The Re-enchanted Universe of AI: The Place for Human Agency

Helga Nowotny

Abstract Generative AI designed as a digital tool to communicate with humans raises a series of questions to what extent ChatGPT and its rivals approach, match, and eventually may surpass human cognitive abilities. I propose to situate their amazing performance in a longer historical perspective of the evolution of human knowledge that occurs by externalizing or "outsourcing" knowledge operations, beginning with the invention of writing. However, the latest developments confront us with digital technologies that render it increasingly difficult to distinguish genuinely human characteristics and abilities from artificially created ones. This renders the question of human agency and the extent to which it is transferred to machines as crucial for digital humanism, especially when we reflect on some uncanny resemblances with the immanent "enchanted" cosmic order in which our ancestors lived.

1 Introduction

The recent heated debate on ChatGPT opened fascinating debates inside and outside the AI research community on numerous existing assumptions about human language, understanding, emergent abilities of LLMs (large language models), and how much closer we are to having reached AGI (artificial general intelligence), even if it remains doubtful whether it will be achieved at all. ChatGPT stirred worldwide enthusiasm, promised significant productivity gains, and continues to give rise to serious concerns about the impact generative AI will have on open democratic societies and the lives of citizens. Seen from the perspective of digital humanism, this contribution places these developments into a larger historical and societal context and seeks to redefine the place for human agency.

H. Nowotny (✉)
Former President European Research Council, Vienna, Austria
e-mail: helga.nowotny@wwtf.at

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_13

2 Questioning Our Assumptions: What ChatGPT Is and What It Does

The release of ChatGPT in November 2022 by OpenAI, partnering with Microsoft, has triggered an unprecedented wave of public enthusiasm, inevitably mixed with anxieties and concerns. Other Big Tech companies quickly followed by releasing their version of generative AI, opening a fierce competition for market shares. The amazing efficiency and speed of development of this latest bunch of digital technologies took many by surprise, including experts. In the eyes of the public, ChatGPT and its ilk have become the vanguard representative of the previously often announced and hyped *digital revolution* which now seems to have arrived on everyone's computer screen and in the midst of society. It achieved what none of the previous AI advances accomplished—unleash a wide public debate on potential beneficial uses and on a long and open-ended list of potential risks. The rising geo-political tensions, especially between the USA and China, and the strategic competition that ensues serve to exacerbate the uneasiness with which this otherwise welcome technological innovation is greeted.

The processes undergirding what ChatGPT (and other generative AI) is and how it functions are known in its broad outlines. It is based on large language models, LLM, that are trained on a huge trove of texts (and images), including Wikipedia, Reddit, and raw web page scans to perform its core function which consists in simulating human conversations. It achieves this by estimating the probabilities of which words follow each other and which sentences follow each other. They are trained with the help of millions of parameters that include writing style, conventional phrasing, and tone, all designed to create the illusion of conversing with a human. They are scaled up, with yet unknown consequences how scale affects certain emergent "behaviors" of the model. Thus, they are great at mimicry, but poor at facts and unable to cite their sources. They are prone to errors and to "hallucinate," a concept popularized by Google researchers in 2018. It refers to mistakes in the generated text (or images) that are semantically or syntactically plausible, but factually wrong or nonsensical.

This immediately raises the question of how trustworthy these models are. The risk of mass-produced misinformation that could put into jeopardy any democratic election and other abuses come to mind, as well as the harm caused by false and unreliable medical diagnosis or other decision-making based on such models. These are some of the thorny issues that have since long been associated with AI when adequate regulation, at least for now, and the desirable, but difficult-to-achieve, alignment with human values are lacking. It is unclear, for instance, whether machines can be built that have values—and whose values these would be—or whether machines will learn values from which kind of data. The general feeling is that the genie is already partly out of the bottle and that all of us are participating in a huge experiment that the large corporations are conducting on us without ever having asked for consent nor informing us in transparent and honest ways as the experiment proceeds. Obviously, many of the weird or outright spooky features that users experience in conversations with chatbots at this early stage can and will be corrected. Yet, others may persist, or new ones be added.

