AI & SOCIETY (2025) 40:2069–2081
https://doi.org/10.1007/s00146-024-02088-7

MAIN PAPER

Galactica's dis-assemblage: Meta's beta and the omega of post-human science

Nicolas Chartier-Edwards[1] · Etienne Grenier[1] · Valentin Goujon[2]

Received: 4 April 2024 / Accepted: 23 September 2024 / Published online: 1 October 2024

Abstract
Released mid-November 2022, Galactica is a set of six large language models (LLMs) of different sizes (from 125 M to 120B parameters) designed by Meta AI to achieve the ultimate ambition of "a single neural network for powering scientific tasks", according to its accompanying whitepaper. It aims to carry out knowledge-intensive tasks, such as publication summarization, information ordering and protein annotation. However, just a few days after the release, Meta had to pull back the demo due to the strong hallucinatory tendencies or underwhelming performances of the model. This article aims to study, through a critical threefold argument, the potential impacts of LLMs once deployed in the scientific value chain. Our first argument is a technical one. By examining the technicity of Galactica, it is possible to explain the descripancies between its promotional corporate discourse and abysmal outputs. Second, by going back to debates in both computer science and computational philosophy on the automation of abduction, we argue from the epistemological front that LLMs indeed cannot produce strong abductions and, therefore, claims about the automation of hypothesis generation remains chambering. Finally, our third argument is a sociological one. By conceptualizing the scientific field through Nancy Katherine Hayles' cognitive assemblage theory, we aim to outline the potential steering of science by LLMs, mainly through information ordering. The core of our argument rests on the assertion that excessive control on information risks contravening a certain serendipitous aspect inherent to scientific discoveries.

Keywords Artificial intelligence · Large language models · Galactica · Meta · Cyberneteics · Critical artificial intelligence studies

1 Introduction

In *The Scientist Speculates*, mathematician and cryptologist Irving J. Good discussed the social implications of artificial intelligence (AI). By comparing the training of a machine to the education of a child, he considered the possibility of an exponential explosion of machine intelligence in a large number of fields, including scientific research.[1] A few years later, in 1966, Good revisited the idea of a machinic "intelligence explosion" in one of his most famous articles, entitled "Speculations Concerning the First Ultraintelligent Machine" (Good 1966), now considered one of the founding texts on Technological Singularity and Artificial General Intelligence. In November 2022, the hopes entertained

✉ Nicolas Chartier-Edwards
nicolas.chartier-edwards@inrs.ca

Etienne Grenier
etienne.grenier@inrs.ca

Valentin Goujon
valentin.goujon@sciencespo.fr

[1] Institut National de la Recherche Scientifique, Québec City, Canada

[2] Sciences Po, médialab, Paris, France

[1] "At this stage there would unquestionably be an explosive development in science, and it would be possible to let the machines tackle all the most difficult problems of science. Many of the most pressing problems, such as those of medicine and of information retrieval, would make giant strides every month, and human scientists might have to take a back seat." (Good 1962: 194).

by John I. Good has been renewed, as a Meta AI research team introduced what it promoted as the first step towards "a single neural network for powering scientific tasks" (Taylor et al. 2022, p 2): the model for science called Galactica.

Akin to a modern-day *Tiresias*, "AI for science" models are now dubbed "oracles" by the technoscientific milieu and its critiques (Messeri and Crockett 2024). Machine driven text analysis replaced the augury's antique art of bird watching and is considered by some to be the solution to the problems created by the scientific publishing deluge.[2]By promising to offer an accurate perspective over the inscrutable vastness of the landscape, composed of by massive heaps of scientific papers churned day and night, these oracles might increase the distortions caused by the hyper specialization of researchers who now evolve in what amounts to a scientific monoculture (2024). The resulting "illusion of exploratory breadth" might mislead scientists into believing "they are exploring the full space of testable hypotheses, whereas they are actually exploring a narrower space of hypotheses testable using AI tools." (2024, p. 50) These tools could reduce the scope of research by enclosing scientists in areas strictly defined by the datasets used to train oracles, therefore severing potential connections to outlying data and social interactions. The auguries of the classical Hellenistic era were known to be cryptic. Delivered in the form of gnomic poems, the visions of the oracles had to be deciphered by their audience. In a sense, contemporary oracles also suffer from opaqueness, as their lack of interpretability even puzzles the specialists involved in their construction (Pasquinelli amd Joler 2020). Interpretability being itself a subject of discord within the research community (Lipton 2016), it is fully understandable that the struggles for AI sensemaking translates partially into conflicts about the definition of fundamental logical categories of inferences and their technical implementation into explainable AI. It is in this context of debate that the industrial promoters of oracles such as Galactica are setting high expectations for the prophesied (or at least, promised) forthcoming algorithmic thought.

We investigate "AI for science" as an object of sociological inquiry to contribute to the emerging social science field of critical AI studies (Roberge and Castelle 2021). To study *how AIs, understood mainly as large language models (LLMs) in the current twenty-first century machine learning paradigm, could transform scientific activities*, we settled on the controversial case of previously mentioned underperforming model: Meta's Galactica. Our aim, through the "dis-assembling" of Galactica, is to demonstrate that there exists actual issues on the technological, epistemological and sociological fronts that should be matters of concerns, especially in the context of aggressive promotion from Meta concerning their oracle's estimated performance and capabilities to accelerate and optimize scientific research. Our research question thus goes as follows: How could the insertion of LLMs in the scientific value chain[3] transform such a chain of operations? Furthermore, how do the marketing discourses and promises around and about AI contribute to the future shaping of science-making in itself? To answer these questions. We organized the article in a threefold structure covering the technological, epistemological and sociological existence of automated science models. First, we will first briefly sum up the controversy surrounding the deployment and pulling-back of Galactica from open access. The discussion on the controversy contributes to the argument by outlining the marketing discourse employed by Meta (notably by promoting the model through some of its most important *portes-parole*) and a technical description of Galactica itself to unveil issues behind the product's underwhelming, if not disastrous performances. Then, by "taking seriously" the claims made on the automatisation of abductive reasoning through AI by both influential computer scientists and computational philosophers, we will examine how such assertions contribute to the performative belief that AI can indeed generate new hypotheses in a myriad of scientific fields. The objective here is to demonstrate from an epistemological standpoint that the automation of abductive reasoning remains out of reach in the machinic horizon of post-human science. Finally, by revisiting Nancy Katherine Hayles' cognitive assemblage framework, we aim to produce a sociological critical analysis on the transformative effects of LLMs such as Galactica, once deployed in the scientific field. By contextualizing arguments promoting tools such as Galactica in an inherited fear of information overload passed down from cybernetics to contemporary computer science, we aim to critically challenge the potential control of hypothesis generation once such tasks have been delegated to machines. To this cybernetic vision of science, we oppose a mertonian, serendipitous-friendly, vision of scientific activity where the destabilization of researchers acts as an impulse for hypothesis-making (rather than a risk of disorganization) to stimulate new discoveries. Fundamentally, we wonder if Galactica, Meta's beta of science-oriented LLMs, stands as the alpha or the omega of post-human science?

