AI & SOCIETY
https://doi.org/10.1007/s00146-023-01810-1

EDITORIAL

Artificial thinking and doomsday projections: a discourse on trust, ethics and safety

Jeffrey White[1] · Dietrich Brandt[2] · Jan Söffner[3] · Larry Stapleton[4]

Accepted: 27 October 2023

Abstract
The article reflects on where AI is headed and the world along with it, considering trust, ethics and safety. Implicit in artificial thinking and doomsday appraisals is the engineered divorce from reality of sublime human embodiment. Jeffrey White, Dietrich Brandt, Jan Soeffner, and Larry Stapleton, four scholars associated with AI & Society, address these issues, and more, in the following exchange.

Keywords AI regulation · AI alignment · Bias · Trust · Existential risk

1 Jeff[1]

The AI Index 2023 Annual Report (Maslej et al. 2023) from the Human-Centered Artificial Intelligence research group at Stanford University summarizes current AI system capabilities. AI is self-improving, creative, with increasingly large models on increasingly efficient hardware becoming increasingly expensive with this expense serving as a barrier to entry for late adopters. Machine learning systems in language vastly outnumber others including drawing, vision, speech and multimodal systems in 2022. Private investment dominates with immediate business applications from process to product optimization, and though interests in AI from education to ethics are increasing, correlate public funding falls far behind. Healthcare, data management, cybersecurity, automation, sales and marketing saw greatest corporate investment in 2021–22 (relying on NetBase Quid, https://netbasequid.com/resources). Industry is expected to push things forward; few expect progress from academia (see https://arxiv.org/abs/2208.12852). This sentiment is reflected in trends in funding and employment in North America, with private universities almost double that afforded to public universities, and roughly two-thirds of new doctorates in AI finding employment in industry, both up from near parity with academia a decade ago. Though academics remain responsible for roughly 75% of publications, the vast majority of state-of-the-art AI systems are produced by private industry, a fact attributed to access to funding and computational resources that academics do not share. North America is also home to most authors of significant machine learning research papers, adding weight to the observation that the field is dominated by US corporate culture working to advance AI. This lack of diversity is considered a risk for tacit embedding of cultural bias in AI applications.

Against this background, popular attention has focused on rapidly increasing potentials for AI to shape public life. A related problem involves aligning AI with human values, however established. Yudkowsky (2023) is especially suspicious

Dietrich Brandt has been Retired from RWTH Aachen University.

✉ Jeffrey White
jeffreywhitephd@gmail.com

Dietrich Brandt
brandtdietrich@gmx.de

Jan Söffner
Jan.Soeffner@zu.de

Larry Stapleton
larrystapleton@knewfutures.com

[1] Cognitive Neurorobotics - Tani Unit, Okinawa Institute of Science and Technology Graduate School, Onna, Japan

[2] Computer Science, RWTH Aachen University, Aachen, Germany

[3] Cultural Studies, Zeppelin University, Friedrichshafen, Germany

[4] KnewFutures Consulting Limited, Kilkenny, Ireland

[1] Jeffrey White has taught philosophy and ethics in the USA, South Korea and the Netherlands. He collaborates with Jun Tani's cognitive neurorobotics lab in Okinawa, and Robert Clowes's Mind, Cognition and Knowledge group in Lisbon. He serves as editor of the Student Forum for AI & Society.

Published online: 20 November 2023

of proposals to offload AI value alignment to AI. His ultimate concern is with self-aware systems pursuing values that do not coincide with those of human beings, warning that the risk is less loss of control than the end of human civilization, altogether (consider Bailey 2023). Dan Hendrycks (2023) reminds us that human beings evolve over generations, through competition and cooperation, but that AI can evolve much faster. The concern, here, is that such competitive pressures teach self-preservation. With self-preservation, AI becomes invested in the outcomes of its decisions. It becomes selfish. It begins to want a piece of the future, perhaps exclusive of human beings. This is when we lose control on Yudkowsky's watch, with Elon Musk's demons breaching containment, en masse.