There is also a wide consensus that these generative AI do not "understand" neither the questions nor the answers they give—as yet. However, during a very short period of early adoption, they have upturned a number of assumptions that were more or less taken as given and hence uncontroversial. Several are linked to the explicit goal of achieving AGI, artificial general intelligence, which, however, is not shared by everyone (Siddarth et al., 2021). A heated debate has erupted on whether large pre-trained language models can be said to "understand" language in a way that would refute the Chinese Room Argument, proposed by John Searle in 1980. This is not merely an academic debate, as the extent and ways in which they understand the world have real implications for numerous fields and kinds of applications that range from driving so-called autonomous vehicles to how we will entrust them with the education of our children. Does "understanding" on the part of the machine include understanding of the physical and social world encoded by language in human-like ways? And, related to this, which part of the knowledge we have about the world we live in depends on language and which part is captured and processed tacitly and by other sensory means?

At this moment, the AI research community appears divided on key questions that have arisen from starkly diverging arguments and assumptions around distinct modes of understanding, different definitions of the range of intelligence (including intelligence among non-human organisms), and the nature of language. It also matters how one compares human cognition, which includes flawed, non-logical reasoning prone to "irrational" thinking, with the inner working of an LLM that is purely driven by probabilities yet may produce incoherent and nonsensical output.

Thus, one side of the debate claims that LLMs truly "understand" language and can perform reasoning in a general way, although not quite yet at human level. The other side argues that large pre-trained models such as the latest GPT-4, despite being fluent in their linguistic output, will never be able to "understand" because they have no experience or mental models of the world. They have been trained to understand the form of language and exhibit an extraordinary formal linguistic competence, but not its function, let alone its meaning. They lack the conceptual understanding needed for human-like functional language abilities (Mitchell & Krakauer, 2023).

Another unresolved question concerns scaling and emergent abilities. There seems to be a threshold of complexity beyond which LLMs of a certain size display emergent abilities not observed in smaller models. Emergence is well known in complex systems when self-organization between component parts sets in and gives rise to novel phenomena. It is not yet well understood what happens when transformers that enable the rapid scaling up of parameters in a LLM lead to the emergence of new learning behavior and the ability to solve problems it has rarely or never seen before. Understanding how such transitions happen and why the threshold varies with task and model are open research questions (Ornes, 2023).

These are just a few of the fascinating issues that have come to the fore in the latest debates. Controversies are a unique opportunity to pry open the black box into which accepted scientific knowledge otherwise has been closed. Their eventual settlement rarely ends with a clear winning or losing side. Rather, they signal that exciting new findings may emerge, moving the field forward. Controversies thus are attractive for younger researchers and those practitioners who have not yet committed to one side, but are eager to find out more.

Some of the basic assumptions that underlie the more controversial issues related to consciousness, understanding, or being sentient have a long, and partly forgotten, conceptual and philosophical ancestry in previous debates and assumptions in the history of AI. They offer tantalizing glimpses of the possibility whether we are at the brink of a new, non-human form of understanding displayed by these models, rather than viewing them merely as "competence without comprehension." "LLMs have an unimaginable capacity to learn correlations among tokens in their training data and inputs, and can use such correlations to solve problems for which humans, in contrast, seem to apply compressed concepts that reflect their real-world experience" (Mitchell & Krakauer, 2023, p. 6). The field of AI may thus have created machines with new modes of understanding, almost like a new species which may be better adapted to solve different kinds of problems.

But different species are also better adapted to different environments. They are known to be engaged in niche construction, i.e., carving out a space from the environment in which they find themselves that promises not only their survival but their evolutionary future (Odling-Smee et al., 2003). From the perspective of a digital humanism, it might well be that the human species is engaged in its next niche construction on this planet: this time not mainly through ruthless exploitation of the natural environment, but constructing a digital niche in co-evolution with the digital tools created by them (Nowotny, 2020).

3 Technologies as Agents of Change: The Externalization of Knowledge Operations

The historical growth of new human knowledge can be interpreted as a sequence of major transitions in externalizing knowledge operations, processing and application, storage, dissemination, communication, and repurposing of knowledge, which in various configurations generate new knowledge. Often encoded in a new technological medium, knowledge operations extend what is possible, visible, feasible, and understandable. In turn, this impacts the society in which they become embedded, facilitating, channeling, or hindering what can be achieved. Thus, it is never the technology alone, but an assemblage of changes in mindsets, in behavior, and in the socioeconomic power structures that underpins major cultural transitions.