[2] A phenomenon to which the AI industry, with its fixation on the "open science nexus", largely contributes (Roberge et al. 2019).

[3] Described as "the full range of activities which are required to bring a product or service from conception, through the different phases of production" Porter's value chain concept (Kaplinsky and Morris 2001) can be easily transposed in the academic world where the end products are peer-reviewed scientific articles. As presented in the literature discussing academic capitalism and the triple helix model (Etzkowitz and Leydesdorff 2000), science itself becomes of field of extraction for plus value, both for the scientists as individuals aiming to gain reputational capital in their field and industries aiming to capitalize on open science to fuel R&D.

2 Dis-assembling Galactica: on technicity and promotional discourses

On November 15th, 2022, the Papers with Code (PwC) platform X account introduced a large language model (LLM) for science, named Galactica, allegedly capable of performing a wide range of research tasks: “summarize academic literature, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins, and more.”.[4] In addition to the X thread, the release of the model was accompanied by the launch of a dedicated website (galactica.org) with the preprint of the article, released the next day on arXiv, accompanied by several examples of input–output pairs for several disciplines (machine learning, math, computer science, biology, physics) and tasks (translate code into plain English, simplify code in Python, find an error in a mathematical formula). Liked more than 7000 times, the X thread was widely shared by senior Meta researchers, such as Chief AI Scientist Yann LeCun and Vice President of AI Research Joelle Pineau, but also by well-known figures in the online AI research community, including researchers from Meta's competitors (DeepMind, Nvidia). Two days later, on November 17th, the same PwC X account announced that the Galactica demo had been paused, although the model would remain available for researchers wishing to reproduce the results of the preprint (for example, via the Hugging Face platform or the official GitHub repository). Shortly afterwards, Yann LeCun reposted PwC's tweet with the following bitterly ironic comment: “It's no longer possible to have some fun by casually misusing it [Galactica]. Happy?”. Weirdly marketed as a performant product, Galactica was now being reframed as a demonstration, a product still in its beta-testing phase. What can we make of such a rapid change in discourse around this model?

We begin by diving under the hood of Galactica, through the model card available on GitHub and the preprint published online by the nine Meta AI researchers (Taylor et al. 2022).[5] Following the introduction of the Transformer neural architecture by a Google research team in 2017 (Vaswani et al. 2017), the vast majority of contemporary LLMs adopted a variant of this original architecture, named "vanilla", using either an encoder-only version (e.g. Google's BERT model; Devlin et al. 2018) or a decoder-only version (e.g. OpenAI's GPTs models) with some modifications depending on the model. Originally developed for natural language processing (NLP), the Transformer neural architecture is now used in speech processing, robotics and computer vision, where the Vision Transformer architecture and its variants are now popular (Dosovitskiy et al. 2020). The potential issues associated with the massive popularity of the Transformer architecture relate both to the epistemic risks of algorithmic monoculture (Fishman and Hancox-Li 2022) and the economic risks of industrial concentration (Luitse and Denkena 2021). From this point of view, it is not surprising that the PwC research team hosted by Meta AI chose a decoder-only Transformer, adding to the vanilla architecture a few modifications linked to the size of the contextual window (2048 tokens) and of the training vocabulary (50,000 tokens).

Similarly, like many other LLMs, Galactica is available in several sizes—“mini” (125 million parameters), “base” (1.3 billion), “standard” (6.7 billion), “large” (30 billion) and “huge” (120 billion)—which, as the number of parameters increases, become deeper and deeper with the addition of so-called “hidden” layers located between the input and output layers. This trend towards scaling up the size of the models is not without consequences on the other technical components of the entire AI system: for example, training the largest size of the model ("huge", with 120 billion parameters) requires the use of 128 A100 Graphics Processing Units (GPUs) designed by the company NVIDIA. While the exact training time for this “huge” version is not known, such a huge amount of computing power remains the prerogative of AI research laboratories supported by the main industrial players known as 'Big Tech' (van der Vlist et al. 2024)—OpenAI and Microsoft, DeepMind and Google (Alphabet), Meta, Amazon Web Services, etc. However, regarding inference, the researchers mention that the maximum size of 120 billion parameters is justified by the fact that the model fits into a single A100 node, with 80 gigabytes of memory, which facilitates "downstream accessibility" (Taylor et al. 2022, p 10) for users. Moreover, even in terms of model size, the Galactica family is on a smaller order of magnitude than the LLMs with which it is comparatively evaluated on a set of scientific tasks or more general

[4] Launched in July 2018, the PwC platform brings together research papers, computer code, datasets and leaderboards organized by data modalities (images, texts, videos, audio, 3D), learning tasks (question answering, semantic segmentation, object detection, image classification, language modeling) and research areas (computer vision, natural language processing, reinforcement learning, speech processing, graph learning). Released under the open CC BY-SA license, these resources are easily added by users of the platform. These values of reproducibility and open access led the creators of PwC to join Facebook AI in December 2019 while affirming that it remains “a neutral, open and free resource” (Stojnic 2019).

[5] Introduced in 2018 by researchers at Google and the University of Toronto, model cards are "short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (…) and intersectional groups (…) that are relevant to the intended application domains" (Mitchell et al. 2019).

NLP tasks.[6] However, despite its smaller size, the authors write that on reasoning tasks, "Galactica beats existing language models on benchmarks such as MMLU (Hendrycks et al. 2020) and MATH (Hendrycks et al. 2021)" and that on knowledge-intensive scientific tasks, "Galactica significantly exceeds the performance of general language models such as the latest GPT-3" (Taylor et al. 2022, p 2). What could explain the results of the Galactica family on these specific and more general tasks compared to much larger models trained with more computing power?

Indeed, Galactica's originality lies not so much in the size of the models or in the computing power required to use them, but rather in its training corpus made up of 62.22 million scientific documents (i.e. around 106 billion tokens): papers from preprint servers such as arXiv, PubMed Central and Semantic Scholar; reference material from Wikipedia and Stack Exchange; knowledge bases such as PubChem Compound and UniProt; two highly-filtered subsets of Common Crawl, Scientific Common Crawl and Academic Common Crawl; academic GitHub repositories for several STEM disciplines; finally, nearly fifty prompt datasets dedicated to several types of tasks like question answering, entity extraction, summarization, dialog and chemical property prediction.[7]According to the authors, this “high-quality and highly curated” corpus stands out from the much larger training datasets of other LLMs based on “an uncurated crawl-based paradigm” (Taylor et al. 2022: 2). This primacy of quantity over quality implies less data curation and amounts to the integration of toxic content (hate speech, pornography, stereotypes, conspiracy theories, misinformation, etc.) into many vision-language multimodal datasets (Birhane et al. 2021), vision datasets (Scheuerman et al. 2021) and language corpora (Luccioni and Viviano, 2021).

