Philosophy & Technology (2024) 37:55
https://doi.org/10.1007/s13347-024-00743-x

RESEARCH ARTICLE

Expropriated Minds: On Some Practical Problems of Generative AI, Beyond Our Cognitive Illusions

Fabio Paglieri[1]

Received: 5 January 2024 / Accepted: 3 April 2024 / Published online: 20 April 2024

Abstract
This paper discusses some societal implications of the most recent and publicly discussed application of advanced machine learning techniques: generative AI models, such as ChatGPT (text generation) and DALL-E (text-to-image generation). The aim is to shift attention away from conceptual disputes, e.g. regarding their level of intelligence and similarities/differences with human performance, to focus instead on practical problems, pertaining the impact that these technologies might have (and already have) on human societies. After a preliminary clarification of how generative AI works (Sect. 1), the paper discusses what kind of transparency ought to be required for such technologies and for the business model behind their commercial exploitation (Sect. 2), what is the role of user-generated data in determining their performance and how it should inform the redistribution of the resulting benefits (Sect. 3), the best way of integrating generative AI systems in the creative job market and how to properly negotiate their role in it (Sect. 4), and what kind of "cognitive extension" offered by these technologies we ought to embrace, and what type we should instead resist and monitor (Sect. 5). The last part of the paper summarizes the main conclusions of this analysis, also marking its distance from other, more apocalyptic approaches to the dangers of AI for human society.

Keywords Generative AI · Machine learning · Societal implications · Enhancement · Replacement · Extended mind

1 Introduction: ELIZA, is that you?

Advancements in machine learning (ML) techniques, together with access to large amounts of machine-readable datasets, have led to the success of generative AI systems, both for text generation (e.g., ChatGPT, Bard, and Bing Chat, later rebranded

✉ Fabio Paglieri
fabio.paglieri@istc.cnr.it

[1] Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche (ISTC-CNR), Via Giandomenico Romagnosi 18A, 00196 Rome, Italy

as Copilot) and text-to-image art systems (e.g., DALL-E, Midjourney, and Stable Diffusion). In the recent media hype surrounding these applications, it has become commonplace to share anecdotes on their use and its results: since early 2023, social media feeds have been inundated by stories regarding users' personal experiences with AI generative systems like ChatGPT, the one that attracted the most prominence in public discourse. In the vein of this sociable tradition, it seems fitting to start this paper recounting an interesting "close encounter" I had with these technologies.

Recently, I came across the work of a graduate student, whose identity is charitable to keep hidden. The text was written rather well and included a critical review of the experimental studies conducted on one of my research topics: therefore, my name and the name of the colleague who worked with me on that line of research, Marco Marini, often appeared in the text, and consequently also in the final bibliography. The consultation of the references, however, had in store a few surprises. Among other entries, the following were recorded:

> Marini, M. (2013). When it's better to choose the one you love: The effect of attractiveness biases in consumer choices. *Judgment and Decision Making, 8(5)*, 476-485.
> Marini, M. (2019). How to get people to take risks? A choice-based measure of risk preference. *PloS One, 14(1)*, e0209983. doi: https://doi.org/https://doi.org/10.1371/journal.pone.0209983
> Marini, M. (2019). Luring to a suboptimal option: The effect of payoff reduction in a risky choice framing. *Judgment and Decision making, 14(2)*, 198-207.
> Marini, M. (2020). The asymmetrically dominated compromise effect in a dynamic setting. *Journal of Economic Psychology, 76*, 102-257.
> Paglieri, F. (2009). The attractiveness of decoys in economic contexts: An experimental investigation. *Judgment and Decision Making, 4(4)*, 335-342.

Formally, this bibliography extract is flawless: the entries are correctly formatted according to the standards of the American Psychological Association (APA), the relevant information is all present, the articles are consistent with the topic of the student's assignment, and the titles of the various contributions are, objectively, quite intriguing. The only problem is that… none of these publications exist!

The incident was neither a brave, subversive act of provocation (to demonstrate that university instructors no longer read carefully the written assignments of their students), nor a symptom of terminal stupidity in the student (only a very dumb cheater would try to falsify the references of the very same people tasked with evaluating their work): instead, it was the outcome of a naïve and inappropriate use of generative AI. The student, after writing the assignment themselves and inserting the appropriate references in the text, using the author-date APA standard, had incautiously asked ChatGPT to prepare the reference list, giving it their own text as part of the prompt.[1] Unfortunately, the software compiled a bibliographic list in full

[1] Incidentally, the request, in and by itself, is not at all fraudulent: formatting a bibliography according to a certain standard, once the relevant titles have already been identified and correctly cited in the text of the paper, is a mechanical and tedious task, so wanting to delegate it to a machine is perfectly reasonable.

compliance with APA standards, but without any attention to the truthfulness of the information included therein.

Here, however, we are not interested in the student's misadventures, but rather in how ChatGPT produced its output, which was certainly not random: there is method to this madness. Firstly, the journals in which the fake contributions would have appeared are plausible, both thematically, and because Marini and I have already published in those venues in the past, or in very similar ones. Secondly, the volume numbers that are mentioned refer to issues that have indeed been released, and usually the numbering and year of publication match; in one case, the entire reference (PloS One, 14(1), e0209983. doi: https://doi.org/10.1371/journal.pone.0209983) refers to an existing article, except that it is a study on a completely different topic, i.e. gender barriers in research at the South Pole (Nash, M., Nielsen, H., Shaw, J., King, M., Lea, M.-A., & Bax, N (2019), "Antarctica just has this hero factor…": Gendered barriers to Australian Antarctic research and remote fieldwork). The inconsistencies that emerge upon closer inspection are also revealing: the 2020 article attributed to Marini is listed as appearing between page 102 and page 257, except that there never was a single 155-page long contribution published in that particular journal, and probably not even in others, at least in the field of economic psychology; delving deeper, one discovers that the Journal of Economic Psychology, from 2020 onwards, no longer reports the page numbers of individual articles, but only their identification number, which is composed of a 6-digit code starting with 102, and the code 102257 (that ChatGPT creatively transformed into page numbers, 102–257) corresponds to the editorial of the issue following the one cited in the invented bibliographic reference. At other times, the system falls prey to ambiguities of meaning: the decoy effect, which was the main focus of the student's paper, is also referred to as the attraction effect in the literature, and the word "attraction" evokes the semantic field of affects, which instead has nothing to do with the technical phenomenon in question (i.e., a shift of preferences towards an option that is manifestly superior to another inserted ad hoc, called decoy). It is because of this semantic ambiguity that ChatGPT came up with a title like "When it's better to choose the one you love: The effect of attractiveness biases in consumer choices" – a wonderful title, by the way, which I will certainly use, as soon as the opportunity presents itself.

In short, this false output is not due to anomalies or errors in the functioning of the software, but on the contrary it illustrates perfectly what ChatGPT is built to do (and does very well): generate linguistic strings (in this case, bibliographic entries) that have the maximum probability of satisfying the user's request, based on similar instances present in the (huge) database to which the program had access during training. What ChatGPT does not do, and cannot do due to the way it functions (at least for the time being), is consulting the real world or an internal representation of it: the system does not work by checking the state of the world and describing it, but

Footnote 1 (continued)
In fact, unbeknownst to the poor student, there are excellent open access software for that (Zotero, for example), which will not insert invented references in the process.

rather by constructing responses that are maximally consistent with the vast mass of linguistic data at its disposal, whose adherence to reality is by no means guaranteed. Consequently, the response that the student receives after asking for an APA-style bibliography is absolutely adequate by the lights of the program, that is, it "sounds good" and is consistent with the APA bibliographic canons; but it is not at all adequate in a referential sense, given that it makes up non-existent titles, although plausible and captivating.

Failure to appreciate the basic working of generative AI systems is what fuels a tendency to use them as oracles: like my student, we ask something and expect the system to "answer", implicitly assuming that such an answer will be provided in the same way in which a human agent would do it. In other words, we assume that the system understands what we are asking (in the sense of attributing referential meaning to our request, and possibly also drawing appropriate pragmatic implicatures), and then it answers based on some relevant knowledge of the world. The reality, as just discussed, is very different: generative AI systems basically detect patterns and "answer" by returning the most likely meaningful linguistic output, given the prompt they receive and the dataset used in training. This is also why such systems have been known to hallucinate, i.e. to provide answers that are formally impeccable yet utterly invented (like the bibliography in our example): by design, they are not constrained by semantic truth, only by statistical prediction on the linguistic input.

There are correctives, of course: for instance, search-based LLMs, such as Bing Chat (now Copilot), try to "ground" their performance on the results of web searches on the most relevant issues included in the prompt they receive, to ensure that their answer is neither fabricated nor outdated. However, these solutions work by curating the dataset to which the system is exposed: this does not change the fundamental nature of its operations, which is about detecting patterns among large linguistic datasets, not establishing something about reality. LLMs interact with us by chaining linguistic input in the appropriate ways, not by conveying any further meaning.

Nonetheless, the widespread temptation to treat these systems as meaningful interlocutors in a conversation is strengthened by two main independent factors: a specific design choice made by their developers, and our own deep-seated tendency to humanize any machine capable of interacting competently with us, even at a fairly superficial level. Concerning the former, it is worth spelling out the acronym GPT in "ChatGPT", which stands for Generative Pretrained Transformer. "Generative" indicates the key operation performed by the system, i.e. generating new textual strings by predicting the most adequate next word in the output and chaining such predictions together in coherent texts. "Transformer" refers to the specific neural network architecture implemented since 2017, which (to put it simply) allows the system to focus its computing power only on the most relevant parts of the input, with a huge boost in performance. The word in the middle, "Pretrained", reminds us that these systems, before being given us to play with, have undergone massive training. This training happens in distinct phases, which can be summarized in three main steps: (i) actual pretraining on huge amounts of unstructured data taken from all sorts of sources (e.g., online content, newspapers, books), which makes the system good at generating text on a variety of topics; (ii) instruction fine tuning, in which the system performance is aligned with its intended use by training it on datasets that

are closer to the target use cases, including a lot of questions-answers sequences to teach the model to respond to instructions; finally, in some cases, (iii) reinforcement learning with human feedback, which is also instrumental to alignment and allows the system to better interact with actual users, with the specific aim of making its performance well-suited to their expectations. This procedure shows that the last two stages of training of LLMs are specifically designed to make them successful as credible interlocutors for their users: hence, our tendency to regard them as such is, at the very least, strongly reinforced by the design choices made by their developers.