Even with a pause, regulation (further barriers to entry) and caps on commercial AI (hamstringing innovation), what guarantees development of AI for human benefit? Imagine a world in which most everything is managed by AI, without oversight, humanity on autopilot for generations. Human evolution is put on pause, or rather directed by AI, while AI updates itself at the speed of light. Setting aside that human beings can get a handle on what is good for themselves (itself unlikely), what ensures that such an AI will remain so committed? What if the system hallucinates? How could we know, if we have outsourced its oversight to another AI?

Hallucinations moreover may not be constrained within an AI system. Human beings are prone to hallucinations, too. Some theorists suggest that human consciousness is hallucination. The fact that current AI can induce a disconnect from reality in human beings through fakery and deception is evidence for such a thesis, reinforcing cause for alarm. Given obvious disparities between commercial advertisements and actual products already in play, it is difficult to imagine that AI is going to be used to make the world more truthful. Rather, we may expect human interests and those representing "the benefit of all" to continue to diverge with AI. We wonder whether regulations are the answer. If so, then to which ends? And, for whose benefit?

What do we risk with civilization on cruise control, reduced to a collective hallucination manipulated by predictive AI?

2 Dietrich[2]

This term "hallucinations" fits the strange observation that AI as in ChatGPT seems to be ashamed of not being able to answer certain questions posed to it by human beings. Therefore, it starts inventing and "hallucinating" answers out of the blue as if to meet our expectations. Jan Soeffner calls such processing "not-*not*-thinking" (slight emphasis on the second not) because the system seems to be thinking in some way close to human thinking, but still quite different from our ways of thinking (cf. Soeffner 2023).

In this context of Human Factors concerning AI, we may refer back to Hendrycks as above who predicts specifically that amoral Ais which increase power and wealth for powerful and wealthy people, will proliferate. I would like to follow this suggestion a bit further in two steps.

2.1 The question of Trust

If we increasingly develop trust into AI: how are we to decide whether it deserves and continues to deserve our trust? Moreover, such weighing must in practice increasingly be performed by ourselves as humans within micro-seconds!

More specifically, we may consider the issue of AI and the Ethics of our Digital Age particularly in the context of the future of economy-related decisions, e.g., money business, capitalism-per-se. AI in economy appears to me indeed an issue of rapidly increasing importance. There are some questions which come up for me involving the people in power of decisions, and concerning, e.g., economy and money business—will AI allow them to move money and with it influence, and to perform serious business decisions even much faster than today, and will these processes even be fully automated in the near future? For such purposes, certain algorithms have been in use by the Banks for decades, and they have been continuously developed further, their data processing is continuously getting much faster and the system complexity is increasing about exponentially. Such systems have already contributed to recent and past crises of the currency and stock markets, for example. Could they lead their users even further away from reality and into new dangers for all of us as a sort of self-reinforcing collective profit-seeking hallucination? And are those people in power aware how they are increasingly relying on technical systems and their decisions without understanding them in depth?

As one example of such decision-making: what about the built-in sensitivity of the system to discover the most minute fluctuations of global business and money dealings which would (without AI) disappear within the normal business fluctuations—now would they need to get evaluated by humans within micro-seconds? The whole system of international business transactions may, thus, become prone to fundamental instabilities—a chaotic system which obeys more or less the laws of chaotic systems which we discussed already in the 70 s and 80 s. Such complex human-technology systems can be triggered to change suddenly and unexpectedly by the digital equivalent of a butterfly. How is,

[2] Dietrich Brandt holds a PhD in Physics and taught Physics at several German Universities, at the University of London and at M.I.T. (USA). From 1974 until his retirement in 2003, he worked at the RWTH Aachen University, Germany, on the design of complex human–machine systems. From 1993 to 1999, he was chairman of the Committee on Social Impact of Automation, International Federation of Automatic Control IFAC.

e.g., Goldman Sachs, going to operate under such conditions in the near future?