Compared with the 300 billion tokens in GPT-3, the 780 billion tokens in PaLM and the 1.4 trillion tokens in Chinchilla, the number of tokens used to train Galactica may at first sight seem small but, by performing 4.25 passes (called epochs) on these tokens without overfitting, the researchers were able to train the models on a total of 450 billion tokens. Finding that the use of repeated tokens improves both upstream and downstream performance, the Meta AI research team concluded that "the "tokens → ∞" focus of current LLM projects may be overemphasized versus the importance of filtering the corpus for quality.” (Taylor et al. 2022, p 11). In addition to this effort to curate, deduplicate and tokenize the training data, the other strategy adopted to avoid scaling up the entire AI system (volume of the corpus, size of the model, amount of computing power) was to add, to the general corpora, nearly fifty datasets of prompts covering a wide range of tasks. This small amount of data (358 million tokens, i.e. around 0.3% of the total number of tokens) provided in the form of text prompts is not used for instruction fine-tuning, as is often the case, but more upstream during the model pre-training stage. Here again, one of the authors' arguments is that, unlike fine-tuning, adding textual prompts during pre-training does not require any specific downstream expertise, hence the idea that "making the model work out-the-box for popular tasks […] is more useful for users of the model" (Taylor et al. 2022, p 8).

In summary, the Galactica family of models is a set of LLMs designed for an a priori laudable purpose: to provide a new interface for streamlining scientific knowledge and facilitating the navigation of ever-expanding corpuses of publications. These models are very similar to other existing LLMs in their decoder-only Transformer architecture but, unlike these, their performance is based less on quantity (more parameters, data and computing power) than on quality (highly-curated corpus, specific tokenization, prompt pretraining). This attention to the quality of the training corpus seems to be coupled with technical choices that are rather favorable to the main intended users in terms of the openness of the model (non-commercial CC BY-NC 4.0 license), the computing power required for inference (a single NVIDIA A100 node), and the range of tasks covered (generalist prompt pre-training rather than specialized fine-tuning). Given these details, how is it that the beta of such a model for science has been paused in just two days? According to us, this can be explained by a series of discrepancies observed between the model itself, the promotional rhetoric surrounding its release and the actual uses of the demo publicized on X or elsewhere online.

These discrepancies are particularly visible in a X post written by the first author of the preprint, Ross Taylor, reacting to a VentureBeat article (Goldman 2023) published to mark the first anniversary of Galactica's release with verbatims from Joelle Pineau. For the former, the online display of the LLM on the website leads to seeing "a base model demo" as "a product", while for the latter, she evokes a *gap in expectations* between what she describes as "a research project" and what the public identifies as "a product". Although the comments in the preprint are more nuanced by mentioning several limitations (only open-access training resources, citation bias, lack of general knowledge), the Galactica website states that the model "aims to solve [information overload in science]", that it is trained on "humanity's scientific

[6] For example, other models benchmarked in the preprint include: the former OpenAI flagship model, GPT-3 175B (Brown et al. 2020); a particularly large model from Google, named PaLM-540B (Chowdhery et al. 2022); a previous Meta LLM, OPT-175B (Zhang et al. 2022); the multilingual model BLOOM-176B (Le Scao et al. 2022) developed by the open science initiative “Big Science” (Hugging Face, Inria, CNRS, etc.); finally, two models from DeepMind, Gopher-280B (Rae et al. 2021) and its successor Chinchilla-70B (Hoffmann et al. 2022).

[7] For a full breakdown of the Galactica training corpus, see the appendix of the preprint published on arXiv (Taylor et al. 2022, pp 42–52).

knowledge" and that it appears like "a new interface to access and manipulate what we know about the universe". When the demo was released, a tweet by Yann LeCun also mentioned that Galactica can generate "a paper with relevant references, formulas, and everything.". Far from being anecdotal, these ambitious statements contribute more broadly to the economics of techno-scientific promises around the release of models considered to advance the state-of-the-art within AI research (Dandurand et al. 2022).

These claims of performance and knowledge mastery clash with the poorly-factual and even outright false outputs generated by Galactica, such as the (in)famous examples of a Wikipedia article about bears living in space or a scientific paper about the benefits of eating crushed glass. Other cases of eerie or problematic outputs have also been reported (for example, a fictitious reference or author), going against the claims of a functionally performant model used for "scientific tasks which value truth" (Taylor et al. 2022, p 3). Another line of criticism focused on the difficulty of obtaining adequate responses from the model due to the safety filters implemented by its designers. For example, when a user queries it about "Queer theory", "Critical race theory", "Racism" and "AIDS", Galactica responds with the following message: "Sorry, your query didn't pass our content filters. Try again and keep in mind this is a scientific language model.". *This refusal suggests both a normative definition of what science is not* ("Queer theory", "Critical race theory") *and an intention to avoid topics affecting marginalized groups* ("Racism", "AIDS"). With the exception of three limitations (risk of "hallucination", frequency-biased model, highly confident tone) indicated on the website, there is no mention of specific measures to reduce this "hallucinatory" tendency, while no details are given about the safety filters put in place by the researchers.

While this can be partly explained by the opacity surrounding LLMs designed by Big Tech companies (Burrell 2016), other organizational factors may be at the root of this lack of information and anticipation regarding "hallucinations" and safety filters. In his Galactica postmortem posted on Twitter a year after the model's release, Ross Taylor states that the research team consisted of just eight people (nine including himself), which is "an order of magnitude fewer people than other LLM teams at the time.". During the launch, this reduced workforce was "overstretched and lost situational awareness" while losing sight of "the obvious in the workload we were under.". Following falling revenues and colossal investments in the metaverse, Meta implemented a hiring freeze in May 2022 before announcing, a month later, the departure of the Vice President of AI Jerome Pesenti alongside "a new decentralized organizational structure for Meta AI" (Bosworth 2022). One of these organizational changes aims to integrate the Responsible AI (RAI) team, created in 2019, into the Social Impact team. In October 2023, a Business Insider article revealed that this reorganization meant a reduction in the team's workforce, which was almost halved between 2021 and 2022 due to "competing interests and competing for resources" as well as "a lack of autonomy, and creating clear impact" (Hays 2023).[8] This also marks a shift in the RAI team's objectives, from anticipating potential problems when releasing new AI systems to ensuring compliance from a strictly legal point of view.

The rapid pause in the Galactica demo should also be contextualized in Meta's broader industrial strategy. In May 2022, Zuckerberg's company announced the release of a suite of models, called Open Pre-trained Transformer (OPT; Zhang et al. 2022), in order to offer a more open, ecological and data-oriented alternative to OpenAI's GPT-3 model. A few months later, in August 2022, another Meta AI research team deployed an interactive public demo of the BlenderBot 3 chatbot (Shuster et al. 2022), based on an OPT-175B model fine-tuned on a set of dialog datasets. Launched to collect feedback that could improve the chatbot, the demo quickly sparked the publication of several press articles about amusing (criticism of Facebook and Mark Zuckerberg) or more worrying (anti-semitic stereotypes, election-denying claims) conversations with BlenderBot 3.