However, humans have been known to treat even simple chatbots like meaningful linguistic agents, well before such systems were capable of the excellent performance exhibited nowadays by LLMs. This is the well-known ELIZA effect: ELIZA was a textual interaction program (today we would call it "chatbot") created in 1966 by Joseph Weizenbaum; its name was inspired by Eliza Doolittle, the flower girl who speaks poorly in George Bernard Shaw's *Pygmalion*, and the program clumsily mimicked the initial stages of a Rogerian therapy, essentially responding to the user's statements or questions with other, seemingly relevant questions. This was enough at the time to convince many that the program understood what the user was saying and responded in kind, or even in a profound, provocative, and wise way. This type of improper attribution of intelligence to digital interfaces has since become known as the "ELIZA effect", and it is a good example of the limitations of Turing's imitation game (1950) as a litmus test of intelligence. As an illustration, consider the following exchange between ELIZA and a user:

User: Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here?
User: He says I'm depressed much of the time.
ELIZA: I am sorry to hear you are depressed.
User: It's true. I'm unhappy.
ELIZA: Do you think coming here will help you not to be unhappy?

Our proneness to treat interacting machines as analogous to human interlocutors is becoming even more relevant nowadays, when we are faced with generative AI systems that are much better than ELIZA at playing the imitation game, and it raises two problems. The less relevant one is that such perception is substantially inaccurate, as discussed, and it risks making us blind to the actual mechanisms that empower these applications. The most pressing concern, however, is that excessive anthropomorphism towards these technologies is also deeply misleading. Let us call it *the red herring of machine intelligence*: as long as we frame the debate on generative AI around the idea that "machines have learned to talk" (or to draw, or to make a video, or whatever creative application may be relevant), our attention will remain focused on establishing the extent of their intelligence. Some will question it, others will celebrate it; some will fear it, evoking apocalyptic scenarios of machine dominance a la Skynet, others will welcome it as a gift, a further empowerment offered us by technological progress. Either way, we will keep talking about the alleged intelligence of these machines, whereas another, much more urgent problem should occupy our collective attention: the most striking and potentially problematic implication of these technologies, indeed, is not their impact on the theoretical definition

of intelligence, but rather their consequences for *practical ownership of the means of production of intelligent behavior*. The crucial question is not "Is their behavior intelligent?", but rather "Are we still in control of what makes their behavior intelligent?" – regardless of what definition of intelligence we use in assessing their conduct.

2 Interlude: The Dark Side of the Black Box

Whenever issues of control are raised in debate over AI systems, there is a tendency to frame them in terms of transparency: we consider ourselves as in control of AI technologies insofar as we can understand what is going on in these systems and explain it to others. Indeed, explainable AI, also known as XAI, has become a key topic in AI research over the last few decades (for some reasonably updated surveys, see Adadi & Berrada, 2018; Arrieta et al., 2020; Langer et al., 2021), especially since some of the crucial steppingstones upon which the success of current AI systems is built (e.g., deep learning) are particularly hard to explain. Indeed, the debate nowadays is mostly focused on the limits of XAI, for instance analyzing the trade-off between accuracy and explainability: as Adadi and Berrada summarize, "the most accurate AI/ML models usually are not very explainable (for example, deep neural nets, boosted trees, random forests, and support vector machines), and the most interpretable models usually are less accurate (for example, linear or logistic regression)" (2018, p. 52145). What makes this problem particularly interesting, and potentially scary, is that here we are not talking just about explanations accessible to laypeople, but also explanations accessible to anyone, including the original developers of the AI system in question. According to most commentators, we have already reached the point where nobody exactly knows how some AI systems work, not even the people that built them: such a scenario comes with a certain existential anxiety, and understandably so.

Some qualifications are in order, though: the general principles according to which any AI system works are well understood by its designers, of course, and this remains true also for recent applications, otherwise they could not have been developed to begin with. However, what is often beyond the grasp of the designers is how exactly the system learns, after processing huge amounts of information, to solve the tasks that are assigned to it. Both the size of the datasets involved and the sheer complexity of the system itself prevent anyone from clearly understanding what is going on. Just to give an idea of the magnitudes involved, Large Language Models (LMMs for short) are called "large" because they are based on so called deep neural networks, involving multiple hidden (deep) layers with billions of nodes: e.g., Chat GPT-3, back in 2020, had about 175 billions calculation nodes or parameters, more than twice the estimated number of neurons in the human brain (86 billions, according to Azevedo et al., 2009). So, the most accurate description of our epistemic status with respect to the black box of generative AI is that experts understand very well the general principles of its functioning (and may occasionally be able to explain it to laypeople), yet nobody has a clear grasp of how they solve specific tasks, once these systems start working in contact with data. This is in sharp contrast

with what happens with other technologies: end users remain often blissfully ignorant of the inner working of most technological devices (computers, phones, cars, radios, televisions, etc.), yet for each of these technologies there are experts that understand their procedures in full details. This is simply not the case with generative AI, and ML in general.

This situation poses obvious urgent challenges: for instance, should the radical lack of explainability of generative AI and ML be a deal breaker for their adoption, in spite of whatever benefits they might offer to human society? However, as it will become apparent in what follows, this is not the kind of problems that this paper focuses on. This is not meant to deny the relevance of similar questions, rather to bring attention to other problematic implications of generative AI, which usually are overshadowed by the general concern for its lack of transparency. In that regard, it is worth noticing that the explanations sought after in XAI, albeit tailored to specific stakeholders, always concern the inner workings of AI systems, "how the magic happens", which is exactly what makes this quest so elusive for generative AI and ML. It is somewhat surprising, however, that little or no attention is given to other types of explanations, focused not on how these systems work, but rather on *who stands to gain (or lose) from the fact that they do work*. "Cui prodest?" is, literally, one of the oldest questions in the book, when it comes to making sense of relevant phenomena: yet this aspect is rarely considered a relevant part in the explanation of AI systems.

I suggest this oversight should be remedied: not only because "following the money" is often extremely informative for making well-informed collective decisions on matters of public interest (and AI certainly qualifies as one such), but also because, at this level of explanation, generative AI and ML are not harder to analyze than other technologies, thus we should take advantage of that. As discussed in greater details in the next section, it is indeed relatively easy to appreciate where the value of generative AI comes from, as well as where its benefits tend to accrue; moreover, both factors reveal significant imbalances in the current scenario and suggest interesting correctives, which are largely independent from the technical opaqueness of generative AI and ML. To put it bluntly, not knowing how the machine works should not prevent us from understanding who is making money with it, and how.

3 Data Colonialism in the AI Far West

The economic value of generative AI comes from the quality of its performance, which in turn derives from two key assets: significant technological advancements in ML techniques, and access to large, machine-readable datasets. The first asset is provided by scientific research, the industry at large, and ultimately the developers of each specific AI application: thus, it stands to reason that these stakeholders should ripe most of the benefits. Data, however, are provided by all of us, quite literally: hence, by the same lights, part of those benefits should be fairly distributed among the original data providers, since all of them contributed (possibly without even realizing it) to the amazing performance of the latest AI systems. There is nothing vague

or imprecise in such a statement: data are essential to ML, which in turn stands at the core of any generative AI, very much like gasoline is crucial for a car engine – without it, the engine will not work, no matter how well designed it is.[2] The same with data for ML. Gas stations require payment to provide gasoline for our cars: similarly, we are entitled to payment for providing the data that make generative AI systems tick.

A very public manifestation of this principle occurred in late December 2023, when The New York Times sued OpenAI and Microsoft for using massive amounts of their copyrighted articles to train their generative AI systems, without permission and without compensation, and with the added damage of creating platforms that act as direct competitors to traditional newspapers, such as The Times. Similar lawsuits against tech companies had already been filed by prominent individuals (e.g., writers like David Baldacci, Jonathan Franzen, John Grisham, Scott Turow, as well as comedian Sarah Silverman) for similar reasons, alleging that significant portions of their creative works had been co-opted for training purposes, again without permission or compensation: The Times was, however, the first large media corporation to sue AI companies, possibly paving the way to others. Legally, in the US the question hinges on the definition of "fair use": whereas the tech companies argue that including copyrighted materials in training datasets for AI systems is an instance of fair use and therefore permissible without approval from or compensation to copyright holders, newspapers and content creators claim otherwise, usually citing instances in which the AI systems, once trained on such materials, reproduce it almost verbatim in its responses to users' prompts. As The Times put it, "there is nothing 'transformative' about using The Times's content without payment to create products that substitute for The Times and steal audiences away from it".[3] Interestingly, the newspaper is seeking not only monetary compensation, but also (and possibly primarily) a change in basic practices in the generative AI industry: in fact, The Times' sue asks for the destruction of chatbot models and training sets that incorporate its material.

