*Erschienen in: GRAHN, Håkan, Hrsg., Anton BORG, Hrsg., Martin BOLDT, Hrsg. 35th Annual Workshop of the Swedish Artificial Intelligence Society SAIS 2023. Linköping, Sweden: Linköping University Electronic Press, 2023, S. 79-84. ISBN 978-91-8075-274-9 https://doi.org/10.3384/ecp199008*

# Understanding Large Language Models through the Lens of Artificial Agency*

Maud van Lier[1]

***Abstract*— This paper is motivated by Floridi's recent claim that Large Language Models like ChatGPT can be seen as 'intelligence-free' agents. Where I do not agree with Floridi that such systems are intelligence-free, my paper does question whether they can be called agents, and if so, what kind. I argue for the adoption of a more restricted understanding of agent in AI-research, one that comes closer in its meaning to how the term is used in the philosophies of mind, action, and agency. I propose such a more narrowing understanding of agent, suggesting that an agent can be seen as entity or system that things can be 'up to', that can act autonomously in a way that is best understood on the basis of Husserl's notion of indeterminate determinability.**

## I. INTRODUCTION

For the past few months, many news items have been devoted to the recent developments in AI-research. Large Language Models (LLMs) like ChatGPT have stunned their users with their ability to produce texts that are almost indistinguishable from texts written by humans. Especially in the production of standard texts, it seems only a matter of time (if it is not the case so already) that these systems will outperform the majority of human writers in writing (standard) texts. At the same time, the mass use of LLMs has shown some of their flaws as well. Where the texts that LLMs produce might seem meaningful qua content *to us*, semantics plays no actual role in the word-generation processes of LLMs. Generated texts can thus contain a variety of mistakes that may not be obvious to us at first glance, making their uncritical use problematic.

In a recently published article, Floridi [1, pp. 4–5] states that "the implications of LLMs and the various AI systems that produce content of all kinds today will be enormous. (...) Some jobs will disappear, others are already emerging, and many will have to be reconsidered". What Floridi seems most interested in, though, is the many challenges that philosophers face in trying to understand the nature and role of this new kind of artificially created 'agents'. We have, according to Floridi [1, pp. 5–6], succeeded in creating a new form of agency, one where 'the ability to act' has been successfully decoupled "from the need to be intelligent, understand, reflect, consider or grasp anything". LLMs like ChatGPT are, then, what Floridi [1, p. 6] would call "intelligence-free agents". Even though I agree with Floridi that the societal integration of LLMs raises many new and challenging philosophical questions, I do not agree with him that this is because we have, in creating these systems, "liberated agency from intelligence" [1, p. 6].

In this paper, I will defend two claims. First, I argue that current LLMs are not 'intelligence-free' in the way that Floridi claims that they are — as not requiring cognitive processes to produce meaningful output — and that therefore we have not yet succeeded in liberating agency from intelligence (section II). Second, I argue that rather than focusing on whether or not LLMs are intelligence-free agents, the more interesting question is whether they can be called 'agents' in the first place. Where it is not unusual to refer to artificial systems as agents in computer science and robotics, these agents have had little to do with the kind of complex entities that are generally referred to as agents in the philosophies of mind, action, and agency. The recent successes of LLMs, however, suggest that this second, and more narrow understanding of agent might (soon) be applicable to LLMs (and other artificial systems) as well. Yet, to be able to determine what artificial systems could then be called such agents, and under which conditions, we need a proper understanding of what it means to be an agent in this more restricted sense. As I will show in section III, Floridi's agent account is too broad to be used for this purpose and I therefore propose a more restricted notion of agent that is able to capture what being an agent in this second sense entails.

## II. COLLABORATIVE AGENTS

As mentioned in the introduction, Floridi [1, pp. 5–6] states that in creating AI-systems like LLMs, "we have decoupled the ability to act successfully from the need to be intelligent, understand, reflect, consider or grasp anything". In my understanding, Floridi means by this that LLMs are able to produce 'meaningful' (in the broadest sense of the word) content, without needing to understand why or how — or even *that* — this content is meaningful. LLMs can thus produce the same end-product (the text) as we can, without having to rely on the cognitive processes we use to produce this end-product. They can act without thinking — they are "intelligence-free agents" [1, p. 6].