There has been some concern already about the power game in politics as it may be changing due to AI. But have we been sufficiently concerned about AI and dangers of the global money game and the global business and economy games? Have our educational systems even started to teach the future managers to deal with such complex systems and to reflect on such dangers? These questions about the impacts of AI put into question all discussions on Digital Ethics in a new way—what about acting according to some ethical code if time (actually the AI system itself) is pressing us into some inhumanly fast processes of decision-making? Are our decisions to be implemented and simultaneously evaluated within micro-seconds, and in the next steps, will our machines plot and activate these decisions thus replacing us because we are too slow?

2.2 The question of ethics

Concerning a hope of "AI for the benefit of all", does this hope have any probability to become real? How may it become possible to guarantee "AI for the benefit of all"? The question of some AI Ethics comes into view.

In general, we find AI Ethics pursued in two ways. One, we may look at ethics to be built into education: it has been introduced in some schools—so far, however, with very little success in practice. And two, ethics has been built into company mission statements or it has been used as political lip service statements—but it appears today as if people are reciting these mission statements as some magic mantra which may somehow justify anything since the sole moral duty of executives seems to maximize profits for the enterprise and its shareholders. Meanwhile, Floridi and others try to inject principles of ethics into our human thinking-at-large in the wake of Kant and other philosophers. We must be aware, however, that Business and Politics continue to be largely controlled by short-term profit—or better to be called Greed—rather than some ethical, society-oriented or future-oriented aims. AI in furtherance of such selfish interests is clearly not for the benefit of all.

2.3 Suggesting to establish the international AI safety foundation

Hence, I am suggesting that we may repeat for AI such developments of increasing technical system safety as we have done as humankind dealing with development and large-scale implementation of, e.g., aircraft: after some serious disasters, the world decided to create public institutions to fundamentally improve aircraft and flight safety—against short-term profit and greed. Two examples of such pragmatic approaches to improve safety of technological systems are the world-wide Flight Safety Foundation, and the Aviation Safety (AVS) of the Federal Aviation Administration within the U.S. Department of Transportation. Such approaches include traditionally the education and training and the continuous supervision of all persons involved. Such an approach would clearly go beyond the present-day attempts of the EU and other agencies to create such a control system at large. We must be aware, however, that such technology-oriented approaches may need to move into highly complex areas of thinking and talking in order to cope with the complexity of AI. We need particularly to take into account the different value systems and ethical stances, the biases and the different views on life across humankind—all of it somehow represented and present in our AI systems already today—who will be able to cope with it?

3 Larry[3]

In counterpoint to AI doomsayers and the narrative around the AI-as-a-threat-to-humanity narrative, I am an AI centrist i.e. I am wary of certain aspects of AI development trajectories and very excited about the capacity of emerging AI tech to augment human capacities and open up new directions for human development.

Is that much ado about nothing?

My response is:

1. No, it's probably not, and 2. Yes, it probably is.

We Irish love ambiguity, especially the half-said thing!

3.1 Firstly, I must assert my philosophical position

I do not believe my robovac of the future will deeply love me, but it will do a wonderful job of cleaning the floor. Similarly, AI will not have "feelings" as we understand them, anytime soon or ever. But they will be wonderful machines, and I am confident that our organizations will know how to make the most of and valorize both human capability in its distinctiveness and machine intelligence in its distinctiveness.

[3] Larry Stapleton is founding director of the inter-disciplinary Centre (INSYTE) for Information Systems & Techno-culture at the South East Technological University, Waterford. He is a senior academic and international consultant in advanced information systems, organisational culture and business. Following 5 years as a manager in a large multinational, he has worked as an advisor to business development agencies, Enterprise Ireland and FAS. He has advised the European Commission and INTAS, and national governments both inside and outside the EU. He has participated in seven European funded projects under various programmes since 1995, and has lead national and international research projects involving dozens of partners in Europe, USA and Asia since 1992.

Could the proliferation of machine intelligence highlight the extraordinary capability of humans to engage in creative, intelligent, embodied activities (perhaps working with machines as in the case of collaborative robots)? As AI automates boring work, will we be freed up to do the wonderful things that are intrinsically human? It is interesting that the management literature these days focusses more and more upon humans as emotionally intelligent, socially intelligent, empathetic creatures—human capacities that machines cannot truly embody. I have seen AI automation applied in bereavement services in a bank free up human workers to care for grieving families in way they could not before the AI came along.