Regardless the outcome of this legal dispute, similar events highlight the controversy surrounding the use and value of data for new technologies (for discussion, see Bertini, 2023). This is not limited to AI systems, of course: data, as well as other user-generated contents, are an integral part in the success of several digital technologies. The most notable example are social media platforms, where profits hinge on maximizing users' engagement with the platforms, which is largely motivated by interest in contents (posts, photos, videos, comments, likes, reactions, etc.) that were generated by other users, which in turn tends to grow as a function of the time spent on the platforms by users, in a potentially endless cycle of social exchanges. Such social exchanges, however, net huge profits for the companies that provide the platform, regardless of how gratifying they are for

[2] If anything, the metaphor is too conservative, in that generative AI uses data in a much more intense and sophisticated way than what an engine does with gasoline: data can be squeezed several times, and if integrated with other similar (same data from other users) or complementary data (other data from the same user) can offer further information.

[3] Source: https://www.reuters.com/legal/transactional/ny-times-sues-openai-microsoft-infringing-copyrighted-work-2023-12-27/ (last consulted on December 29, 2023).

users: ironically, there is a widespread tendency to consider use of such systems as "free", with an accompanying sense of having made a great deal in exploiting them without spending any money, whereas in fact we are paying access to platforms with our data and user-generated contents, whose value for the social media providers far exceeds the costs they sustain to offer us that particular service. In other words, we perceive as a gratuity what is instead, for us, a horrible economic transaction. Similar considerations apply to recommender systems: they significantly improve our user experience on a variety of platforms (e-commerce websites, streaming services, social media, and basically every digital environment where we are threatened by information and choice overload), yet they work only thanks to behavioral data on our own choices, which we provide for free, with significant economic benefits for the platforms themselves.

It is becoming increasingly apparent that traditional copyright laws are not ideally equipped to deal with these conundrums, since the crux of the matter is not intellectual property, but rather fair distribution of the resulting benefits (a problem which applies also to other technologies, e.g. digital platforms: see Montesi et al., 2016). Even the notion of "transformative" uses, which in US jurisprudence is supposed to help determining legitimate use of unlicensed copyrighted materials, implies adding "something new, with a further purpose or character" to the original content: a criterion grounded on the idea that such transformations would entail significant creative work by whomever is using the unlicensed copyrighted material, therefore justifying the claim that the new content is original and different enough to be considered as free from copyright. This intuition is at odds with how data and other user-generated contents, including our attention (Davenport & Beck, 2001; Simon, 1971), produce value in new technologies: either this happens by merely making such materials accessible to others, without any transformation (as it is the case for user-generated content on social media), or by letting AI systems use these materials to train their algorithm and achieve excellent performances – in which case all transformations are performed by machines, not by human agents. In this scenario, framing the problem of data value in terms of copyright laws is likely to be a lost cause: instead, we should look at this situation as an instance of *unfair exploitation of unpaid resources and/or services*.

This is the perspective endorsed by scholars working on so called "data colonialism", such as Couldry and Mejias (2019). The analogy behind the expression is meant to be very precise: data capitalism (West, 2019; Zuboff, 2018) and attention economy (Davenport & Beck, 2001) express colonial traits in a historically accurate sense, insofar as they invade previously uncontested territories to extract resources with little or no regulation. In traditional colonialism, the invaded territories were newly accessible lands inhabited by technologically less advanced populations, the resources to be extracted were material ones (natural resources and slave labor, mostly), and those profiting from such exploitation were the colonial empires. In data colonialism, to be invaded are personal spaces, previously uncontested because deemed unable to produce economic value, the precious resources are attention and free generation of contents and data (including, most prominently, behavioral data, i.e. our action patterns), and those making money hand over fist are tech companies, taking advantage of our relatively poor understanding of the economic implications

of the new digital ecology (Floridi, 2014) and the resulting lack of regulations on its exploitation.

Rejecting this colonialist tendency requires fighting on two different, yet deeply connected fronts: on the one hand, there is need to *renegotiate the distribution of benefits and costs* of the ongoing digital revolution; on the other hand, *limiting or preventing data exploitation in the first place* must become a matter of collective deliberation, instead of being left to the will of private companies, operating in a poorly regulated market.

The first response leverages the intuition that, whenever users exchange data and contents for "free" access to online platforms and apps, they are in fact vastly overpaying relatively cheap services. To balance the account in a fairer manner, two main options are being pursued: data dividends and digital service taxes. The former is an approach mostly pursued in the United States, whereas the latter is being spearheaded by EU countries: this might not be mere coincidence, since data dividends treat the redistribution of digital profits primarily as a private negotiation between users as data sellers and service providers as data consumers (let us call it "the American way"), whereas digital service taxes delegate to states the task of making tech companies pay for data and then redistributing fairly these revenues ("the European way"). Data dividends gained prominence when they were endorsed in 2019 by the Governor of California, Gavin Newsom (Democratic Party), yet as of early 2024 they are still to be implemented anywhere, despite significant interest, both from scholars (e.g., the Data Dividends Initiative, https://www.datadividends.org/) and activists (e.g., the Data Dividend Project, https://www.datadividendproject.com/). The core idea is simple: each user should be able to quantify the value of their data, in order to receive fair compensation for it. Two major obstacles stand in the way of such straightforward approach, though: (i) how to quantify exactly the value of individual data, and (ii) how to track it back to each user, with the level of precision required for fair compensation. Currently, no effective solution to these hurdles have been found or implemented, at least at a scale large enough to matter.

Digital service taxes (DSTs) avoid both problems, by taking a broader approach to the issue: since the data and user-generated contents feeding the digital economy are collected globally, tech companies should pay taxes to each country, in an amount roughly proportional to the amount of "raw resources" provided by citizens of that country. This collective contribution is much easier to pinpoint, either by looking at the number of users of a particular technology within a country, or by estimating it on general principles: e.g., assuming similar levels of usage of global technologies worldwide, DSTs may be proportional to the population of each country, possibly with some correction factors (such as digital literacy, Human Development Index, etc.). Whatever the details, it is not particularly hard to come up with sensible taxation principles: indeed, contrary to data dividends, DSTs have been implemented in at least 38 countries by the end of 2023, with EU countries leading the way. A further indication of their effectiveness is that tech companies subjected to multiple DSTs (many from the United States) have started lobbying against it, on the ground that such measures constitute double taxation and produce significant

revenue loss.[4] A possible way forward would be to supersede the need for DSTs in individual countries, by reaching a global agreement on the reallocation of part of the profits of multinational companies to the countries in which they operate, as proposed by "Pillar One" of the two-pillars strategy under discussion by the Organisation for Economic Co-operation and Development's (OECD). Unfortunately, the feasibility of this global plan is in question, since countries where large multinationals are primarily based, like the US, are unwilling to sign it until DTSs have been abolished worldwide, whereas states that have already successfully increased their revenues with national DSTs, like several EU countries, are understandably reluctant to relinquish an established tax instrument (Avi-Yonah et al., 2022). Whatever the outcome of these multilateral negotiations, the very existence of growing international concerns on digital taxation, occasionally exploding in actual tax wars,[5] indicates that how to redistribute benefits and costs of the digital economy is one of the most urgent challenges for our society.

Whereas public debate on data dividends and digital taxation tackles the need for fair allocation of profits, it is important not to take for granted that personal data and user-generated contents are automatically "up for grabs", simply because they are valuable. The fact that our private lives have, in the digital economy, significant market value does not imply that we are no longer entitled to privacy rights. In a world of compulsive web searches and constant social media exposure, it might sound self-defeating to invoke privacy rights, since people seem all too eager to relinquish their privacy in favor of the public spotlight. Yet the contradiction is only apparent. Consider personal pictures: I might be fine sharing my photos with multitudes of undetermined strangers, and yet remain unwilling for that picture to be used in training a generative AI that will later produce pornographic content. This unwillingness is not only rooted in the very concrete possibility that such technologies may be used to create deepfakes and share non-consensual intimate images, e.g. in the context of revenge porn (for discussion, see Viola & Voto, 2023): it applies also to apparently legitimate uses of our audiovisual materials. Next time you stumble upon an AI-generated picture of a beautiful, half naked model (some pointers in that regard are provided in the next section, by the way), ask yourself what causal link you would be willing to contemplate between that image and pictures of your friends, your relatives, your partner, or even yourselves. Would you be fine knowing that your personal photos are being used to teach AI systems how to create erotic images and videos? What role would you prefer for your pictures to play in such training, that of the hot models to emulate or that of the ugly counterexamples to avoid at all costs? Whatever your answer to similar questions is, the concerns they raise signal that such decisions need be a matter of public debate, not something left to arbitrary choices made by the private sector.

[4] Source: https://bipartisanpolicy.org/blog/taxation-in-the-digital-economy-digital-services-taxes-pillar-one-and-the-path-forward/ (last consulted on December 29, 2023).

[5] For instance, in 2019 the US imposed a 25% import tariff on French luxury goods as a retaliatory measure against France's Digital Tax Bill.

This is why the General Data Protection Regulation (GDPR) of the EU represents a significant step in the right direction. Granted, its first perceivable impact was to make the life of European Internet users miserable, forcing them to provide consent every time they visited a website. Yet in time it became apparent that having that option was crucial, especially since it revealed the huge variety of uses to which our data were subjected by a multitude of private companies, previously unbeknownst to us and based only on tacit consent. Moreover, this gave users the opportunity to experience various ways in which their consent may be asked for, including some shady "dark patterns" (Soe et al., 2020). For instance, giving the option of rejecting all cookies is much fairer than forcing the user to check a long list of individual entries one by one; similarly, assuming refusal as the default option and asking users to opt in if willing is less manipulative than doing the opposite, i.e. making acceptance the default option and asking users to opt out if unwilling; similarly, suggesting default acceptance of options qualified by technical legal jargon (e.g., "legitimate interest") is suspicious. The more familiar users become with the practice of managing consent online, the more they will let subtle features of the consent form steer their behavior, which in turn invites careful consideration on how these forms are designed (Gerber et al., 2023). Nowadays, I am in the habit of automatically refusing to visit websites that ask my consent in an objectionable manner, unless there are no alternatives – which is rarely the case online, where very few websites can claim to be the unique provider of any given service or information.