In this section, I argue against this view by providing two objections to a labeling of LLMs as intelligence-free agents. First, I argue that current LLMs can only *act* in collaboration with (other) agents — humans — and that these other agents *do* make use of cognitive processes. This co-ability to act is thus not truly decoupled from intelligence. Second, and in line with this, I argue that the agent that can perform such a collaborative act is a *collaborative*

*Support by VolkswagenStiftung grant Az:97721 is gratefully acknowledged.

[1]M. van Lier is with the Philosophy Department, University of Konstanz, 78464 Konstanz, Germany maud.van-lier@uni-konstanz.de

*agent* that consists of a LLM and a human component. As long as a human is still supervising the act (or even part of this collaborative agent), the collaborative agent is not 'intelligence-free'. Taken together, these two objections then amount to a refutation of Floridi's claim that, at present, we have "liberated agency from intelligence" in creating LLMs [1, p. 6].

### A. No Agency without Collaboration

In a recent book, Russo [2, p. 18] points out that "scientists use and interact with machines throughout the whole process of knowledge production". Russo emphasizes that this knowledge production is therefore a *co*-production that takes place in a techno-scientific *practice*, one in which both parties (scientist and machine) play a fundamental role. I now want to argue that the 'act' of producing meaningful texts by LLMs should be understood in a similar fashion: LLMs do not generate texts *by themselves*. In practice, they *co*-produce texts together with their user. I use the word 'production' rather than generation here, as it emphasizes the active roles of both the system *and* the user of this system in the production.

Floridi [1, p. 2] himself points out that in learning to use ChatGPT, one must, among other things, learn "how to use the right prompts (...), check the result, [and] know what to correct in the text produced by ChatGPT". Without a prompt, ChatGPT will not generate a text. Without the right prompt or critical feedback, ChatGPT will not generate *meaningful* texts. At the moment, then, ChatGPT, or any LLM for that matter, does thus not generate texts in isolation. Rather, producing (meaningful) texts in these cases is a *collaborative production*, one that necessarily involves both the (human) user and the system. Let us now return to Floridi's statement: "we have decoupled the ability to act successfully from the need to be intelligent, understand, reflect, consider or grasp anything" [1, p. 6] . Is this truly the case when LLMs generate texts? I would argue against this. At present, LLMs can only produce a meaningful text *together* with their user, where the user still needs to be able to understand and reflect on the consequences of a particular prompt and on what it means for a text to be correct or meaningful. Yes, the system itself can generate texts without understanding, but the entire *practice of text production* requires the prompts and feedback of intelligent users.

### B. No Agency without Intelligence

Meaningful text production by LLM and user, when seen as an 'act', should thus be understood as a *collaborative* act. Only agents can act, so what kind of agent can perform this act?

In an attempt to close the 'responsibility gap'[1] that emerges when automated systems harm humans, Nyholm [5] shows the usefulness of the concept of *collaborative agency*. He argues that

> It does indeed make sense to attribute significant forms of agency to many current robotic technologies, such as automated cars or automated weapons systems. But this agency is typically best seen as a type of collaborative agency, where the other key partners of these collaborations are certain humans [5, p. 1203].

Nyholm calls all collaborators in a collaborative agency agents, but not all of them are independent or autonomous.[2] Collaborative agency is instead understood as a hierarchical model "where some agents within the collaborations [the automated systems] are under other agents' [the humans] supervision and authority" [5, p. 1203].

Given the earlier points made about the co-production of meaningful texts by LLM and user, I would say that this kind of text production is a paradigm example of an act performed by a collaborative agent. There is even a hierarchy where the user supervises the production by giving prompts and feedback. Calling current LLMs agents, I would say, is thus under the understanding of them being the *non-autonomous* and *non-independent* components of a collaborative agent. The acts of this collaborative agent are supervised by the human component, where this supervision still requires the human user to understand what makes a text meaningful, and what prompts would or would not work to create such a text.