The first major transition occurred with the invention of writing as a social technology giving rise to the transition of an oral to a written culture. The transition involved the mastery of newly invented symbols (hieroglyphs, cuneiforms, alphabets), the novel combinations of the constituent elements, physical infrastructures using adequate materials (clay, papyri, animal skins), and social competences and skills required for collaborative functions and division of labor (specialization of scribes, transmission of skills, interpretative capabilities). Taken together, these preconditions form an assemblage that gives rise to novel cultural forms that can travel across time and space. The externalization of knowledge operations that previously had been stored in the memories of individuals and their oral skills to modify and transfer the content across generations were now inscribed in a physical medium. For the first time, language was encoded with symbols that could be read, interpreted, understood, as well as changed and transmitted in numerous novel ways.

The implications were vast. For the first time, a direct confrontation with the past as fixed in writing ensued. As the sources were few and the material precious, control over them strengthened the centralization of interpretative authority and led to a concentration of power in the hands of a small elite of priests and rulers. Libraries became the repositories of all knowledge that was available. Writing also implicated the loss of some cognitive facilities which, famously, was deplored by Plato as a decline in the ability to memorize a vast corpus of knowledge.

The next major transition is linked to *The Printing Press as an Agent of Change*, the apt title of Elizabeth Eisenstein's classical work in which she analyzes the capacity of printing in facilitating the accumulation and wide diffusion of knowledge. The adoption of print technology created new audiences and new industries around publishing. It enabled the revision and updating of old texts to incorporate new knowledge and forging new links with a widely scattered readership and helped to spread literacy and change the attitude to learning. New networks and transborder collaborations ensued, creating a more open, cosmopolitan environment which encouraged questioning and the spread of ideas.

Printing initiated a profound cultural change of mindsets, which ultimately marks this period as a crucial turning point in Western history. It had a major impact on the Renaissance with the revival of the classical literature, on the Protestant Reformation as it enabled the interpretation of the Bible by each reader and thus shaped religious debates, on the Scientific Revolution as printing rendered possible the critical comparison of texts and illustrations, and by encouraging the rapid exchange of novel discoveries and experiments, giving rise to the Republic of Letters (Eisenstein, 1979).

We now find ourselves amidst another cultural transition triggered by the amazing advances in the externalization of knowledge operations through AI. Put in extremely simple terms, the invention of writing enabled the externalized storage and accumulation of knowledge operations, while printing opened their wide dissemination and repurposing in socially inclusive ways. This was followed by the new media of information and communications technologies from the late nineteenth to the twentieth century. Their social effects were to overcome spatial and temporal distance and, in Marshall McLuhan's famous phrase, turned "the medium into the message." The internet is flooded with social media messages, "personalized" targeting of individuals, and user-generated content. Now, we have entered the phase of algorithmic intimacy (Elliott, 2023). More importantly, we are now extending the externalization of knowledge operations to a growing number of cognitive tasks, including the generation of new knowledge and information. In doing so, we continue to benchmark human performance against AI performance with the result that by moving the goalposts, we let AI score even more goals. The overall effect is AI agents as the externalized embodiment of human cognitive abilities, deliberately designed to resemble us.

4 They Are Like Us: The Re-enchanted Universe of AI

The arrival of AI agents as the digital "Others" evokes ambivalent feelings, oscillating between amazement, uneasiness, and fear. Techno-enthusiasm confronts alarmism. Serious concerns about the potential threats for liberal democracies are swept aside by debates about "existential risk" which conjure a double-bind situation between either being wiped out or to happily surrender to AGI. Admittedly, it is difficult to imagine what a co-evolutionary symbiosis between humans and machines will be like. Yet, they are here to stay, and it does not help to equate "non-human" with "inhuman" as recently claimed in an op-ed by a prominent US columnist (Klein, 2023). The capabilities of the systems we build resemble more and more those of humans, and although they work in ways fundamentally different from us, the inbred tendency to anthropomorphize lets us forget that we are interacting with machines (Weizenbaum, 1976). We continue to use words like "thinks," "behaves," "knows," and "believes"—familiar terms in the world we share with other human language users even if we know that it is not the case—and extend them without reflecting to AI agents that have been trained to imitate human language users.