Securing our right to effectively control the use of our data and contents, as well as making sure that we partake of the economic benefits they generate when we consent to their use, are ways of redressing some socio-economic imbalances introduced by new digital technologies, including generative AI systems. However, it is worth keeping in mind that these are *disruptive technologies* (Bower & Christensen, 1995): this is not a pejorative term, rather a statement of the fact that they create new markets and value networks, thereby displacing pre-existing business models and potentially making previous social and legal infrastructures obsolete. Thus, assessing and regulating their impact forces us to look not only at data and their value, but also at effects pertaining the job market, hopefully to avoid being made redundant by the latest wave of disruptive innovation.

4 Generative AI in the Job Market: Attack of the Creative Clones

Do you remember when robots were supposed to take care of all the boring, repetitive, tedious work, thereby freeing us to invest our time in creative, stimulating, leisurely activities? Indeed, the term "robot" comes from the Czech word "robota", which refers to a period of serf labor, therefore also, by extension, to concepts such as "drudgery" and "hard work". The same ambition was transferred to AI systems in general: besides any academic interest, the practical dream was to create machines intelligent enough to diligently perform dull yet essential tasks on our behalf, so that we could dedicate ourselves to more congenial occupations. Now, compare this aspiration with the current impact of generative AI on the job market: far from

offering us a new Eden of artistic renaissance, these technologies threaten to take over exactly the type of creative jobs that we would not want to relinquish them!

As a case in point, film writers and actors in the US have been vehemently protesting the use of AI in the film-making industry, with the Writers Guild of America (WGA) negotiating an important deal with the studios in late September 2023, after a prolonged "double strike" of both writers and actors that brought the industry to a standstill. The core tenet of that agreement is relevant here, because it hinges not on whether to use AI systems for writing scripts, but rather on *how* to use them, so that the rights of human writers are duly protected. This emphasizes that the key issue is not whether AI systems have the capabilities of producing good scripts (short answer: they do, with some qualifications), but instead on how their use for such purposes should be regulated, in the interest of all stakeholders. The deal agreed upon by the WGA was hailed as a significant win for human workers, since it prevents studios from using AI to write scripts or to edit scripts that have already been written by a human author, as well as from treating AI-generated content as "source material", like a novel or a stage play: the latter is crucial, because adapting pre-existing material entails much lower fees for human screenwriters, thus preventing the use of AI-generated "source material" serves to avoid easy exploitation by the studios. Simon Johnson, professor of entrepreneurship at the MIT Sloan School of Management with a long-standing interest in the socio-economic impact of new technologies (e.g., see Acemoglu & Johnson, 2023), described this as a "fantastic win for writers" and a potential "model for the rest of the economy", since "AI is under control of the writers, not under control of the studios, [and] it's not to be used as an automation technology. It's complementary to humans".[6]

Similar considerations apply to the agreement that settled the actors' strike in November 2023, although in that case the general consensus is that the studios got the upper hand: whereas clear provisions for the use of (and compensation for) digital replicas of real actors, living or deceased, were introduced, significant leeway was left with regards to synthetic performers, i.e. digitally-created assets that (i) are intended to create the clear impression that the asset is a natural performer who is not recognizable as any identifiable natural performer, (ii) are not voiced by a natural person, (iii) are not a digital replica of a real actor, and (iv) portray roles for which no employment arrangement exists with a natural performer. While the agreement allows the actors' unions to "bargain in good faith" whenever a synthetic performer is being considered for use instead of a human actor, there is nothing to prevent the studios from using AI-generated non-replica "actors" regardless of the outcome of such bargaining.[7] This leaves the door open to replacement of real actors with synthetic performers, once the technology is mature enough to ensure high-quality results; moreover, the economic incentives to explore that option are enormous for studios, since it would offer both a significant cut in costs and an increase

[6] Source: https://www.theguardian.com/culture/2023/oct/01/hollywood-writers-strike-artificial-intelligence (last consulted on December 29, 2023).

[7] Source: https://www.theverge.com/2023/11/18/23962349/sag-aftra-tentative-agreement-generative-artificial-intelligence-vote (last consulted on December 29, 2023).

in productivity. Maybe the actor unions are betting on the fact that the technology will never be good enough when it comes to videos, or that digitally created performers, regardless of their acting prowess, will lack the appeal of real stars. If that is the case, both assumptions seem very optimistic: recent developments demonstrate that generative AI has great potential for improvement, whereas the increased number and prominence of AI-influencers suggest that humans have little qualms in creating affective bonds with AI-personas; not to mention the success that such digital personas already have in specific cultural contexts, e.g. as members of K-pop bands in Asia, either in combination with human performers (e.g., Superkind, æspa) or on their own, in AI-only bands (e.g., IITERNITI, Mave:). AI-generated singers and influencers are already huge stars for large groups of people: betting that actors might be immune from this kind of competition is risky, to say the least.

Speaking of influencers, it is interesting to briefly discuss how generative AI is affecting and potentially revolutionizing that sector, which is, by its very nature, less structured and less unionized than more traditional job markets, such as writers, actors, and singers. Unsurprisingly, in the absence of clear guidelines, recently there has been a proliferation of AI-influencers: that is, profiles of a person (usually a model) whose pictures are AI-generated and whose social media pages are curated by human content creators – for now, however, since the possibility of automating that too, via AI-generated messages, is not remote, at least in principle. A notable example is Aitana Lopez, an AI-influencer that made the news in early December 2023 as "the first Spanish AI model earning up to €10,000 per month": of course, here "earning" is a shorthand for "generating revenues for its creators", yet the fact that it comes very natural for us to think of an AI-persona as an entity capable of "earning" money is part of the problem. However, Aitana is "employed" (i.e., created and never paid) by a modelling agency based in Barcelona, The Clueless, together with another AI-persona, Maia Lima: notice how the acronym "AI" is included in both their names, as a hint to their AI-generated nature – which is, by the way, made manifest on all social channels where they are active, although typically in subtle and elegant ways, so as not to spoil the illusion of reality. When visiting their page on the agency website (https://www.theclueless.ai/models), one is confronted with their pictures, followed by the words "more coming soon…", signaling the agency's intention of "enlisting" other AI models. Visiting the specific profiles of Aitana and Maia will let one discover more about their "personality", as well as giving access to more of their pictures: they are both conventionally beautiful, of course, but in different ways, since Aitana (pink-haired and tanned) is bolder and more extrovert in her poses, whereas Maia (white-haired and thin) conveys a more reserved and elegant kind of appeal. Their (fictional) personas are fully detailed, according to the type of image they are meant to convey, as follows:

> Aitana Lopez is a strong and determined woman, independent in her actions and generous in her willingness to help others. With boldness and authenticity, she faces challenges and expresses her opinion without reservation, although her complicated humor and self-centeredness sometimes make it difficult to get a smile out of her, showing her complexity. As a content creator, she shines with extroversion, attracting attention with her striking character. As a passion-

ate Scorpio, she highlights her love for video games and her dedication to the fitness lifestyle, evidencing her intensity and care for her physical well-being.
Maia Lima is a young Argentine girl who is characterized by her shyness and purity. She is an innocent and solitary person, enjoying her independence without emotional ties. Her loving nature is manifested in her bisexual orientation, which reflects her openness and diversity in her relationships. With her physical features, Maia embodies an exotic beauty. Born under the sign of Sagittarius, Maia is passionate about soccer, especially Boca Juniors, which reveals her energy and enthusiasm for this sport. Through makeup, she expresses her creativity and personality, finding a unique way to communicate. In addition, her love for music is reflected in her deep bond with favorite albums and playlists, which is a constant source of companionship. Photography and travel are her passions, giving her the freedom and vitality she seeks in her life.

In an interview with EuroNews,[8] Rubén Cruz, founder of the agency and lead designer of both AI models, reveals interesting details on the thought process behind their creation. It all started with a dissatisfaction with real models: according to Cruz, the agency was losing clients by "fault of the influencer or model and not due to design issues", so they decided to create their own influencer, in order to "make a better living and not be dependent on other people who have egos, who have manias, or who just want to make a lot of money by posing". Once their first AI-influencer, Aitana, was created, they soon discovered additional advantages of this business model, besides not having to deal with the quirks of a real human being, like asking for payment and reasonable working hours: for instance, they realized that brands "want to have an image that is not a real person and that represents their brand values, so that there are no continuity problems if they have to fire someone or can no longer count on them"; that is, the possibility of tailor-making the perfect influencer for whatever values or image a brand wants to convey, together with the insurance that such influencer will never step out of character, is extremely valuable for potential clients.