Let us now return to Floridi's claim. Can we say that 'we have liberated agency from intelligence'? I would again at present argue against this. LLMs are by themselves neither independently nor autonomously able to produce meaningful texts. Rather, using Nyholm's concept, current LLMs are agents in the sense that they form a part of a collaborative agent. This agent, in its entirety, is not intelligence-free, because its human-component supervises and guides the act that is performed.

## III. ARTIFICIAL AGENTS

In the previous section I pointed out that, given the current involvement of human users in the eventual *meaningful* output of LLMs, a labeling of these systems as 'intelligence-free' agents seems to misrepresent the actual practice of their use. This was not to say, however, that what these systems can do is not a big step forwards in the development of more autonomous and advanced AI-systems. Quite the contrary: current LLMs show more than ever the need for the right conceptual tools to capture what it is that they can do better than, or different from, other systems, and what (dis)similarities they do or will share with us humans. What makes them interesting, though, is not that they can perform tasks without 'intelligence', but rather that they can carry out very complex tasks in an almost completely independent

[1]The term 'responsibility gap' refers to the fact that "there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume the responsibility for them" [3, p. 177]. For a critical view on whether or not there are such gaps in the first place, see also [4].

[2]Nyholm [5, p. 1202] states that his paper "questions the tendency to view these types of agency [the ones attributed to types of robots] as instances of autonomous or independent agency".

manner. This makes them appear more and more agent-like to us.

As mentioned in the introduction, using the term agent to describe artificial systems in computer science and robotics is quite common, but this use has had little to do with how the term is generally understood in the philosophies of mind, action and agency. Agents are understood in these fields as entities that things can be 'up to', that can behave autonomously in some sense. The quality of the output of LLMs, as well as the relatively simple prompts that they need to generate very complex output, suggest that — at least in the near future — we might be able to develop systems that can properly be called agents in this philosophical sense.

However, for us to actually be able to make an agent/non-agent distinction between artificial systems, it seems that we need a notion of agent that is more restricted than Floridi's use of the term. This is because, in Floridi's account, every system that has any kind of transformative effect on its environment can be called an agent, meaning that every artificial system already counts as an agent.

In this section, I will first present Floridi's notion of agent and compare it to an understanding of agent that is loosely based on agent accounts in philosophy. I show that Floridi's concept of agent does not carry any explanatory power on its own and is therefore not the preferred choice for a study of artificial systems as agents. Instead, I propose a more restricted reading of agent, one that, *already on its own*, can provide us with insight into, and standards for, current and future AI-systems.

### A. The Explanatory Power of Agency

In *On the morality of artificial agents*, Floridi and Sanders [6] defend the view that 'artificial agents' (understood as artificial systems) can be involved in moral situations both as moral patients and as moral agents. To make this argument, they first spell out what they mean by 'agent'. According to Floridi and Sanders, a human being (Jan) is an agent

> if Jan is a system, situated within and a part of an environment, which initiates a transformation, produces an effect or exerts power on it [the environment], as contrasted with a system that is (at least initially) acted on or responds to it, called the patient [6, p. 9].

As Floridi and Sanders [6, p. 9] state themselves, in this understanding of an agent, "there is no difference between Jan and an earthquake". The domain of systems that one can thus attribute agency to in this account is very large, since any system that causes some change in its environment counts as an agent.