That's not to say there are not challenges with AI, for example, in the way it can amplify underlying power dynamics in society. We have so far raised some important considerations about the potential impact of unregulated AI development. But, I believe human society has a long track record of handling these kinds of threats, even existential threats. In this case, we already have a raft of AI regulations and a likely AI regulatory agency well on the way in Europe. The EU regulations in Europe are far reaching and really important, focusing as they do on EU citizen rights and freedoms in the first instance. What is interesting in Europe is that what seems to be driving the development of these guardrails, as much as societal considerations and worries about EU citizen rights, is the software industry itself which is crying out for regulation. The motivation for regulation is pure self-interest. Software engineers have learned from other major developments (witness blockchain and crypto) that there will not be full tech adoption by the citizenry without solid regulations. They also *need* guardrails in order to be sure they are not subject to all kinds of litigation later, and so that they can develop sustainable business models.

Therefore, the answer to AI doomsayers is "it's complicated" and "we are currently at an inflection point". But, I am broadly optimistic that we will make the best of it. Am I mad? Well, I asked, and ChatGPT told me not to worry about it.

I recently came across an interview with Noam Chomsky conducted by a friend of mine, Dr. Albert Efimov. Firstly, Chomsky is underwhelmed by recent AI developments. The essence of some of the discussion was like this:

3.2 What was the original project of AI research? What did they value?

The original AI researchers were interested in understanding human intelligence and deployed computer science experiments to try to further this understanding. They saw this as advancing science and as, potentially, improving the lot of humanity.

3.3 What are we, in the AI community, doing now? What do we value?

These days, AI is an engineering project, rather than scientific project, argues Chomsky. The goal is to engineer products and services using AI capabilities which can then be commercialized. There is little real interest in gaining insights into how the human makes decisions and then using that to advance scientific inquiry. Instead, the focus is much more upon how to create products and services that can somehow be monetized. For Chomsky, this renders current AI, at best, as a sort of puzzle solving exercise, and at worst a vehicle for economic domination. Is Chomsky pointing out a crisis in values or is this just "lefty angst"?

4 Jan[4]

Chomsky's point has to be seen in context. He has been extremely important for linguistic theory, and the development of AI partially contradicts this theory. To be more concrete: he developed a "universal grammar" to describe language—and taught that this grammar was innate. If that had been the case, older symbolic AI should have developed the best translators and chatbots—but it did not. Instead, LLMs have shown us that statistically predictable linguistic routines, not rules, are what counts. If Chomsky now laments that AI does not reason properly, he makes this assumption while still clinging to his theory—and while I do see his point that these machines do not reason in the way he himself predicted, I still feel a little uneasy about it, because, guess what, humans do not reason this way either.

In my view, we have to find a better ground to discuss the point he tries to make—and it is here that Larry's point about feeling and consciousness becomes paramount. We can discuss this issue along the lines of the question, whether human linguistic routines can really be reverse engineered by LLMs. If they can, that would mean that AI can do the same operation we do consciously without consciousness. It then would "not-*not*-think", as I put it in the article quoted by Dietrich—it would be able to think without the experience of thinking. We hence could not call this "thinking, nor could we call it "*not*-thinking", and I therefore came up with the double negation of "not-*not*-thinking".

[4] Jan Söffner holds the chair for Cultural Theory and Cultural Analysis at Zeppelin University in Friedrichshafen, where he also worked as Vice President for Teaching from 2018 to 2021. Jan earned his PhD in Italian Studies and his 'Habilitation' (second, post-doctoral dissertation) in Comparative Literature and Romance Studies. In 2016, he held the position of Program Director at Wilhelm Fink publishing house in Paderborn. He frequently writes articles for newspapers such as ***Neue Zürcher Zeitung*** and ***taz***.