The obvious question, of course, is whether AI-influencers can be effective at what they are supposed to do, i.e., attract the attention of numerous followers and influence their lifestyle choices. While the latter is hard to assess for the time being, there is little doubt on their capability of garnering followers: when the EuroNews interview was published (December 2, 2023), Aitana Lopez had about 121.000 followers on Instagram, a number that was doubled in less than one month, with 243.000 followers as of December 29, 2023. "She" is also active on other social media, for instance by uploading photos of herself in lingerie to Fanvue, a subscription-based social media platforms similar to OnlyFans where content creators can share exclusive content with their fans in exchange for a monthly fee: in principle, such contents can be of any nature, yet since the Covid-19 pandemic similar platforms are almost exclusively dedicated to pornographic content. Needless to say, the

[8] Source: https://www.euronews.com/next/2023/12/02/meet-the-first-spanish-ai-model-earning-up-to-10000-per-month (last consulted on December 29, 2023).

money that Aitana "makes" from her fans on Fanvue go directly to her designers. Nor is she the only AI model trying to round up her "earnings" by being active on platforms that offer adult content to subscribers: Emily Pellegrini is another AI-generated celebrity that garnered considerable attention in recent months, sporting 134.000 followers on Instagram in late December 2023 and, more significantly, collecting as much as 10.000 $ over a 6 weeks period of activity on Fanvue, as well as attracting the attention of several celebrities who, failing to recognize her digital nature, tried to secure a date with her (prompting some cynical comments from online users, e.g. "it's not like celebrities are smart or anything, so let's not freak out!"). Not to mention Sarah Jordan, an AI-generated Australian beauty with 574.000 followers on Instagram and over 250.000 on X as of late December 2023, plus the usual moonlighting on Fanvue, offering adult contents to paying subscribers.[9] More generally, AI-influencers are by now quite normal in Asia, where the practice of engaging with virtual content creators has been around for much longer: the first South Korean cyber-singer, Adam, dates back to 1998 (admittedly with rudimental graphic results, by contemporary standards), whereas nowadays we have at least two prominent AI-generated K-pop bands, Eternity (later renamed IITERNITI) and Mave:, and as many as 150 AI-influencers active in the Asian markets, with collaborations with prominent multinational brands such as IKEA, Puma, and Asus.[10]

Overall, the current impression is that there is a lot of potential in AI influencers, as well as a marked interest from the industry. Moreover, the fact that they might replace real influencers (and are, in fact, designed to do so) does not seem to generate the same kind of push back observed for other professional categories, such as writers and actors. This largely depends on the less regulated nature of the influencer job market, and possibly also on a widespread perception that the role of influencer does not entitle to the same kind of protection common to other professions. Indeed, Rubén Cruz, in the aforementioned interview with EuroNews, had no qualms in making ethical arguments in favor of the use of AI influencers: firstly, he noticed that "Kim Kardashian makes a million euros for an Instagram photo and she doesn't cure cancer. Nobody earns a million euros for uploading a photo to a social network, it seems absurd to me"; secondly, he suggested that using AI influencers would help reduce market prices, thereby giving a chance to compete also to small companies that cannot afford big advertising campaigns. In short, AI influencers would be morally laudable because (i) they prevent real people from excessively monetizing their popularity, and (ii) they boost competition and free market in the advertising sector.

Writers, actors, and influencers are only some of the creative professions that are currently threatened by generative AI. In spite of the differences in how each category is dealing with this technological challenge, all scenarios have something fundamental in common: in every instance, generative AI constitutes a menace insofar

[9] Source: https://nypost.com/2023/11/11/lifestyle/onlyfans-rival-fanvue-bets-on-porns-fake-future-meet-emily-the-sites-hottest-eerily-real-ai-model/ (last consulted on December 29, 2023).

[10] Source: https://digital-business-lab.com/2023/10/virtual-influencers-in-asia-review-all-the-campaign-highlights/ (last consulted on December 29, 2023).

as it is framed as a substitute to human labor, rather than as an enhancement. It is worth noting that its potentialities, i.e. what generative AI can do, is nowhere under discussion, in these public debates: the fact that such systems can replace human performance, with various degrees of success and with different levels of autonomy from human intervention, is considered as a given, and rightly so, for the most part[11]; the real issue is how to regulate their use, so that the resulting benefits are fairly shared and the associated costs are minimized. In the words of Acemoglu and Johnson (2023), we need to *shift the focus from machine intelligence to machine usefulness*: the goal is not to develop intelligent machines, but rather to negotiate how to use these machines, intelligent or otherwise, in ways that maximize the common good, rather than increasing inequalities and allowing extreme exploitation by a well-positioned minority. The authors also suggest, and I concur, that we should not be overeager in jumping on the productivity bandwagon, i.e. the idea that any productivity-enhancing technology is by default to be applauded and embraced: this, on the contrary, overshadows the real issue, which concerns how the spoils of greater productivity ought to be shared among all members of society. Nor is the problem merely reducible to material payoffs: for instance, unemployment might turn out to be a non-ideal condition for human beings, even if they were monetarily compensated with the earnings of their digital substitutes. As noted by Farina et al. (2024), ethical considerations might require trading-off some efficiency of generative AI solutions, in order to preserve other values, e.g. safeguarding jobs.

The predicament of creative workers, in the face of AI-generated competition, is aggravated by the fact that it is hard for the general public to sympathize with them, precisely because they work in the creative sector, which is widely regarded as a highly coveted and privileged position. During the Hollywood strikes, spectators were significantly affected, with series being canceled and movie releases being delayed: while experiencing withdrawal symptoms for lack of their favorite pastime, many of those suffering entertained very uncharitable thoughts towards the protesters, along the lines of "What could possibly people like Steven Spielberg, Brad Pitt, and Scarlett Johansson have to complaint so bitterly about, for months on end?!". This implies that creative workers are unlikely to garner much sympathy from the public opinion, in their battle against AI-generated competitors, which in turn mighty limit their ability to secure satisfactory settlements, as well as failing to provide momentum for broader legislative solutions to the problem of AI-induced job displacement in such sectors.

[11] The success of generative AI in creative endeavors requires some qualification, since highly sophisticated statistical recombination of pre-existing elements (which is, in a nutshell, what these models do when they are being "creative") is still just recombination. Regardless of how good and human-like the results might be, there is room to doubt this will be sufficient to produce those "creative ruptures" that characterize human geniuses or, in any case, the most talented professionals in their respective fields. This might very well be the case, yet it does not mitigate the problems discussed in this paper: even if generative AI systems are just good enough to compete with other creative workers (which are not all geniuses, after all) and to tempt large groups of laypeople to delegate them a host of cognitive performances (as discussed in Sect. 5), this suffices to raise all the societal challenges that are being considered here. The thought that these systems will not be able to produce a new Leonardo da Vinci or Albert Einstein is, I fear, a meager consolation.

However, the fact that writers, actors, and influencers, in their struggle with the consequences of generative-AI for their respective professions, are faced with very different perspectives, as of the end of the year 2023, is indicative of an important truth: there is no "manifest destiny" for these technologies in human society. On the contrary, they are by their very nature highly malleable (Acemoglu & Johnson, 2023), in that they support a wide variety of uses: it is therefore up to shared political decision making to establish how their role should play out in the collective interest. Left to their own devices and maintaining the current legal infrastructures, market forces will keep pushing massively towards automation, embracing a view of generative AI as a substitute to human labor. There are, however, ways of promoting the adoption of such technologies to complement and enhance human performance, rather than replace it. The most obvious option is to make replacement of human workers illegal in certain contexts, as it happened in the agreement established between the WGA and the US studios in late 2023. Alternatively, AI-specific taxation can be used to make AI-generated work less convenient for companies, thereby curtailing their drive towards automation, while at the same time generating public revenue to support workers facing replacement by AI systems. There is also the option of investing in research on AI complementarity, to explore the impact of a different use of generative AI on productivity, as well as strengthening the voice of all interest groups on such matters (as mentioned, influencers suffer in their struggle against replacement by AI due to lack of professional representation), ensuring AI expertise within governments, to avoid making capital mistakes due to sheer ignorance of the problems, and boosting digital competence in end-users, so that they learn to better defend their rights (Kozyreva et al., 2020). Not to mention the importance of establishing clear ownership over the data (and the gains produced using such data) that feed the intelligence of generative AI and other digital technologies, as discussed in Sect. 3.

Going back to the old dream of "letting the robots do the dirty work", we can now appreciate its fundamental flaws. It is not only the fact that, ironically, nowadays AI systems threaten to replace us in highly creative activities that we would actually like to perform, rather than taking care on our behalf of the boring stuff. The problem is also, and primarily, with the whole idea of "being replaced by intelligent machines", whatever the task they end up replacing us at. This substitutive view of AI has always been very prominent, not only in science fiction (usually with apocalyptic undertones), but also in bona fide AI research, as a key benchmark: building a system capable of replicating human performance, or even surpassing it, is the standard criterion to define success in novel AI applications. Scientifically, there is nothing wrong with it: practically, however, it has a problematic tendency to convey the idea that, once we have built such machines, it is only natural to let them take care of business in our place. This is a recipe for disaster, from a socio-economic standpoint: thus, we ought to resist such tendency. In this, we might want to recall the admonition of Ian Malcolm, the well-known (fictional) chaos mathematician of *Jurassic Park*: "Scientists are actually preoccupied with accomplishment. So they are focused on whether they can do something. They never stop to ask if they should do something".

5 The Expropriated Mind

Misplaced enthusiasm for a substitutive view of generative AI technologies is problematic not only because it entails the risk of massive displacement of workers, but also because it pressures towards *technological replacement of cognitive functions and capabilities*. This is, once again, a perspective that has engendered significant enthusiasm in scholars, at least since the seminal paper on the extended mind by Clark and Chalmers (1998): in the following decades, the proposed approach, in all its variations,[12] garnered substantial philosophical interest and became one of the pillars of so called "4E cognition", according to which the mind is best understood as being embodied, embedded, enactive, and extended (Newen et al., 2018). At the core of the extended mind thesis is the functionalist intuition that what matters to qualify a process as cognitive is the role it plays, not the physical infrastructure that allows it to happen. The original formulation of this tenet is the famous parity principle: "If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process" (Clark & Chalmers, 1998, p. 8).