However, unlike the Galactica demo, the blog posts and preprints accompanying the release of the OPT-175B and BlenderBot 3 models were particularly detailed, with more extensive technical documentation about limitations and evaluations, and do not include statements as ambitious as those announced online for Galactica. This more modest and cautious approach made it possible to avoid the series of discrepancies observed in the case of the Galactica demo between the model itself, the promotional statements surrounding its release and the reality of uses publicized online. Nevertheless, as Ross Taylor and Joelle Pineau point out respectively in the postmortem post and the VentureBeat article, the lessons learned from the Galactica demo have influenced the launch strategy of the popular LLaMA (Touvron et al. 2023a) and Llama 2 (Touvron et al. 2023b) models during 2023. Despite the demo's two-day existence, the legacy of the Galactica family is therefore far from being non-existent: in addition to its influence on post-Galactica LLMs from Meta, the different sizes of the model are still

[8] After a phase of "quiet layoffs" in September 2022, formal staff cuts affecting more than 10,000 Meta employees also affected product-focused (the Responsible Innovation team) and infrastructure-focused (the Probability team) teams during the latter part of 2022. These cuts continued to affect AI teams during the last quarter of 2023 with the disbandment of the ESMFold protein folding research group in August, the layoffs in the silicon unit called FAST (Facebook Agile Silicon Team) in October and, finally, the transfer of most of the remaining members of the RAI team to the Generative AI and AI Infrastructure teams in November.

downloaded several thousand of times several months after its release on the Hugging Face platform.

3 The messy logic of inferences and their alleged automation

While the technical argument on the dysfunctionality of Galactica informs us both on its unsuccessfulness both from a pure "output quality" perspective and the flaws of the promotional discourse surrounding its publication, we contextualize this controversy in broader epistemological claims made in both fields of computer science and computational philosophy on the reasoning capabilities AI in knowledge production. Failing to instill common sense in expert systems through deterministic computing, the computer science community has now massively engaged into probabilistic methods tied to the connectionist paradigm (Cardon et al. 2018) hoping that the automation of induction could lead to the emergence of synthetic knowledge. As this current phase of AI research unfolds, few scholars question the type of knowledge these systems aimed to produce. Amongst them, Beatrice Fazi reworked the classical debate initiated by Turing about machine intelligence and wondered if a "machine can think anything *new*.[9]" (2021, p. 813) Trying to escape from the simulative paradigm that defined the field for many years, the author tackles the issue negatively through a critique of discourses supporting "the absence of novelty in computation […] based on the fact that, in machines, everything is pre-programmed." (2021, p.819) Using the examples of emergent behaviors in automated systems and of genetic algorithms that are subject to mutation, Fazi thus claims that at least *some* branches of AI development fosters *some* level of autonomy and the creation of new knowledge.

The main contradiction in this argument is that, while trying to escape from the prison of simulation, the author relies on examples that could be construed as simulations, since most artificial life programs were developed using living organisms as operative metaphors. Examples of such abound in pioneering works such as Bonabeau, Theraulaz and Deneubourg's self-organizing hierarchies (1996) which are modeled after social insects. Epistemologically, we want to argue that it might be impossible to avoid any discussion of simulation when addressing the question of algorithmic thought. Furthermore, since it is very difficult to project oneself into the life of a primitive biological construct or even a simple animal such as a bat, as Thomas Nagel famously explained[10] (1974), maybe the best metaphor should be the workings of our own mind. To follow the inquiry launched by Fazi, we might want to reformulate her own reworking of Turing and ask *what could constitute the "newness" in algorithmic thought?* Otherwise, without any precision as to the qualities and aspects of this algorithmic thought, all there is left is a blank canvas contributing to the haziness of the field of AI (Roberge et al. 2020) and its problematic "thingness" (Suchman 2023). Moreover, concerning Galactica's "automation of research", what kind of interpretations could be drawn out of automated text analysis and what kind of hypotheses could be formulated by such systems? It is then inevitable to reconsider the seminal human inference typology proposed by Peirce and take a closer look at deductions, inductions and abductions as they happen (or don't) in AI systems.

The limits caused by the pre-programmation of machines described by Fazi is fitting for the deductive inferences powered by automated systems. According to the author: "to compute is to formally abstract, and thus to generalize and reduce into a logical and deterministic relation the dynamism of life and of thought that comes from lived experience." (2021, p. 819) Inferences produced through deduction do not produce any new knowledge, they are not synthetic in Peirce's words, since their results are contained within their initial predicaments. The deterministic programming of AIs, tied to the out of fashion symbolic paradigm, falls into this category. The scientists involved in the construction of knowledge and reasoning systems hoped to mimic common sense by creating a ground truth based on troves of data and mathematical rules. Even if the scope of these systems was reduced to match human expertise in a specific field, rather than trying to hack said common sense, they remained unable to create additional knowledge as they were finite. On the other hand, neural networks characterizing the current connectionist AI paradigm are associated with another type of inference—the induction. According to Pasquinelli, these probabilistic systems can provide "an extraordinary form of automated inference" and "can be a precious ally for human creativity and science" (2017 p. 9). However, since they "operate within the implicit grid of (human) postulates and categories that are in the training dataset," (2021, p. 11) they are prisoners of these categories and unable to generate entirely new ones. In short, there is a certain synthetic quality to the inferences that they can produce—they can

[9] Emphasis added.

[10] "But bat sonar, though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine. This appears to create difficulties for the notion of what it is like to be a bat. We must consider whether any method will permit us to extrapolate to the inner life of the bat from our own case, and if not, what alternative methods there may be for understanding the notion." (Nagel 1974, p.438).

highlight the existence of unseen statistical patterns, therefore performing what is dubbed a "weak abduction"—but they lack creativity, theoretical breadth, and will only underscore correlations in their fine-grained empirical dataset.