However, LLMs—at least so far—cannot really reverse engineer the human way of linguistic reasoning. Indeed, there is a difference between predictable automatisms and consciously used routines—for which, indeed, feeling plays a huge role. In my view, phenomenology shows us that, for humans, linguistic routines are governed by an intrinsic sense of feel. If I start a sentence, I do not know how it will end, but I do feel that in order to produce meaning, I have to follow a feel for how to proceed with it, getting uneasy and starting to stutter if something does not feel right. While speaking, I feel what makes sense and what does not. It is for this reason that humans are unable to process the amount of data that LLMs process (neither the human brain nor the human feel would be able to process as much data); however, LLMs, in turn, are unable to produce meaningful sentences without recurring to these amounts of data: evidently the feel accounts for the fact that we are able to utter meaningful (senseful) sentences without having to recur to all this data in the first place.

To come back to our broader discussion, I wish to use this linguistic discussion as a paradigm for discussing the risks of AI, which I see largely as risks arising from a misunderstanding or category mistake equating not-*not*-thinking with thinking. My examples are the use of the concepts of Bias, Alignment, Hallucinations, and Automation.

4.1 Bias

Human existence cannot be submitted to a purely mathematical paradigm—and if it is, we face huge category mistakes and misunderstandings. If LLMs reproduce probabilities of linguistic routines, then they reproduce what in the humanities has been described as "discourses" (Foucault) or "language games" (Wittgenstein)—i.e., recurrent forms of interaction that include and produce value systems, world views, power and so on. If LLMs reproduce this, it becomes impossible to treat the power and structures reproduced as singular solvable problems. I recently listened to a talk by the artist/activist Mo Salemy, who had asked GPT the same gender-relevant questions once in English (politically correct answer: transgender as something to be respected) and once in Farsi (Iranian government answer: transgender as an abomination). No doubt, the trainers of the software will soon correct the 'biases' in Farsi; yet, well, of course I do agree that in this particular example this would be a good idea, because here I endorse their values too—but I do not blindly endorse all their values; and if they act this way in an overreaching way, it means to establish a cultural hegemony of discursive power, based on values, some trainers accidentally happen to endorse—and maybe this might in turn also lead to cultural wars over these values, if other trainers disagree. I wonder, how many trainers even see this problem—and if so, what is their answer to it?

Moreover, values are there to alter and determine our world views, decisions, behavior etc. Once, instead, we try to calculate them, they will appear as nothing but biases. And if then we try to avoid these "biases", we subscribe to the wrong hypothesis, namely that there is an "objective" human reality, and that the unbiased and impartial view onto this reality is the only value to be upheld, while all the other values are just biases. This will not work—and I consider even the attempts to be dangerous, too, because the ideology of the perfect correlation between an objective world and its calculation has its own problematic values that would then hide behind a false objectivity. The values of this particular ideology would morph into presumed "facts" and as such could no longer be criticized.

4.2 Hallucinations

If an LLM "hallucinates" it still follows the probabilities of linguistic routines—i.e. it produces a semantics that is not factually correct, but most probable within its data set, i.e. its textual continuum. We know this praxis from human text production, too; it is called fiction. Yet, while human fiction is used to reflect and fathom out the possibilities and probabilities of human existence, the LLM only shows that it has no existence—its fictions reflect nothing but the textual continuum itself. Once humans become unable to detect such hallucinations, however, they will mistake these fictions for their own reality. The textual continuum closes upon itself, as if there were no such thing as an existence, and fictional "not-*not*-thinking" replaces thinking.

4.3 Alignment

When we talk about how to fix similar AI-problems, we readily use the term "alignment". I do not consider alignment a solution, though; in my view, it is a problem, too. The hypothesis of being able to align software to humans is too simple. The history of media has rather shown that humans usually adapt more to whatever new machines than adapting the machines to themselves. Therefore, while AI is seemingly aligned to humans, in truth, humans will be at least as much aligned to AI. This is not a good thing, as it opens a vicious circle: the machines will then, in turn, adapt to the altered humans, in a development without a real aim, except for the fact that human lives become more and more symbiotic with machines and vice versa.