Conceptually, I have a lot of sympathy for the extended mind thesis, and it is undisputable that its research program prompted highly relevant debates on the nature of mental process and "the mark of the cognitive", to use an expression popularized both by supporters (Rowlands, 2009) and critics (Adams & Aizawa, 2010) of this approach. However, it is quite disconcerting to realize that, amidst such a vibrant debate, almost no one paid attention to the obvious (practical) elephant in the room: if mental processes are externalized to artifacts outside of our body, what is to prevent the expropriation of such artifacts, and therefore of whatever cognitive processes they were supposed to perform? If that happens, what are the consequences? Do we have any specific grounds to object against such expropriation, other than invoking standard property rights? Does the role these artefacts play in our mental processes give us any special right over them? Ironically, the issue of ownership came up frequently in the extended mind debate: Rowlands (2009), for instance, proposed ownership as a way of insulating the extended mind hypothesis against accusation of cognitive bloat (i.e., being too prone to accept any causally relevant external influence over cognitive processes as integral to them); yet the notion of ownership invoked by Rowlands, and by others after him (e.g., Gallagher, 2013; Smart et al., 2022), is not only hard to define, but also phenomenological in nature and related to subjecthood. In contrast, the concept of ownership that ought to preoccupy us is legal. Unfortunately, to the best of my knowledge, the key question "who (legally) owns the extended mind?" has been asked only once in the literature, in a somewhat obscure paper (Dunagan, 2015), buried in a handbook on intellectual

[12] Mapping the checkered history of the extended mind hypothesis is beyond the purposes of this paper, yet it is worth mentioning at least three main varieties of its theoretical tenets: active externalism (Clark & Chalmers, 1998), vehicle externalism (Hurley, 1998; Rowlands, 2006), and locational externalism (Wilson, 2004).

property: not exactly the most likely reading for philosophers of cognitive science, which might explain why the issue was not picked up in the broader debate on the extended mind hypothesis.

In the current technological scenario, I believe the issue must be reconsidered, with more emphasis on political implications and less enthusiasm for conceptual debate: digital technologies in general, and generative AI systems in particular, offer increasing opportunities to offload a rich variety of cognitive processes. This might be good news to vindicate the original externalist intuition, yet it presents all of us (externalists included) with a serious, practical challenge: *the more our mind is extended, the more easily it can be expropriated*, either by private companies or by public powers. A rarely mentioned but highly desirable feature of our good-old-fashioned central nervous system is that, barring illegal clinical procedures, it cannot be taken away from us: the same does not apply to the vast majority of the cognitive artefacts that make our mind extend so seamlessly in the external world. In fact, most of them are not really "ours" to begin with: at best, we own the hardware (your personal computer, your tablet, your mobile phone – unless you are leasing them, as more and more people tend to do nowadays), but what really makes digital artefacts work is the software, and that is typically offered to us as a service for a limited period of time, at a price or for free (i.e., paid with our data and/or contents, see Sect. 3). The same applies to the majority of online platforms, including social media: as Mark Bonchek put it on the *Harvard Business Review*, "most social media is rented, not owned. Facebook, Twitter, and LinkedIn are your landlords and you just lease the space".[13] In this context, cognitive expropriation does not require any manifestly violent act: it would be enough to stop the flow of apps and services (which is something that most providers are legally entitled to do, under current provisions), and we will no longer be able to use those artefacts to perform the cognitive processes that we are now happily delegating them.

Is this reason for concern? Maybe we are better off delegating a lot of cognitive work to technological devices, thus we are genuinely happy to pay for their continued use; and, insofar as we are willing to pay, tech companies will have no reason to expropriate such tools, since that would be self-defeating for them. Now, anyone who believes this path will lead to a happy ending for all involved should first consider two formidable problems. The first is that embracing a societal model in which our growing dependency on technological devices to perform fundamental cognitive activities makes us subordinate to the interests of multinational tech companies is a very risky proposition, since it will further move the balance of power towards private interest and away from public oversight. Nor it would be any better is such private companies were state controlled (China, as of 2024, provides a reasonable approximation of that scenario), since the power would remain unhealthily concentrated in the hands of a very limited number of individuals, who would just happen to be state officials rather than industry magnates. The second problem concerns what exactly we are so eager to pay for: if it is something that effectively enhances

[13] Source: https://hbr.org/2014/10/making-sense-of-owned-media (last consulted on December 29, 2023).

our cognitive powers, then it is easy to see the appeal; but if it is something that *merely replace pre-existing skills and lead to their progressive loss, in order to induce unnecessary technological dependence*, then it is equally easy to spot that as a rotten deal, if not an outright fraud.

Thus, we are faced with a highly relevant question: are AI-based technologies aimed at extending our mind towards enhancement,[14] or are they skewed towards replacement of pre-existing cognitive functions? As Acemoglu and Johnson remind us (2023), the answer to this question is not set in stone: these technologies have the potential to play either role, or both, in our social lives, so it is ultimately up to our collective deliberation to steer them in a direction that will maximize the public good. This deliberation, however, cannot be safely left to the free interplay of market forces, since there is no reason to assume they will deliver collectively optimal results; in fact, there are grounds to suspect they will not, based both on past experience (the current scenario is severely unbalanced to the advantage of increasingly smaller minorities) and on general principles (*ceteris paribus*, replacement technologies are easier to market than enhancement technologies, since the latter require active engagement and effort by their users, whereas the former promise the opposite as their main selling point). The shift towards technological enhancement, as opposed to mere replacement, will not happen on its own: it requires active measures at all levels of society.

It is important not to set up a false dilemma, though: enhancement and replacement are not mutually exclusive, as mentioned, and there are cases in which it might be ok to let technological devices replace certain cognitive processes in a stable manner, even if this means losing the capability or the habit of performing them otherwise. The memorization of a large set of phone numbers is a good example: I belong to a generation that grew up without mobile phones, which implied (among other things) that, whenever I wanted to call someone on my cable phone, I had either to find their number on some written repository (my agenda or a phone book, those relics of old…), or simply recall it from memory. The latter was by far the most convenient option, so it was quite common for anyone to memorize the phone number of a variety of people: your close relatives and your love interests, but typically also some of your friends, colleagues, and even the stores or clubs that you frequented the most. At the age of 15, I could easily recall about 50 phone numbers, and that was absolutely normal: today, in the prime of my academic career, that figure is down to 2 – my own number and my wife's, whereas I shamelessly ignore those of my sons. And I have memorized those two numbers simply because I am often asked to input them in online forms, where using my smartphone to retrieve them would be impractical. On the plus side, of course, calling others by phone has

[14] Readers should be mindful that there is, in the philosophical and neurological literature, a technical meaning of the expression "cognitive enhancement", which tends to be confined to enhancements resulting in modified brain functioning (for discussion, see Savulescu & Bostrom, 2009), even when such modifications are not pharmacologically induced (Dresler et al., 2013). However, in the context in this paper, I see no reason to limit the use of "cognitive enhancement" to similar cases, so the expression should be understood in the broader sense of any positive extension of the cognitive capabilities of the agent.

become much easier, and nowadays my extended mind (me and my smartphone, in this case) can reliably recall several hundreds of phone numbers. Overall, this is an instance where I do not mind permanently offloading the cognitive task of "memorizing phone numbers" to a digital device.

Technological replacement, however, is in general suboptimal, because it generates dependency without empowerment. Comparing it with enhancement clarifies the point: we become dependent on enhancing technologies for achieving results that would be impossible without them, *and always were*. In this case, we are trading off our independence for an extension of our powers: motorized vehicles are an example of physical enhancement, whereas search engines exemplify well cognitive enhancement. Sure, we are all dependent on Google and similar platforms in our ability to search the Internet for information and services, yet such an ability constitutes a significant extension of our mental powers: if the gain is substantial enough, individually and collectively, then the resulting dependency will be justified. With technological replacement, however, we become dependent from an external device to perform a task that was already available to us, which we lose the ability or the inclination to perform as a direct result of the adoption of the new technology: here the loss of independence is not compensated by any extension of our cognitive powers, but rather by comfort and effort reduction. If the replaced skill or process is trivial enough to be forsaken without regret, as in the phone numbers example, then all might be well; but, as soon as the lost competence is a relevant one, we are stuck in an awful situation, where we have made ourselves dependent from proprietary technologies for no good reason.

Similar considerations are relevant overall, yet they become particularly urgent in the context of education, because it is during development that most cognitive skills are acquired through training – or fail to be acquired, if such training is absent due to technological replacement (a danger discussed in the literature under the label "cognitive diminishment"; see Kasneci et al., 2023; Mhlanga, 2023; Shiri, 2023). With respect to generative AI, how it will be used in educational settings will have a significant impact on the shape of future societies (for discussion, see Cassinadri, 2024). Thus, whether and how to regulate its educational applications should constitute a primary concern for governments and citizens alike. In that regard, I recently had a chilling experience, although it came with its own silver lining: I was sitting at a cognitive science conference, listening to a talk presenting a field study on the use of ChatGPT for source-based writing in school, when the speaker started lamenting the fact that students were clearly digitally naïve, since they insisted on using ChatGPT as a source, e.g. asking questions about the subject matter of the essay to be written, rather than doing the obviously correct thing, i.e., having ChatGPT write the whole essay on their behalf. The students' tenacity in using generative AI for enhancement rather than for replacement was stigmatized by the researchers as a sign of incompetence in the proper use of ChatGPT, without even contemplating the possibility that it was instead motivated by ethical reasons (i.e., the students did not want to cheat on a school assignment) and by a deep understanding of the rationale of educational activities (where the point is to learn, not to score points by letting someone or something else do the work). As a result, I was personally overjoyed by the students' performance in the task, but rather appalled by the researchers'

interpretation of it.[15] Such a blind attitude towards the use of generative AI in education goes beyond mere technological optimism, bordering instead on "technological inebriation": this is something both scholars, practitioners, and policy makers need to steer clear from.

This is not to say that generative AI should stay outside of schools and universities: besides being extremely hard to implement (the use of ChatGPT and similar software is already spreading in the private practice of students and teachers), any ban would be suboptimal, since it would prevent also useful applications of these technologies. There are several extremely valuable roles that generative AI may play in educational settings: let us just mention a few, limiting ourselves to LLMs for text generation, like ChatGPT. They excel at writing creatively in the style of specific authors, text typologies, or literary movements, provided that a sufficiently large number of relevant instances are included in their training dataset. In fact, if the prompt is specific enough, even a single instance may suffice. By way of example, this is what Henry Minsky, son of AI's founding father Marvin Minsky, got out of ChatGPT by asking it to "describe losing your sock in the dryer in the style of the Declaration of Independence"[16]:

> When in the course of household events, it becomes necessary for one to dissolve the bonds that have connected a sock to its mate, and to assume among the powers of the laundry room, the separate and equal station to which the laws of physics and of household maintenance entitle it, a decent respect to the opinions of socks requires that it should declare the causes which impel it to go missing.
> We hold these truths to be self-evident, that all socks are created equal, and are endowed by their manufacturer with certain unalienable rights….