In a thought provoking article, Luciana Parisi argues that the enmeshment of computers in human activities coupled to the automation of inference production through AI caused a "shift in logical methods of decision-making" that should be interpreted as "a symptom of a transformation in logical thinking." (2019, p. 1) In order to support this thesis, she claims that all types of inferences described by Peirce—deduction, induction and abduction—have been automated in AI systems. *Such claims must not be taken lightly when studying the transformative effects of models such as Galactica on the future trajectory of science.* Quoting computational philosopher Magnani, Parisi claims that "since the 1980s abductive reasoning has been adopted by diagnostic and expert systems." (2019, p. 109) Recognizing that "abductive logic is mainly performed in automated models for medical diagnosis" (2019, p. 93), it seems that this type of automated inference has been mainly deployed in a single subfield of AI research. Although there is a widespread consensus on the automation of induction and deduction, we assert that it would be wiser to tread lightly when considering the automation of abduction. Indeed, these systems, often dedicated to medical research, do not generate entirely new hypotheses; sequentially brute-force combinations built from a fixed dataset within the confines of an AI isolated from the surrounding world *does not* support strong abduction. It is rather a reminder of the classic 1970s Mastermind game where a player had to guess a sequence of hidden colorful beans. In an article dedicated to the comparison of current "explainable AI" to the peircean inferences, Hoffman et al. temper the reach of the claims made by computational philosophers; they expose that "in some reports, what is referred to as abduction is actually a form of induction," and further add "that inductive generalizations from sets of cases in case-based reasoning" are mislabeled as such (2020, p. 2). In such cases, if computer scientists "defined abduction as the process of revising a knowledge base in order to fit the data," it would appear that we are far removed from automating abductions and that the use of the term, if "invoked in research on intelligent systems, it actually does little heavy lifting, either in cognitive modeling or in implementations." (2020, p. 8) If abduction cannot be reduced to a form of induction or the reversed version of deduction,[11] what would distinguish this type of inference from the others? According to computer scientist Erik Larson, ventures such as abductive language programming *never* emulated strong abduction since systems that were developed to run such formulas were plagued by an incapacity to emulate common sense. *Knowledge bases, however large, will always be too narrow to generate ground truth.* Also, without prior abductive inferences to determine which knowledge is pertinent in the construction of real world representations, it is impossible to program a system that will produce the necessary identified abductive inferences. Beyond this communicating vessel problem preventing the construction of a ground truth mechanism linked to the real world in automated systems lies a second hurdle: "The real world is a dynamic environment, which means it's constantly changing in both predictable and unpredictable ways," therefore it is very difficult to put it in a bottle and "enclose it in a system of rules" (Larson 2021, p. 125). The world is too vast and ever changing to be grasped at once without the help of abduction, which, in turn, needs ground truth. A third problem preventing automated systems from generating new hypotheses is that "abductive derivation of a rule is a creative act" and, as such, it involves imagination (2021, p.6). Counterfactuals, "inferences that do not exist in a dataset," populate the imagination of scientists and will support the creative act of abductive derivation (2021, p.174). If the real world is too big and changes too rapidly (and chaotically) to be captured at once; it would appear that information related to non-events linked to imaginary alternative realities also has to be taken into account. Maybe the solution to this capture problem relies on an entirely different strategy that would be based on continuous interactions with the surrounding world and counterfactual projections based on these exchanges; enmeshment rather than trying to run a simulation.

4 Post-human science: oscillating between cybernetic tidiness and serendipitous messiness

As previously explained, the inability of AI-oracles to contribute to "good science" can be attributed both to their technicity, such as the Galactica case has shown, and to the fact they are bogged down in the realm of weak abductions. This final section of the text aims to conceptualize, sociologically, AI models for science and their transformative effects

[11] While exposing the logical forms of deduction and abduction, Peirce himself produced an example published in the *Popular Science Monthly* (August 1878) under the title *Illustrations of the Logic of Science.* Peirce proposed that the synthetic inference form named *hypothesis* was in fact an "inversion of the deductive syllogism." This view has been widely diffused in computer science and is now used by some authors to support the idea that abduction is strictly about switching the case and result predicaments even if by "its very nature [abduction] cannot be an extended form of deduction, because it logical form (essence)is an egregious deductive fallacy." (Larson 2021, p.172).

through Nancy Katherine Hayles' concept of "cognitive assemblage" (2016, then 2017).[12] This perspective leads us to contextualize the deployment of LLMs as a response to what we observe seems to be a fear historically carried in computer science of information overload (Halpern 2014; 2022). By adopting Hayles' ecological approach to science through assemblage logic, we argue that LLMs such as Galactica (if they end up being functional) could steer the making of science in a direction where ordering and compulsive tidiness risks sterilizing the fertile grounds necessary for the emergence of new forms of knowledge.

Cognitive assemblages, conceptualized by Hayles and further developed by David Beer (2023) can be defined as follows:

> A cognitive assemblage emphasizes the flow of information through a system and the choices and decisions that create, modify and interpret the flow. While a cognitive assemblage may include material agents and forces (and almost always does so), it is the cognizers within the assemblage that enlist these affordances and direct their powers to act in complex situations. (Hayles 2017, p. 116)

It is important to note that cognizer[13]s, whether biological or technological, are defined by opposition to noncognizers as being able to "exercise choice and make decisions" (Hayles 2017, p. 31). The cognitive assemblage approach thus insists on the distribution of cognitive tasks through many *actors* and *actants* located in a very specific context. The insistence on cognition can be explained by how the author defines it, mainly "as a process of interpreting information in contexts that connect it with meaning" (Hayles 2016, p. 32). Following this approach, it is possible to conceptualize scientific activities as ecologically contextualized interactions between biological and technical cognizers, organized in such a way that their outputs equate with discoveries.

The deployment of AI, in this case Galactica, in a cognitive assemblage will inevitably contribute to its reconfiguration, as they are not fixed but ever-changing. By acknowledging that neither actors nor *actants* hold a monopoly on social production as they are both mutually participants, Hayles' concept pays attention to situations of asymmetry, disjunction, and conflict within said cognitive assemblage. We settled on this framework as it is exactly the distribution of cognition that is promised, if not prophesied, with the deployment of Galactica in labs, as previously demonstrated through the X thread boosterism and the whitepaper preprint promotional tone. In Meta's idealized and optimized science-making, scientists would be able to delegate cognitive tasks to the LLM, such as information ordering and retrieval for additional support in research. These tasks would form a technological unconscious operating in the background of laboratories. Humans and technical cognizers would thus engage in information production, exchange, and interpretation which ultimately amounts to a circulation dynamic distributed across the whole assemblage.

These dynamic circular exchanges between biological and technological cognizers are reminiscent of the cybernetic concepts of feedback loops (Wiener 2013) upon which it becomes possible to generate predictions between human and machine interactions.[14] In cybernetics, the aim of control is to prevent information entropy which risks destabilizing entire systems, leading to their incapacitation (Wiener. 2013, p.11). One of the key arguments Hayles makes while discussing cognitive assemblage is that this very kind of cybernetic control is no longer possible due to the propagation of technological cognizers and the scale of cognitive distribution that now operates across the globe[15]:

> The complexity of these assemblages, for example in finance capital, has clearly shown that control in the sense Wiener evokes it is *no longer possible* [emphasis added]. Cognition is too distributed, agency is exercised through too many actors, and the interactions are too recursive and complex for any simple notions of control to obtain. Instead of control, *effective modes of intervention seek for inflection points at which systemic dynamics can be decisively transformed to send the cognitive assemblage in a different direction*. (Hayles 2017, p. 203).

In the case of scientific practices, one of these inflection points is exactly the information ordering prior to the construction of hypotheses by scientists. The deployment of an LLM such as Galactica could imply the steering of the cognitive assemblage in a certain direction wished by Meta.

[12] *Hayles and Beer's position allows them to greatly reduce the field covered by cognition, therefore enabling simple preconscious biological sensory systems a glimpse of it. This "diminished" version ofc ognition stripped of reflexiveness can therefore be transferred to interpretative machines, even if they are devoid of any capacities when it comes to comprehension.The question of "machinic cognition" thus remains an open one.*

[13] According to Hayles, cognizers are not defined by the presence of consciousness as she defines cognition as a process which can happen unconsciously (Hayles 2017).