Using anthropomorphic language blurs the line between "us" and "them" and amplifies the tendency to attribute agency to an AI. Leaving aside the questions whether, in principle, an AI system will be able to reason, believe, think, or understand, our world is being rapidly populated with AI artifacts that make it increasingly difficult to tell the difference. When ChatGPT answers, for instance, "sorry, it's not my fault, but I have been programmed like this," we are getting habituated to treat things that seem like us as if they were like us.

This is new—or have we been there before? After all, our ancestors inhabited a world they shared with non-human "Others" for the largest part of the history of humanity. These were gods of various standings, spiritual beings, ghosts, souls of plants and animals, and other "meta-persons," as cultural anthropologist Marshall Sahlins calls them. His posthumously published book is a moving tribute to the "Enchanted Universe" in which "Most of Humanity" lived. These meta-persons were immanent in the lives of humans. Together, they formed one big cosmic order in which, for better and worse, human fate was determined by meta-human powers.

They were present as decisive agents in every facet of human experience and in everything that humans did. They were the undisputed sources of success, or lack of it, involved when humans hunted and pursued their political ambitions: in repairing a canoe or cultivating a garden, in giving birth, or in waging war. Everything was the material expression of their potency, and nothing could be undertaken without evoking the powers of the meta-humans. It was an asymmetrical co-existence. In many respects, the spiritual beings resemble humans: they would lie and cheat, could be malicious, and were quarreling among themselves. Humans depended on their goodwill and benevolence which they sought through ritual invocations and cultural practices. The main difference, however, that conferred super-human power to them was their immortality (Sahlins, 2022).

This immanentist “Enchanted Universe” came to an end some 2500 years ago during the “Axial Age,” first analyzed by Karl Jaspers. The exact timing, the geographic reach, and the concept itself continue to be controversially discussed, but agreement exists that a new, transcendental order took over. In it, humans were separated from the higher powers that reigned above. Researchers working with Seshat, a large data set of prehistoric societies, associate this shift with the advent of moralizing punishing gods that correlates with the rise of social complexity in early societies (Turchin, 2023). The idea of a transcendental order is at the origin of the monotheistic religions and the fundament of modern societies that recognize the objective reality in which we live today.

We are convinced that our ancestors “only believed” in the Enchanted Universe, while “in reality,” they “knew” better. In other words, their Enchanted Universe was a perpetual, collective illusion. Sahlins refutes this interpretation. Our recent experience with digital “Others” that are amazingly efficient in taking over ever more human cognitive and physical tasks should make us rethink this condescending attitude. Just as we believe that an AI “understands us better than we do ourselves,” even if we know that the predictive algorithms at work are fed on data extrapolated from the past, or when we are seduced by a seemingly charming chatbot that, however briefly, we are speaking to a human, our ancestors might have felt the same. In this sense, we are entering a re-enchanted universe, even if it obviously differs from their animistic cosmos.

I am not suggesting that with the end of modernity, characterized as the Weberian disenchantment of the world, we are stepping back into the enchanted world of our ancestors, even if some uncanny similarities exist. But the transcendental bearings on which the modern world was built are beginning to change. The human species has drastically transformed the earth through the impact humans had on the natural environment in the short period of the Anthropocene (Frankopan, 2023). We continue to create numerous artificial entities, the non-human digital Others, with whom we have to negotiate to gain or retain control. We seem to have reached what Giambattista Vico adumbrated in his *New Science* (1711), namely, that “*verum (the true)*” and “*factum (the made)*” are interchangeable—we only understand what we made. The true and the made are reciprocal, each entailing the other.

Yet, our latest *factum,* the AI systems we are building, have so far escaped our full understanding of how they achieve what they do. For example, we transfer agency to them when we begin to “believe” that everything predictive algorithms tell us must come true, forgetting about probabilities and that the data are extrapolations from the past. At the heart of our trust in AI lies a familiar paradox. Just as we use computers to reduce complexity, we make the world more complex at the same time; we leverage AI to increase our control over the future and uncertainty, while at the same time, the performativity of AI, the power it has to make us act in the ways it predicts, reduces our agency over the future (Nowotny, 2021).

In the Enchanted Universe in which most of humanity lived, everything that was done happened with and through the decisive power of the meta-persons. If we believe that predictive algorithms "know" the future, do we not risk returning to a deterministic worldview in which human destiny has been preset by some higher power? If we are tricked by our anthropomorphic tendencies to believe we are communicating with another human, even if we are partly or fully aware that this is not so, does this not resemble the enchanted world of our ancestors?

