hebrew University of Jerusalem); Jack Copeland (University of Canterbury, New Zealand); Itzhak Fried (University of California, Los Angeles); Toshio Fukuda (Nagoya University and IEEE); William Hopkins (Georgia State University); Marc Mézard (Ecole Normale Supérieure-PSL, Paris); Melanie Mitchell (Santa Fe Institute); Helga Nowotny (ETH Zurich); Zaven Paré (Universidade Federal de Juiz de Fora); Martin Rees (University of Cambridge); Oron Shagrir (Hebrew University of Jerusalem); Shimon Ullman (Weizmann Institute of Science); Xiao-Jing Wang (New York University); Karen Yeung (University of Birmingham), Ada Yonath (Weizmann Institute of Science); Robert Zatorre (McGill University). **Fellows**: Laura Candiotto (University of Pardubice); Alex Cayco Gajic (École Normale Supérieure- PSL, Paris); Patricia Coelho de Soarez (University of São Paulo);Evandro Cunha (Federal University of Minas Gerais); Massimiliano Di Luca (University of Birmingham); Diego Frassinelli (University of Konstanz); Andre Fujita (University of São Paulo); Jakub Growiec (Warsaw School of Economics); Benjamin Guedj (Inria, France and University College London); Suranga Kasturi (Indiana University); Yasutomo Kawanishi (Nagoya University); Philipp Kellmeyer (University of Freiburg); Michael Livermore (University of Virginia); Deshen Moodley (University of Cape Town); Ithai Rabinowitch (Hebrew University of Jerusalem); Talina Ralitera (French Alternative Energies and Atomic Energy Commission); Oksana Stalnov (Israel Institute of Technology); Henry Taylor (University of Birmingham); Melvin Wevers (University of Amsterdam)↩

2 : https://www.nature.com/articles/d41586-023-00288-7 - https://www.nature.com/articles/d41586-023-03266-1 - https://futureoflife.org/open-letter/pause-giant-ai-experiments/↩

3 : The fourth Intercontinental Academia (ICA4) "Intelligence and artificial intelligence" was organised in 2021/2022 by the international network of University-Based Institutes for Advanced Study (UBIAS). https://www.intercontinental-academia.org/↩

4 : https://dighum.ec.tuwien.ac.at/dighum-manifesto↩

Lahlou, S., Las Casas, E., Bouin, O., Boucekkine, R., Rabinovici, E., Candiotto, L., Copeland, J., Cunha, E., Di Luca, M., Frassinelli, D., Fried, I., Fujita, A., Fukuda, T., Growiec, J., Guedj, B., Kasturi, S., Kellmeyer, P., Livermore, M., Mézard, M., … Yasutomo, K. (2023). Collective intelligence is needed to ensure beneficial Artificial Intelligence. In *Proceedings of the Paris Institute for Advanced Study* (Vol. 12). https://doi.org/10.5281/zenodo.13588574
2023/13 - Intercontinental Academia 4 - Article No.32. Freely available at https://paris.pias.science/article/collective-intelligence-is-needed-to-ensure-beneficial-artificial-intelligence - ISSN 2826-2832/