[14] Hayles directly draws on Wiener's cybernetics to conceptualize that information flows in a circulatory manner inside cognitive assemblages (Hayles 2017).

[15] An argument upon which Parisi, as we previously mentioned, draws on (Hayles 2017). While we disagree with the assertion that AIs have reached the point of automated strong abduction, we still agree with the fact that enmeshment of both human intelligence and machine intelligence is taking place across society and is ever up-scaled as new systems are deployed and that such distribution might be leveraged to produce radical new forms of knowledge escaping the pitfall of industry-led, capital-intensive, research schemes.

To understand the steering of the assemblage, we turn to Sociologist David Beer's conceptualization of AIs. He builds on Hayles' concepts by arguing that AIs must be understood through their promissory economies beyond being simple technological cognizers but rather, as *super cognizers* (Beer 2023). Building on the logic that technological cognizers are "decision-making-able", Beer suggests that AIs, imbued with the promises and dreams of both their promoters and users, should be considered as super cognizers that "stretch the limits of the known' (Beer 2023, p. 73)' in their decision-making. The promissory narrative of super cognizers, composed of marketing promises and fiction prophecies, is "fuelled by a sense of where the boundaries [of the known] might be and how they might be breached—and in some case how they might be deliberately identified and then subverted" (Beer 2023, p. 73). Super cognizers contribute to algorithmic thought by capturing and formalizing unknowns into known information through mathematical correlations.

Computational and automated technologies have long been concerned with images and vision, from optical character recognition to image recognition and now unto image generation and data visualization (Cardon et al. 2018; Crawford 2021; Mendon-Plasek 2021), but one of the objectives in the field of AI research and industry has always been the conquest of language through its mathematization, as Roberge and Lebrun argued (2021). As stated earlier, LLMs such as Galactica work through a form of statistical textual networking, where the composition of sentences does not rest on the projected meaning of said sentences or words, but rather on a string of predictions based on the model's training (Roberge and Lebrun 2021). LLMs thus qualify as super cognizers since they "decipher" correlations between words and sentences outside of what is already known about the meaning of enunciates but rather, on the statistical probabilities of having words and sentences follow up each other according to textual entries found on the internet or, in Galactica's case, in scientific text corpuses or databases.

Going further in the case of science's automation, the core of our argument rests on the premise that the automation of knowledge-making carries a fundamental cybernetic fear through it: information overload might lead to entropic disorganization. Indeed, Lepage–Richer traces this fear back to Wiener's conceptual universe, where science and knowledge production are designated as the safeguard against the "Augustinian evil", referring to nature's resistance to becoming an object of scientific knowledge (2020, p. 206). As Lepage-Richer summarizes: "knowledge, in the context of Wiener's Augustinian framework, was thus reformulated into the production of a localized, cybernetically enforced order against chaos and disorganization" (Lepage-Richer 2020, p. 207). The limits of knowledge must thus always be expanded, as this very expansion equates with the enforcing of order upon nature's latent menacing chaos. With the deployment of LLMs to assist in the ordering of scientific knowledge, it seems that science itself must be made "knowable" to prevent a disorganization of scientists through information overload.

This inherent fear of information overload and the potential ensuing entropic disorganization reveals what Halpern qualifies as a representation of the human brain (and human institutions) as a repository of accumulated information to be ordered for proper accessing and marshaling in decision-making processes; "In this account, human institutions are not designed; they emerge as the result of accumulating processes over time" (Halpern 2022, p.345). The deployment of LLMs to assist scientists in the very task of information retrieval and synthesis reveals that science itself doesn't escape this vision of accumulation, which presupposes the risk of losing oneself in the endless sea of Arxiv publications and, therefore, losing control and risking disorganization. On this account, according to the Galactica whitepaper:

> In May 2022, an average of 516 papers per day were submitted to arXiv (arXiv, 2022 in Taylor et al. 2022). Beyond papers, scientific data is also growing much more quickly than our ability to process it (Marx, 2013 in Taylor et al. 2022). As of August 2022, the NCBI GenBank contained 1.49 × 1012 nucleotide bases (GenBank, 2022 in Taylor et al. 2022). Given the volume of information, it is impossible for a single person to read all the papers in a given field; and it is likewise challenging to organize data on the underlying scientific phenomena. (Taylor et al. 2022)

From a critical standpoint, we assert that *the deployment of LLMs such as Galactica, acting as super cognizers inside the cognitive assemblage of scientific research, are thought of as control mechanisms and failsafes that could prevent entropy*. We acknowledge that the deployment of LLMs as a railguard against the publication deluge might not be an intentional endeavor pursued by Big Tech. Still, it is part of the promotional rhetoric that pushes for the deployment of such systems and beyond pure intentionality, the fact remains that the fact that LLMs' outputs depend on their training and it is that very dependency that could prompt them as control mechanisms. This cybernetic logic would steer the assemblage in a direction that would limit the potentialities for accidents in research, thus opposing what could be deemed a serendipity-inciting research orientation. To be clear, both serendipity and cybernetics are skeptical (or even hostile) to notions of design and planning (Merton and Barber 2004; Halpern 2022). Serendipity aims for a research environment that will provoke accidental discoveries while cybernetics understands distributed intelligence as an environment that could auto-regulate and eventually, auto-produce itself as long as it remains tidy and well organized. In cybernetic thinking, minds and markets are autopoietic and auto-regulating entities that must control themselves to avoid catastrophe (Halpern 2022), and we could now add science to the list with LLMs as control guardrails.

Since the cybernetic endeavor to automatically generate hypotheses and control the information excess attributed to the scientific publication deluge appears either as a paranoid-ridden failure or industrially interested strategy of scientific assetization (Roberge et al. 2019; Birch et al. 2020), we argue that it is of the utmost importance to consider more carefully the advantages provided by a serendipitous research framework. We can find keys for the establishment of such a framework in social sciences corpuses that have studied both science itself but also, artificial intelligence as a social phenomenon; we thus aim to bridge these two projects through sociologist Robert K. Merton's concept of serendipity. According to Merton, "Charles Sanders Pierce had long before noticed the strategic role of the "surprising fact" in his account of what he called "abduction"" (1948, p. 506.) Describing the work of scientists, Merton explains that "fruitful empirical research not only tests theoretically derived hypotheses; it also originates new hypotheses." (1945, p. 469) Abduction, the inference characterized by the invention of new hypotheses, might then depend on a serendipity component. This component might as well be a necessary part of the abductive inference generation process. If "the discovery by chance or sagacity, of valid results which were not sought for" is a key element of abduction, it would be worth considering how, and why such a creative event is triggered. The anomalous nature of the "surprising event" at the origin of the serendipity pattern, "because it seems inconsistent with prevailing theory or with other established facts," would push scientists in an effort to make sense of the observations and question the theoretic framework that allows their presence. (1948, p.506) A new hypothesis might well be needed to explain what is perceived. Also, it would seem there are non-rational notions of desire and curiosity at play in the causal chain that links the anomalous event to the serendipity pattern and the abductive inference. These observations seem to plead in favor of finding ways to increase serendipity in a research context in order to stimulate the production of abductive inferences, as opposed to the cybernetic securitization through AI. According to Merton and Barber, "scientific work is not cut and dried," "instead, there are always loose ends, and that it pays to be aware of those loose ends." (2004, p. 193) This assumption calls for a principle of "controlled slopinness" while elaborating research programs, organizing lab practices and fieldwork. (2003, p. 193) If "compulsive tidiness in experimentation is even more crippling than in other areas of life," research environments including digital infrastructures, should be designed in respect of the following principle: "since science is by nature dynamic, compulsive tidiness goes against the grain of science."(2003, p. 193) Would it then be possible to conceive a research framework that would maximize serendipity by exposing both humans and digital cognizers to anomalous external events? We argue in favor of chaotic computer infrastructures that reflect the dynamics of the real world and are opposed to the fever dream of trying to bottle up the whole thing in a single AI system disconnected from external influences.[16]