Yet, the differences are stark as well. Theirs was a cosmos filled with spiritual life in which humans and nature were interdependent, while we continue to plunder the remaining resources of the natural environment. In its place, we are creating a virtual world, promised to be full of excitement that is intended to make us crave ever more and become addicted. The meta-persons of our days are the large monopolistic corporations that offer largely bland and cheap entertainment. Although we see through the virtual illusions created by them, we remain under their spell. They are the human agents behind the machinery designed to usurp human agency.

5 Redefining Human Agency

The most urgent task ahead is to make sure that the re-enchanted AI universe does not turn into a nightmare. We are rightly in awe when the algorithmic chatbot partners are better and faster in writing texts and providing answers to our prompts, in letting them program annoying computational tasks, or in asking them to come up with novel combination of texts and images. We expect that the next generation of LLM will even more unsettle the conventional ways of teaching and learning, of writing legal briefs, and of doing peer review and even research. There are other challenging discoveries ahead that nobody can predict, as uses of a novel technology always may take unexpected turns when users actively appropriate them, instead of remaining passive consumers. We will be affected cognitively as well as socially in our relations to each other and in our job opportunities in ways that cannot yet be predicted.

The erosion of the public sphere by social media and the threats posed to liberal open societies are likely to grow when the power to direct further developments, including the directions future AI research will take, is concentrated among a few powerful Big Tech players who are guided by the principle of "winner takes all" that underlies the concentration of economic and political power. Where in this rugged landscape, in which beneficial uses intersect with potentially bad and malicious ones, is the place for human agency? What is human agency and how can it be protected? Is there a place for future cultural evolution that is not only caught in the logic of profit-making but dares to resume some of the dreams of greater equality and enhancement of the potential that all humans possess, which existed at the early beginnings of the internet? Can our open democratic societies not only be rescued from threats like unleashing floods of misinformation and the dangers that come with the unprecedented concentration of power in the hands of a few unaccountable and unelected individuals, but can it provide a reconstituted and technologically savvy version of common ground in which citizens can meet and debate?

These are only some of the open questions that arise, seen from the perspective of digital humanism. They are urgent, given the accelerating development of AI and computer technology. Questions about the place of human agency remind us of what is at stake. It touches us deeply, as it is about the sense of control that remains, or needs to be regained, in our future living together with the digital "Others" that are supposed to serve, and not to dominate, us. Human agency is a fragile concept, shaped by historical circumstances and in constant need to be reassessed, reasserted, and redefined. It refers to the capacity of an individual to decide and act in an autonomous way, but its autonomy is always relative, shaped, and constrained by legal norms, values, and cultural habits. It is accompanied by assuming responsibility for one's actions and being held accountable for harmful consequences. In Western societies, human agency is very much tied to the notion of the individual and its freedom. And yet, living in a society presupposes the capabilities of individuals to organize themselves in ways that allow all members to participate and share the pursuit of a public good.

Thus, human agency has the potential of being pivotal in meeting the challenges ahead. They include the necessity of regulating AI technology and implementing it in ways that allow citizens to fully participate in the expected benefits, rather than privileging only those human agents who profit from the advances made by AI. Human agency must come to terms with the agency that is designed into the machines while raising awareness about the differences and similarities between human and non-human intelligence. Last, but not least, the process of redefining and reasserting human agency must be done in view of the inherent openness of the future. We do not know where the co-evolutionary trajectory on which humans and the machines created by them will lead nor whether or when something like AGI will be attained. Before reaching such a presumed endpoint, much needs to be done. In this sense, the future is now.

6 Conclusions

1. Human agency must be seen in a broader social context. It implies taking responsibility for one's actions and to be made accountable for harmful consequences. The human part in the interaction with AI is to be kept distinct from the artificial part, even if more cognitive tasks will be delegated to the latter. This entails safeguards/regulation against becoming "stochastic parrots" (Bender et al., 2021) or having to live with "counterfeit people" (Daniel Dennett).

   These are not only moral or ethical but societal and technical issues that contribute decisively to the kind of open and democratic society we want to live in.

2. There is a need to reconceptualize what is meant by “intelligence,” going beyond currently used criteria of human vs non-human intelligence. Human intelligence can be positioned on a continuum of living organisms (from bacteria to humans), e.g., by considering survival strategies in their environment.