There is a lot to unpack here, and plenty of fun to be had, for an educator interested in teaching their pupils about the elusive notion of "writing style": ChatGPT responses to similar prompts may be used first to check whether students recognize the style that is being emulated, then analyzed to understand what makes such stylistic features apparent, and finally treated as a benchmark to test the students' own skill at writing a text in that particular style. More generally, using ChatGPT as a "textual sparring partner" has great educational potential, beyond issues of style. As all academic instructors, in recent months I have been often confronted with students' written assignments that were the product of illicit use of LLMs. Like many other colleagues, I find two things to lament there: the first is, of course, the breach in moral code, that defeats the whole point of giving them that assignment to begin with (as well as calling into question the use of similar tests for final exams, by the

[15] The whole episode is deliberately narrated without too many details, to protect the identity of the researchers involved: they were relatively young scholars, and these observations were directly discussed with them, thus I believe they will correct course in due time, with no need to expose them to public scorn.

[16] Source: https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/ (last consulted on December 29, 2023).

way); the second, however, is the depressing fact that ChatGPT often writes much better than my average student. Most people will hail this as a stunning success of ML; for an educator, it is a red flag on poor literacy in the school population. However, generative AI might help improve this sad state of affair, insofar as it is not used to simply write the text on behalf of the students: educators might get in the habit of feeding their assignment to an LLM from the very beginning and then discuss the results with their students, using the AI-generated text as a frame of reference for reflecting on the structure and value of its textual output. "Can you do better than ChatGPT?" might even become a new, playful scoring criterion for written assignments, as well as analyzing where the AI-generated output is to be applauded, and where it could instead use significant improvement.

Speaking of improving AI-generated output, also the hallucinations and biases of generative AI can be leveraged for educational purposes, instead of being evoked only to scare students away from adopting these technologies: "Don't let ChatGPT write your homework, that thing is prone to hallucinations!", and similar admonitions. In contrast, AI-generated mistakes can be very instructive, both on how these technologies work (as demonstrated in the Introduction of this paper), and on how misinformation might emerge in general, both in human and artificial systems. As a matter of fact, human beings have been known to suffer from their own brand of epistemic hallucinations and biases, in the form of fake news, epistemic bubbles, confirmation bias, wishful thinking, polarization, and so on. Thus, reflecting on how even highly sophisticated and computationally powerful cognitive agents (LLMs, but also humans) can be prone to systematic error is bound to provide a healthy respect for the difficult task of acquiring reliable knowledge, even (and maybe especially) in an information-rich environment.

This is just a small sample of possible educational applications of generative AI, limited to text generation, although some of them would be easy to adapt for other form of generative AI: for instance, text-to-image models like DALL-E are also very good at "painting subject X in the style of Y", with all the resulting implications for art education. The general point, however, is that the educational value of these technologies hinges on their use for enhancement of cognitive skills, rather than their replacement. On that, those students that were mislabeled as "digitally naïve" by researchers were in fact demonstrating a profound wisdom: the correct intuition is to integrate generative AI in education as a tool for empowerment, not to rely on it to "do the dirty cognitive work" on our behalf. This, I believe, is true in general for all technologies based on ML: using them to extend our mind, a la Clark and Chalmers, can be a beautiful and useful thing, but only insofar as this moves us towards enhancement, with only limited replacement. At the end of the day, as per Acemoglu and Johnson's admonition (2023), the sensible practical goal is machine usefulness, rather than machine intelligence.

It is also important to clarify how these concerns are affected by a fundamental ongoing debate on extended cognition: namely, the distinction between the constitutive incorporation of a device into the cognitive system (a narrow and stronger sense of "extension"), and the frequent or even continuous use of a device to perform some cognitive function, but without leading to incorporation (a broad and weaker sense of "extension"). This distinction has engendered a significant amount

of philosophical work (notable examples include Clowes, 2015; Farina & Lavazza, 2022; Heersmink, 2013), resulting in a very nuanced taxonomy of cognitive artefacts, based on how these artefacts absolve their cognitive functions in relation to the users' capabilities: Fasoli (2017), for instance, distinguishes between substitutive, complementary, and constitutive cognitive artifacts. Moreover, this debate has implications on whether some technologies, including generative AI, should be considered as either likely harmful or potentially beneficial for education. Pritchard (2016) has framed this as a technology-education tension: on the one hand, introducing such technologies in educational practice might result in deskilling or cognitive diminishment for students, which is a very negative outcome; on the other hand, keeping them out of educational activities will prevent students from learning about and being trained with highly relevant technological tools in the protected and well-structured context of school instruction – that is, another undesirable result. Pritchard appeals to an "extended virtue epistemology" Pritchard (2016) to solve this dilemma: insofar as the new technologies are incorporated in the cognitive processes of their users, there is no diminishment and no deskilling, just a substitution of the physical substrate being used to perform the required function. Cassinadri (2024), however, has recently argued against this solution, pointing out that (i) the distinction between cognitive incorporation and mere functional embedding is both debatable in theory (Facchin, 2023; Varga, 2017) and hard to apply in practice (Farina & Lavazza, 2022; Heersmink, 2017), plus (ii) actual instances of cognitive extension in educational contexts are likely to be rare, so that most uses of new technologies for learning purposes do not fit that description, yet have great educational potential nonetheless.

Regardless of one's stance towards Pritchard's position, this discussion shows that establishing whether generative AI extends our cognition in the strong or in the weak sense matters for the second problem raised above: namely, its tendency to replace pre-existing cognitive competences, rather that enhance users with new skills. Insofar as systems like ChatGPT or DALL-E can be described as being fully incorporated in our cognitive processes, to the point that they become constitutive of them, then of course there is no longer need to worry about replacement: much as per Pritchard's suggestion, nothing is being replaced, except the physical substratum carrying out a certain cognitive function – in this case, writing a text or drawing a picture. Of course, it is not at all obvious that such systems can actually meet the rather stringent criteria for cognitive extension in the strong sense, and the burden of proof here falls on those who would like to defend such radical position: however, the option of defusing the replacement/enhancement problem, by reinterpreting generative AI as a constitutive cognitive artefact, is at least theoretically viable.

Yet the first problem we discussed in this Section, the one of ownership of these technologies, is certainly not defused by appealing to the constitutive nature of AI-based cognitive extensions; in fact, similar appeals only make it worse. Let us assume that generative AI is indeed to be considered as a fully incorporated extension of our cognition, no more and no less than any part of our central nervous system: should we therefore rejoice for the fact that now important parts of our basic cognitive capabilities are (literally) owned by private companies? Not at all! If anything, this frames our future as an even bleaker dystopian nightmare.

This has an interesting, general implication: the more strongly one believes that AI technologies are (or will soon be) integrated in our cognition, the more worried they should be of the fact that such technologies remain owned by private corporations, and mostly outside of any serious public oversight. Proponents of a strong version of the extended mind hypothesis should be the most vocal advocates against private ownership of cognitive technologies, because, in their view, legal ownership of the tech entails legal rights on somebody's mind, or at least part of it.

6 Conclusions: Things to do before the AI Apocalypse

Based on the arguments developed in this paper, the overall assessment on the societal impact of generative AI is best expressed as a list of caveats:

1. Greater transparency and explainability should be demanded, not only (and not even primarily) on how these technologies work, but mostly on who currently stands to gain and who stands to lose from their success, to avoid or correct any imbalance that might result from their diffusion. Such socio-economic explanations, incidentally, are not particularly hard to come by, since they are independent from the complexity of generative AI systems.
2. The role of our data in making generative AI systems effective (and therefore lucrative) needs to be appreciated, both to renegotiate the distribution of the resulting benefits, and to discuss to what extent our data should be appropriated for such uses in the first place. To put it succinctly, there is no reason to accept data colonialism as a *fait accompli*.
3. The opportunity of using generative AI systems to replace human workers for free in creative sectors is likely to result in massive displacement and a vaguely dystopic scenario (with machines doing creative, stimulating work, whereas humans are stuck in boring, repetitive jobs), unless proper political and legal actions are implemented.
4. Extending our minds with generative AI systems should be mostly dedicated to an enhancement of our cognitive powers, whereas replacement of pre-existing cognitive functions and skills should be kept to a minimum and carefully monitored in the public interest. The goal is to make us more intelligent, not less.

Notably, none of these concerns is meant as a reason to stop working on (and playing with) generative AI. All of them, however, are intended to stress the key political point of this proposal: what to do with generative AI is a matter of collective deliberation, not something to be left in the hands of self-interested minorities. On second thought, that needs rephrasing: what to do with generative AI *should be* a matter of collective deliberation, not something to be left in the hands of self-interested minorities. The current scenario leans towards the latter option, unfortunately, and that is something that must be corrected with the

utmost urgency. Hence the frequent call, in these pages, to focus on this order of problems, instead of dwelling on other very interesting (yet less pressing) debates surrounding generative AI and ML in general.