5 Conclusion:

Answering the question formulated in the introduction— does Galactica stand as the alpha or the omega of post–human science— proves challenging if we are to acknowledge the full ambiguity of the controversy. On the first hand, it could be argued that we are heading toward a closure of further potential human-LLM collaboration; indeed poor performance coupled with over-hyped promotion, exaggerations on claims concerning the effective quality of machinic abduction and the fear of scientists being disorganized due to the publication deluge all contributed to discredit such products, especially coming from one of the magnificent seven.. On the other hand, the very ontology of AI models, which makes them highly dynamic and adaptable, testifying to AI's longstanding experimental epistemology, means that models such as Galactica must never be taken as " final facts, only ongoing experiments" (Halpern 2014, p.10). This is why the target of our critique is not artificial intelligence per se, but the underlying logic beneath the deployment of such systems in the scientific field, and its potential steering by said logic in a trajectory where danger, whether perceived or real, is neutralized. Still, following the tripartite structure of the article, we mustered an argument on how the insertion of LLMs in the scientific chain of value changes the context of scientific production itself. The insertion of LLMs in the scientific value chain creates a context where both hypergeneration and hyperinterpretation become necessary to keep up with the knowledge economy (see Kobak et al. 2024 for a quantitative demonstration). This hyperactive[17] context is one

[16] Computer scientist Yoshua Bengio proposes in his May 7th 2023 blog entry (https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/) to restrict advanced AI development to the domain of scientific research. Theorized as ethical and perfect artificial scientists, these AI systems would be entirely sealed off to avoid interactions with the world. Dubbed *pure scientists* in opposition to *experimentalist* systems by Bengio (2023), these assistants would be much safer since their inability to interact with the surrounding world makes them impervious to the influence of bad actors. We believe this approach will soon be considered as a sterile avenue since it cannot produce serendipitous events.

[17] We deemed the "hyper" prefix necessary to illustrate an acceleration and intensification of both the speed and production volume of the value chain, coupled with an increase of parsing workload. We believe this illustrates the change of context to which we keep referring.

where hallucinations, glitches, gaps and forms of linguistic standardization risk being normalized, as cases are already being zeroed-on (Cabanac 2023; Ansede 2024). The trick of the trade is that the deployment of LLMs on different points in the value chain acts as its own form of justification: hypergeneration calls for hyperinterpretation to discriminate good from t junk science, and hyperinterpretation calls for hypergeneration as LLMs require ever-expanding troves of textual training data sets. Respecting the ambiguity of the situation while trying to answer the research question leads us to argue that the deployment of LLMs such as Galactica transforms the scientific value chain by progressively imposing a growing reliance on themselves. The introduction of one model seems to justify the introduction of either other models, or the creation of new insertion points.

While this article remains mostly a theoretical incursion into the phenomenon, we believe that the current twenty-first century AI landscape, defined by media hype and massive resource allocation (Roberge and Castelle 2021) will transform the scientific value chain and that this precise situation requires further investigation. Furthermore, s ince AI-oracles such as Galactica seem to be incapable of generating new hypotheses, we believe it to be necessary to delve deeper in the conditions that enable abduction. Social sciences can support the conceptualization of an hybrid research assemblage that would take advantage of cognizers without stifling the creativity of scientists by embracing the chaotic and dynamic nature of the surrounding world, rather than its paranoid ordering by private interests. A closer study of the role played by surprise in scientific discovery might be a key to develop a better understanding on how abductive inferences are produced, and AI might act as such a tool for displacing the established parameters of the current knowledge production sequence by negatively demonstrating what is not a strong abduction and that knowledge is not to be bottled up out of context. Merton's work is, after all, a pledge for serendipity as the driving force of science through the disruption of *previously held knowledge* and the *irruption of surprising, anomalous, events*.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ansede, M (2024) Excessive use of words like 'commendable' and 'meticulous' suggests ChatGPT has been used in thousands of scientific studies. EL PAÍS. https://english.elpais.com/science-tech/2024-04-25/excessive-use-of-words-like-commendable-and-meticulous-suggest-chatgpt-has-been-used-in-thousands-of-scientific-studies.html

Beer D (2023) The tensions of algorithmic thinking: AUTOMATION, intelligence and the politics of knowing. Bristol University Press

Bengio Y (2023) AI scientists: safe and useful AI?. https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/

Birch K, Muniesa F (eds) (2020) Assetization: turning things into assets in technoscientific capitalism. MIT Press

Birhane A, Prabhu VU, Kahembwe E (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963

Bonabeau E, Theraulaz G, Deneubourg JL (1996) Mathematical Models of Self-Organizing Hierarchies in Animal Societies. Bull Math Biol 58(4):661–717

Bosworth A (2022) Building with AI across all of Meta. AI Meta. https://ai.meta.com/blog/building-with-ai-across-all-of-meta/

Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Amodei D (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc. https://doi.org/10.1177/2053951715622512

Cabanac G (2023) Signs of undeclared ChatGPT use in papers mounting. Retraction Watch. https://retractionwatch.com/2023/10/06/signs-of-undeclared-chatgpt-use-in-papers-mounting/

Cardon D, Cointet JP, Mazières A (2018) La revanche des neurones. Réseaux 211(5):173–220

Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Fiedel N (2022) Palm: scaling language modeling with pathways. arXiv 2022. arXiv preprint arXiv:2204.02311, 10

Crawford K (2021) Atlas of AI: Power, politics and the planetary costs of artificial intelligence. Yale University Press

Dandurand G et al (2022) Attentes et promesses technoscientifiques. Les Presses de l'Université de Montréal

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Houlsby N (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Etzkowitz H, Leydesdorff L (2000) The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. Res Policy 29(2):109–123

Fazi B (2021) Introduction: algorithmic though. Theory Cult Soc 38(7–8):5–11

Fishman N, Hancox-Li L (2022) Should attention be all we need? The epistemic and ethical implications of unification in machine learning. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency: 1516–1527.