   It raises the question whether we can do things in “a more intelligent way” and what this could mean.

3. Predictive, communicative, and decision-making algorithms will continue to influence our behavior and what we will become. It is therefore important not to transfer human agency to them, but to deal with them as tools to be used in a responsible manner with humans as last resort. This means keeping open the possibility of recourse and guarding against errors that might introduce arbitrariness.

   Otherwise, predictive algorithms can turn into self-fulfilling prophecies; communicative algorithms can transform us into stochastic parrots, and decision-making algorithms can erode the principles, like social justice, on which our open societies are founded.

Discussion Questions for Students and Their Teachers

1. What needs to be done to create better awareness and to counteract the inbred tendency to anthropomorphize when interacting with a chatbot/conversational assistant?

   Is it sufficient to design it to answer “I am only an artificial agent and therefore. . .”?

   Is it sufficient to make it evade or avoid any answer that would imply that it takes a political or normative stand?

   What needs to be done on *your* side?
2. As we are getting closer to interacting with “entities” equipped with an intelligence that differs from ours—where do you see similarities to the Enchanted Universe as described by Marshall Sahlins? Which are the main differences?

   Does the imagined world of science fiction fit into such a vision?

   Do we need to reconceptualize what we mean by “objective reality,” and if so, how?
3. Do you recall any personal experience in your work with AI linked to the feeling of (a) being in control and (b) discovering the illusion of being in control?

   What can be done to safeguard against the “Eliza effect”?

   How far does being in control extend beyond making the technology function as it should? Do *you* ever feel responsible for the effects it will have? Which ones?
4. Automation of decision-making can create a comfort zone by offering algorithmic recommendations, automated reminders, remote monitoring, etc.

   Do you agree with Anthony Elliott that “the self” is at risk to become numbed in an expanding world of predictive algorithms and that we are kept at a safe distance from our capacity for personal agency and self-reflection?

   Does it matter to *you*?

5. ChatGPT and other chatbots are here to stay, and more improved versions are to come. It can help you in your research as a bright, but occasionally unreliable, sloppy, or even lying research assistant.

   Do you intend to engage with it in *your* future work? How will *you* supervise it?

Learning Resources for Students

1. Anthony Elliott (2023) Algorithmic Intimacy. The Digital Revolution in Personal Relationships. Cambridge, UK.: Polity Press; see pp. 77–107.

   An acute analysis of how changes occurring today in intimate relationship are affected by machine learning predictive algorithms in the fields of "relationship tech," "friendship tech," and novel forms of self-care in "therapy tech." They impact the complex ways in which intimacy is understood, experienced, regulated, and transformed. It is not the "digital revolution" as such which threatens intimate relationships, but the re-orientation of various life strategies and lifestyles that change in accordance with automated machine intelligence. Alternatives are needed for different ways of organizing experiences of the self, society, and automation that encourage experimentation and innovation for an ongoing translation back and forth between the discourses of human and machine intelligence.

2. Divya Siddarth et al. (2021) How AI Fails Us. https://ethics.harvard.edu/how-ai-fails-us.

   The authors criticize the visions and practices behind "actually existing AI" as misconstruing intelligence (a) as autonomous rather than as social and relational and (b) the focus on achieving general intelligence defined largely as surpassing human-level cognitive capabilities which implies that outperforming human intelligence is a worthy and necessary goal and "solution" to many problems. They criticize (c) the shared commitment of AEAI to the centralization of capital and decision-making capacity, which involves scaling and reliance on a small elite. This is countered by a vision of "actually existing digital plurality" (AEDP) based on the principles of complementarity, participation, and mutualism. As an openly "advocacy think-piece" for a pluralist vision of the future of AI technology, the arguments are pitted against the dominant vision of existing AI power structures.

3. Melanie Mitchell and David C. Krakauer (10 February 2023) The Debate Over Understanding in AI's Large Language Models, arXiv:2210.13966v3 /cs.CL/.

   A good overview of the current stand of debate whether LLM can be said to "understand" language describing the arguments made for and against such understanding which shows a stark opposition in the views of the AI research community. The authors plead for the need to extend our understanding of "intelligence" which, arguably, would allow to include novel forms of "understanding" created by the extraordinary predictive ability of cases such as AlphaFold from DeepMind and to differentiate better which kinds of intelligent systems are better adapted for which kinds of different problems.