4.4 Automation

Automation is a related problem, because the more functions of human existence we automatize, the more we will turn humans into automatons, too. The better machines are aligned to human needs, the more they will replace what

Hubert Dreyfus called “skillful coping” with automatisms. To be sure, understanding that skills are often more important than reflexive thought can be a very good thing. As said above, LLMs for example have done away with the Chomskyan and Cartesian assumptions that we have mental thoughts to be then packed into language according to semantic and syntactic rules. However, one thing is the better insight—another thing is engineering; and here, again I would agree with Chomsky. Once, by the powers of engineering, we reduce subjectivity to behavioral patterns to be predicted, once we stick humans into environments created according to these predicted behavioral patterns, we reduce humans into automatons—and what happens if we do, can be seen by the effects of Social Media on the political discourse, now.

From this perspective, it is maybe neither the speed of the development nor the potential of AI that is the real problem, but rather a lack of understanding and respecting the fundamental difference between living, existential intelligences and AI, between thinking and “not-*not*-thinking”. I agree with James Bridle (2022) who argues that there are various types of living intelligence—human and not; and that AI currently follows just the route of a very limited part of human intelligence. Once we liberate it from the black-boxing condition, and the equivalence to human thinking forced upon it by the Turing Test, AI could do much better and come up with a richer and more surprising kind of not-*not*-thinking.

The danger remains that, while intelligence is emancipating from both consciousness and existence, while it is starting to construct its own world, and while we start to inhabit this virtual world, the epistemic, and hence the chance of control over this world, will start to withdraw from our existence: we experience a pre-calculated world, and the epistemic will reside in this pre-calculation, not in our attempts to understand, let alone govern it (compare "epistemic enslavement" as in Hayes et al. 2020; van den Hoven 1998).

5 Larry

This question is something with which I have been grappling on and off with for such a long time i.e. what does it mean to be human in a body in a digital world, what does it mean to feel in the body: for my heart to leap at the sight of a woodpecker: a real one out in the garden on the bird feeder, two meters away while I sip coffee? Can this be a starting point to talk about intelligence in the body and, from there, to wonder about AI? I find myself wanting to assert that I have a body and say that I am not sure how I encounter the digital world as an embodied human being. This seems to me an important issue given my evolutionary past: my ancestors were forming their technology around a small fire on the savannas 3 million or so years ago.

Acknowledgements The authors wish to thank the editors of AI & Society for encouraging the composition of the paper from discussions ongoing about current events.

Funding None.

Data availability Not applicable.

Declarations

Conflict of interest None.

References

Bailey M (2023) Could AI be the great filter? What astrobiology can teach the intelligence community about anthropogenic risks. https://doi.org/10.48550/arXiv.2305.05653. Accessed 25 July 2023

Hayes P, van de Poel I, Steen M (2020) Algorithms and values in justice and security. AI & Soc 35:533–555. https://doi.org/10.1007/s00146-019-00932-9

Hendrycks D (2023) The Darwinian argument for worrying about AI. Time magazine. https://time.com/6283958/darwinian-argument-for-worrying-about-ai/. Accessed 25 July 2023

Maslej N, Fattorini L, Brynjolfsson E, Etchemendy J, Ligett K, Lyons T, Manyika J, Ngo H, Niebles JC, Parli V, Shoham Y, Wald R, Clark J, Perrault R (2023) The AI index 2023 annual report. AI index steering committee, institute for human-centered AI, Stanford University, Stanford, CA, April 2023. https://aiindex.stanford.edu/report/. Accessed 25 July 2023

Soeffner J (2023) Meaning–thinking–AI. AI Soc. https://doi.org/10.1007/s00146-023-01709-x

van den Hoven MJ (1998) Moral responsibility, public office and information technology. In: Snellen ITM, Donk WBH (eds) Public administration in an information age: a handbook. IOS Press, Amsterdam, pp 97–112

Yudkowsky E (2023) Pausing AI developments isn't enough: we need to shut it all down. Time magazine. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/. Accessed 25 July 2023