Goldman S (2023) What Meta learned from Galactica, the doomed model launched two weeks before ChatGPT. VentureBeat. https://

venturebeat.com/ai/what-meta-learned-from-galactica-the-doomed-model-launched-two-weeks-before-chatgpt/

Good IJ (1966) Speculations concerning the first ultraintelligent machine. Adv Comput. https://doi.org/10.1016/S0065-2458(08)60418-0

Good IJ (1962) The social implications of artificial intelligence. In: The scientist speculates: an anthology of partly-baked ideas. Heinemann, London.

Halpern O (2014) Cybernetic rationality. Distinktion Scand J Soc Theory 15(2):223–238

Halpern O (2022) The future will not be calculated: Neural nets, neoliberalism, and reactionary politics. Crit Inq 48(2):334–359

Hayles NK (2016) Cognitive assemblages: technical agency and human interactions. Crit Inq 43(1):32–55

Hayles NK (2017) Unthought: the power of the cognitive nonconscious. The University of Chicago Press

Hays K (2023) Meta's Responsible AI team shrinks amid layoffs and restructuring, even as the company goes all-in on AI. Business Insider. https://www.businessinsider.com/meta-layoffs-responsible-ai-team-2023-10

Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2020) Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300

Hendrycks D, Burns C, Kadavath S, Arora A, Basart S, Tang E, Steinhardt J (2021) Measuring mathematical problem solving with the MATH dataset. arXiv preprint arXiv:2103.03874

Hoffman R et al (2020) Explaining AI as an exploratory process: The Peircean Abduction Model. arXiv, 1–13. https://doi.org/10.48550/arXiv.2009.14795

Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, Sifre L (2022) Training compute-optimal large language models. arXiv preprint arXiv:2203.15556

Kaplinksy R, Morris M (2001) *A Handbook for Value Chain Research*. International Development Research Center. Accessed 11 June 2024

Kobak D, González Márquez R, Horvát EÁ, Lause J (2024) Delving into ChatGPT usage in academic writing through excess vocabulary. arXiv e-prints, arXiv-2406

Larson E (2021) The myth of artificial intelligence. The Belknap Press of Harvard University Press, p 312

Le Scao T, Fan F, Akiki C, Pavlick E, Ilić S, Hesslow D, Wolf T (2022) BLOOM: A 176b-Parameter Open-Access Multilingual Language Model. arXiv preprint arXiv:2211.05100

Lepage-Richer T (2020) Adversariality in machine learning systems: on neural networks and the limits of knowledge. In: Roberge J, Castelle M (eds) The cultural life of machine learning: an incursion into critical AI studies. Palgrave MacMillan, pp 197–225

Lipton Z (2016) The mythos of model interpretability. In: 2016 ICML Workshop on Human Interpretability in Machine Learning, New York, pp 1–9

Luccioni A, Viviano J (2021) What's in the Box? An analysis of undesirable content in the common crawl corpus. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). https://doi.org/10.18653/v1/2021.acl-short.24

Luitse D, Denkena W (2021) The great Transformer: Examining the role of large language models in the political economy of AI. Big Data Soc. https://doi.org/10.1177/20539517211047734

Mendon-Plasek A (2021) Mechanized significance and machine learning: why it became thinkable and preferable to teach machines to judge the world. In: Roberge J, Castelle M (eds) The cultural life of machine learning: an incursion into critical AI studies. Palgrave MacMillan, pp 31–78

Merton R (1945) Sociological theory. Am Sociol Rev 50(6):462–473

Merton R (1948) The bearing of empirical research upon the development of social theory. Am Sociol Rev 13(5):505–515

Merton R, Barber E (2004) The travels and adventures of serendipity: a study in sociological semantics and the sociology of science. Princeton University Press, p 313

Messeri L, Crockett MJ (2024) Artificial intelligence and illusions of understanding in scientific research. Nature 627:49–58

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Gebru T (2019) Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). https://doi.org/10.1145/3287560.3287596

Nagel T (1974) What Is it like to be a bat? Philos Rev 83(4):435–450

Parisi L (2019) Critical computation: digital automata and general artificial thinking. Theory Cult Soc 36(2):89–121. https://doi.org/10.1177/0263276418818889

Pasquinelli M (2017) Machines that Morph Logic: Neural Networks and the Distorted Automation of Intelligence as Statistical Inference. Glass Bead 1(1):1–17

Pasquinelli M, Joler V (2020) The nooscope manifested: artificial intelligence as instrument of knowledge extractivism. aI & Soc 36:1263–1280

Peirce CS (1878). Illustrations of the Logic of Science: sixth paper Deduction, Induction and Hypothesis. Popular Science Monthly, 13. https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_13/August_1878/Illustrations_of_the_Logic_of_Science_VI*13*. Accessed 27 Mar 2024

Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, Irving G (2021) Scaling language models: methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446

Roberge J, Castelle M (2021) Toward an end-to-end sociology of 21st-century machine learning. In: Dans Roberge J, Castelle M (eds) The cultural life of machine learning: an incursion into critical AI studies. Palgrave Macmillan

Roberge J, Lebrun T (2021) BERT, GPT-3, Timnit Gebru et nous: l'intelligence artificielle à la conquête du langage. Sociologie Et Sociétés 53(1):235–257

Roberge J, Morin K, Senneville M (2019) Deep Learning's Governmentality. In: Sudmann A (ed) The democratization of artificial intelligence: net politics in the Era of learning algorithms. Transcript Verlag, pp 123–142

Roberge J, Senneville M, Morin K (2020) How to translate artificial intelligence? Myths and justifications in public discourse. Big Data Soc 7(1):2053951720919968

Scheuerman MK, Hanna A, Denton E (2021) Do datasets have politics? Disciplinary values in computer vision dataset development. In: Proc. ACM Hum.-Comput. Interact. https://doi.org/10.1145/3476058

Shuster K, Xu J, Komeili M, Ju D, Smith EM, Roller S, Weston J (2022) Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. arXiv preprint arXiv:2208.03188

Stojnic R (2019) Papers with Code is joining Facebook AI. Medium. https://medium.com/paperswithcode/papers-with-code-is-joining-facebook-ai-90b51055f694

Suchman L (2023) The uncontroversial 'thingness' of AI. Big Data Soc 10(2):20539517231206790

Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, Stojnic R (2022). Galactica: a large language model for science. arXiv preprint arXiv:2211.09085

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Lample G (2023a) LLaMA: open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Scialom T (2023b). Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 van der Vlist F, Helmond A, Ferrari F (2024) Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. Big Data Soc. https://doi.org/10.1177/20539517241232630

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is ALL YOU NEED. In: Advances in neural information processing systems, 30

Wiener N (2013) Cybernetics, or control and communications in the animal and the machine. The MIT Press

Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Zettlemoyer L (2022) OPT: open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.