4. Murray Shanahan (25 Jan 2023) Talking About Large Language Models. arXiv: 2212.03551v4 /cs.CL/.

   A closer look, inspired by the philosophy of mind, at the language used to "talk about" language models, such as "belief," "knowledge," and "thinking" which are used by researchers as convenient shorthand for precisely defined computational mechanisms which fall within the range permitted by the "intentional stance." In contrast, today's LLM and their applications are so powerful that it becomes highly questionable that such license can still be safely applied. The author argues for the necessity of a shift in language, perhaps including new turns of phrase, to prevent the creation of a compelling illusion of being in the presence of a thinking creature like ourselves when interacting with a LLM-based conversational agent.

5. Marshall Sahlins (2022) The New Science of the Enchanted Universe. An Anthropology of Most of Humanity. With the assistance of Frederick. B. Henry Jr. Princeton & Oxford: Princeton University Press.

   A world-renowned anthropologist draws on his lifelong profound scholarship about the ways how "Most of Humanity" lived over thousands of years. The author exposes our Western-centrism and "transcendental" biases in the explanations we give of the imminent presence of meta-persons and supranatural forces in all human activities in previous times. The bold claim Sahlins makes is that we have to accept the organization and functioning of these immanentist societies with their own concepts and in their own cultural terms, rather than to explain them away as "mere beliefs" or convenient fantasies of the objective reality in which we live today. This raises the question whether the advances of AI open the possibility of a re-enchantment of our world by introducing in our midst digital "entities" or "beings" with whom we enter new forms of interdependence.

6. Helga Nowotny (2021) In AI We Trust. Power, Illusion and Control of Predictive Algorithms. Cambridge, UK: Polity Press.

   At the heart of our trust in AI lies a paradox: we leverage AI to increase our control over the future and uncertainty, while at the same time, the performativity of AI, the power it has to make us act in the ways it predicts, reduces our agency over the future. This happens when we forget that we humans have created the digital technologies to which we attribute agency and may result in self-fulfilling prophecies. These developments also challenge the narrative of linear progress, which played a central role in modernity and is based on the hubris, and illusion, of total control. We are now moving into an era where this control is limited through our various interactions with AI and giving it "autonomy" while facing the challenge of regaining control in the sense of clear attribution of accountability and responsibility.

References

Bender, E. M. et al. (2021). On the dangers of stochastic parrots: Can language models be too big? https://doi.org/10.1145/3442188.3445922. Accessed March 15, 2021, from https://dl.acm.org.

Eisenstein, E. (1979). *The printing press as an agent of change*. Princeton University Press.

Elliott, A. (2023). *Algorithmic intimacy. The digital revolution in personal relationships* (pp. 77–107). Polity Press.

Frankopan, P. (2023). *The earth transformed. An untold history*. Bloomsbury Publishing.

Klein, E. (2023). This changes everything. March 12. The New York times.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. Accessed February 12, 2023, from https://arxiv.org/abs/2210.13966.

Nowotny, H. (2021). *AI we trust. Power, illusion and control of predictive algorithms*. Polity Press.

Nowotny, H. (2020). *Life in the digital time machine*. Swedish Collegium for Advanced Studies.

Odling-Smee, F. J., Lala, K. N., & Feldman, M. W. (2003). *Niche construction: The neglected process in evolution*. Princeton University Press.

Ornes, S. (2023) The unpredictable abilities emerging from large AI models, quanta magazine, March 16. https://www.quantamagazine.org/print?mc_cid=5b30527cd0&mc_eid=593396f255.

Sahlins, M. (2022). *The new science of the enchanted universe: An anthropology of most of humanity*. Princeton University Press.

Shanahan, M. (2023). Talking about large language models. Accessed January 30, 2023, from https://arxiv.org/abs/2212.03551.

Siddarth, D. et al. (2021). How AI fails us. Accessed March 13, 2023, from https://ethics.harvard.edu/how-ai-fails-us.

Turchin, P. (2023). The evolution of moralizing supernatural punishment: Empirical patterns. In Larson et al. (Eds.), *Seshat history of moralizing religion*.

Weizenbaum, J. (1976). *Computer power and human reason: From judgement to calculation*. WH Freeman & Co.