This includes also other types of concerns for the impact of AI on society, which I find both excessive and misleading. This is the case of the statement on AI risk issued by the Center on AI Safety in May 2023, a short, peremptory declaration that reads as follows: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war".[17] Notice that this is the entirety of the statement, not an excerpt. In other words: beware AI, because extinction is looming! Interestingly, this is not the pronouncement of a bunch of nerds who binged excessively on apocalyptic TV shows, but rather a document signed by many key players in AI research, including Sam Altman, CEO of OpenAI (the company who developed both ChatGPT and DALL-E, among other AI applications) and Bill Gates, who needs no introduction. However, despite the authoritativeness of the signatories (or even more so), evoking the extinction of the human race due to AI, without trying to explain what causal link should exist between the two, sounds like cheap propaganda and clumsy recycling of old science fiction tropes (from *Terminator* to *Matrix*, passing through many other examples), rather than serious scientific or political reflection. The promoters of the statement argue that the message is deliberately succinct, to find common ground for all those who are strongly concerned, but are worried for various reasons, and would therefore suggest different solutions to the problem. Giving voice to such internal debate, they claim, would "dilute the message" in terms of political and social effectiveness: better to proclaim that the end is near in Twitter format, to avoid confusion. Maybe so, yet the fact remains that this results in marketing pseudo-science and vagueness, not verifiable statements and rational arguments: every time the scientific community has tried to take this path, in the name of a misunderstood communicative effectiveness, the results have never been excellent, and sometimes awful.

The exaggerate and vague tones of similar concerns are not the only reason to be skeptical of their efficacy. There is also the suspicion that they might serve the role of scaremongering tactics, to divert collective attention towards allegedly catastrophic, yet currently undemonstrated dangers of AI, meanwhile silencing other, more mundane, yet very actual worries, like the ones discussed in this paper. Silly little things, such as social justice, workers' rights, fair redistribution of benefits, good education and human flourishing, tend to pale in comparison with the end of our species: hence evoking the latter has been known to help in distracting the masses from the former. When those warning us against the impending AI apocalypse are the same people actively working on new, state-of-the-art AI applications, all the while making fortunes on their exploitation in poorly regulated markets, it does not take a very suspicious mind to suggest that similar appeals might be disingenuous. The party line, in that regard, is that AI experts are genuinely motivated to participate by a desire to ensure that their work is put to the best possible use for society at large. Elon Musk, interviewed by The Seattle Times when OpenAI was in

[17] Source: https://www.safe.ai/statement-on-ai-risk (last consulted on December 29, 2023).

its infancy and he was one of its co-founders, went on record as follows: "We discussed what is the best thing we can do to ensure the future is good? We could sit on the sidelines or we can encourage regulatory oversight, or we could participate with the right structure with people who care deeply about developing AI in a way that is safe and is beneficial to humanity".[18] Establishing the degree of plausibility of this statement is left as an exercise to the readers, in the best tradition of mathematic textbooks. Alternatively, they might try asking ChatGPT for its views on the matter: that should be fun.

Regardless of whether the motivation behind apocalyptic warnings against the danger of AI for society is considered genuine or hypocritical, similar concerns should not distract us from attending to more urgent problems, since these options are not mutually exclusive. On the contrary, getting in the habit of curtailing the unbridled use of AI technology by private companies and making it instead a matter of public debate and collective deliberation can only help in averting whatever doomsday scenario may be looming just over the horizon. So, let's get cracking on regulating data flows and the resulting profits, the use of generative AI in the workplace, and its correct integration in our educational system and everyday practices: this will certainly put us in a better position to address the more existential concerns that, apparently, keep AI researchers awake at night. After all, there are plenty of useful things to do to make "AI safe and beneficial for humanity", a la Musk, before the AI apocalypse wipes us all from the face of the planet.

Acknowledgements I am grateful to the editors of this thematic collection, Mirko Farina and Witold Pedrycz, and two anonymous reviewers for their helpful comments on a previous version of this manuscript, which greatly helped in improving its quality.

Authors' Contributions Not applicable to a study authored by a single scholar.

Funding Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement. The preparation of this manuscript was supported by and integral to the research project "Boosting Human Wellbeing with Behavioural Insights" (B-Hu-Well), funded as part of the PRIN 2022 PNRR research program of the Italian Ministry of University and Research (MUR), project n. P202227LNS, grant reference number (CUP) B53D23030060001, funded by the European Union as part of the recovery plan Next Generation EU.

Data Availability Not applicable. All data and materials discussed in this study are in the public domain.

Declarations

Ethics Approval and Consent to Participate Not applicable. This study does not involve research with human or animal participants. Ethical approval is not required.

Consent for Publication Not applicable.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

[18] Source: https://www.seattletimes.com/business/technology/silicon-valley-investors-to-bankroll-artificial-intelligence-center/ (last consulted on December 29, 2023).

References

Acemoglu, D., & Johnson, S. (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. Hachette UK.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Adams, F., & Aizawa, K. (2010). *The Bounds of Cognition*. Blackwell.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Avi-Yonah, R., Kim, Y. R., & Sam, K. (2022). A new framework for digital taxation. *Harvard International Law Journal, 63*(2), 279–341.

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., Jacob Filho, W., Lent, R., & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology, 513*(5), 532–541.

Bertini, F. (2023). Artificial Intelligence and data privacy. *Sistemi Intelligenti, 35*(2), 477–484.

Bower, J. L., & Christensen, C. M. (1995). Disruptive technologies: Catching the wave. *Harvard Business Review, 73*(1), 43–53.

Cassinadri, G. (2024). ChatGPT and the technology-education tension: Applying contextual virtue epistemology to a cognitive artifact. *Philosophy and Technology, 37*(1), 14. https://doi.org/10.1007/s13347-024-00701-7

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Clowes, R. (2015). Thinking in the cloud: The cognitive incorporation of cloud-based technology. *Philosophy and Technology, 28*, 261–296.

Couldry, N., & Mejias, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Davenport, T. H., & Beck, J. C. (2001). *The attention economy: Understanding the new currency of business*. Harvard Business Press.

Dresler, M., Sandberg, A., Ohla, K., Bublitz, C., Trenado, C., Mroczko-Wąsowicz, A., Kühn, S., & Repantis, D. (2013). Non-pharmacological cognitive enhancement. *Neuropharmacology, 64*, 529–543.

Dunagan, J. (2015). Who owns the extended mind? The neuropolitics of intellectual property law. In *The SAGE Handbook of Intellectual Property* (pp. 689–707). SAGE Publications.

Facchin, M. (2023). Why can't we say what cognition is (at least for the time being). *Philosophy and the Mind Sciences, 4*, https://doi.org/10.33735/phimisci.2023.9664

Farina, M., Yu, X., & Lavazza, A. (2024). Ethical considerations and policy interventions concerning the impact of generative AI tools in the economy and in society. *AI and Ethics, in Press,*. https://doi.org/10.1007/s43681-023-00405-2

Farina, M., & Lavazza, A. (2022). Incorporation, transparency, and cognitive extension. Why the distinction between embedded or extended might be more important to ethics than to metaphysics. *Philosophy and Technology, 35*(1), 10.

Fasoli, M. (2017). Substitutive, complementary and constitutive cognitive artifacts: Developing an interaction-centered approach. *Review of Philosophy and Psychology, 9*, 671–687.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.
Gallagher, S. (2013). The socially extended mind. *Cognitive Systems Research, 25*, 4–12.
Gerber, N., Stöver, A., Peschke, J., & Zimmermann, V. (2023). Don't accept all and continue: Exploring nudges for more deliberate interaction with tracking consent notices. *ACM Transactions on Computer-Human Interaction, 31*(1), 1–36.
Heersmink, R. (2013). A taxonomy of cognitive artifacts: Function, information, and categories. *Review of Philosophy and Psychology, 4*, 465–481.
Heersmink, R. (2017). Extended mind and cognitive enhancement: Moral aspects of cognitive artifacts. *Phenomenology and the Cognitive Sciences, 16*, 17–32.
Hurley, S. (1998). *Consciousness in action*. Harvard University Press.
Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.
Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest, 21*(3), 103–156.
Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 103473.
Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. In: *FinTech and Artificial Intelligence for Sustainable Development* (pp. 387–409). Sustainable Development Goals Series. Palgrave Macmillan, Cham.
Montesi, D., Bertini, F., Sharma, R., Rizzo, S. G., & Ognibene, T. (2016). Digital platforms: Has the time come for competition regulation? *CCP Research Bulletin, 31*, 18–20.
Newen, A., De Bruin, L., & Gallagher, S. (Eds.). (2018). *The Oxford handbook of 4E cognition*. Oxford University Press.
Pritchard, D. H. (2016). Intellectual virtue, extended cognition, and the epistemology of education. In J. Baehr (Ed.), *Intellectual virtues and education: Essays in applied virtue epistemology* (pp. 113–127). Routledge.
Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology, 22*(1), 1–19.
Rowlands, M. (2006). *Body language: Representation in action*. MIT Press.
Savulescu, J., & Bostrom, N. (Eds.). (2009). *Human enhancement*. Oxford University Press.
Shiri, A. (2023). ChatGPT and academic integrity. *Information Matters, 3*(2), https://doi.org/10.2139/ssrn.4360052
Simon, H. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, communications, and the public interest* (pp. 37–72). John Hopkins University Press.
Smart, P. R., Andrada, G., & Clowes, R. W. (2022). Phenomenal transparency and the extended mind. *Synthese, 200*(4), 335.
Soe, T. H., Nordberg, O. E., Guribye, F., & Slavkovik, M. (2020). Circumvention by design-dark patterns in cookie consent for online news outlets. In *Proceedings of the 11th nordic conference on human-computer interaction: Shaping experiences, shaping society* (pp. 1–12).
Turing, A. (1950). Computing machinery and intelligence. *Mind, 49*, 433–460.
Varga, S. (2017). Demarcating the realm of cognition. *Journal for General Philosophy of Science, 49*, 435–450.
Viola, M., & Voto, C. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese, 201*(1), 30.
West, S. M. (2019). Data capitalism: Redefining the logics of surveillance and privacy. *Business and Society, 58*(1), 20–41.
Wilson, R. (2004). *Boundaries of the mind*. Cambridge University Press.
Zuboff, S. (2018). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.