There is a growing discussion about what kind of entities belong to the domain of agents, where the proposed entities range from humans as the paradigm case, to animals and organisms [7], [8], to groups [9] and artificial systems [10], [5]. A danger of extending this domain too far is that the term 'agent' as a categorizing concept loses much of its explanatory power. I would argue that this is the case in the account of Floridi and Sanders, since it has become difficult to meaningfully distinguish between for example an agent and a force, or an agent and a cause, when both humans and earthquakes are included in the domain of agents. After all, what a human and an earthquake have in common is that they are both natural systems that are subject to natural laws, whose continued existence affects some change in their respective environments. If this is the commonality that they share, would it then not be easier to call them both a cause or a force, or something similar? One could argue that both the human and the earthquake are the initiators or at least the initial loci for changes in their environment and that this is what makes them into agents rather than a cause or a force. But even here one could make the point that it is (fundamentally) implied in the meaning of a cause and a force that they initiate or affect something. Given the definition of agent that Floridi and Sanders provide then, it is not immediately clear what the added value is of referring to these systems as agents specifically.

This added value of the term agent is better captured in most agent accounts in the philosophies of mind, action, and agency.[3] In such accounts, an agent generally refers to an entity that can act: it can do things that are up to it, and in this sense it can act autonomously. This agent is then contrasted with an entity to which things merely happen, whose behavior — if it exhibits any — is non-autonomous in the sense that it never requires any active involvement of the entity itself. In this understanding of agent, the 'active involvement' (typically seen as goal directedness, intentionality, or control [11, p. 8]) allows one to meaningfully distinguish between an agent and a cause or a force: where both Jan and an earthquake cause changes in their environment, only Jan seems to have some sense of control over *how* he changes this environment.

In Floridi's understanding of agent, no such distinction can be made except by adding an adjective to the term agent that specifies what *kind* of agent we are talking about. Here one can think of examples like a *rational* agent, a *moral* one, or, in Floridi's latest article, an *intelligence-free* one. In his account, then, the explanatory power is shifted from the term agent itself to the adjectives that one uses in combination with the term. This is of course not a problem. The point that I want to make, however, is that this shift is unnecessary since the term agent, in a narrower understanding that comes closer to its use in the philosophies of mind, action, and agency, can carry meaning already on its own. And combining the term with the aforementioned (as well as other) adjectives will only enrich this meaning.

Before introducing my agent account, I first want to show in more detail why adopting a more narrow understanding of agent can be beneficial to AI-research. For now, I will assume that such an account requires at the minimum that certain things can be up to the agent — that it can act autonomously in some sense. Already with this provisional definition, we can (in principle) make a distinction between artificial systems that Floridi and Sanders cannot. In their

[3] See for an overview of the various positions in these fields [11] and [12]

account, both automated system and autonomous systems are agents as long as they initiate a change in their environment in some way. Any distinction between automated systems and autonomous ones can thus not be made clear by them by using the term agent alone.

I have said that this distinction can be made in principle, because we have gained nothing when we do not make clear what we mean by 'autonomous' behavior, especially when contrasted with 'automated' behavior. This distinction is not as clear-cut, because where mere automated systems cannot display autonomous behavior, autonomous systems do display automated behavior quite often. Just think of ourselves, the most paradigmatic agents out there. Much of what we do (and are doing) can be seen as automated behavior. When I walk while thinking for example, or when I am breathing. I could influence these doings, but do not necessarily have to. It is thus not that agents do *not exhibit any* automated behavior, but that they are capable of autonomous behavior too.[4] Where the difference between automated and autonomous behavior seems quite intuitive in living beings,[5] the distinction is less obvious in the case of (complex) artificial systems.

Let me try to explain what I mean. A recent advancement in fields like chemistry and materials science is the self-driving laboratory. Such a laboratory consist of an AI-system and a robotic platform, where the AI-system controls the robotic platform. The functioning of these laboratories is described by researchers as 'autonomous': the AI-system allows for the automation of both the design and execution of very specialized experiments within the limits of this robotic platform [14], [15]. Compared to systems that are merely automated, these laboratories are autonomous in that they can take over tasks that previously only the researcher could do, like the generation of hypotheses, and the design of, and control over, experiments that can test these hypotheses. Here it becomes interesting though, because where, compared to automated systems these self-driving laboratories seem to function autonomously, the question is whether we would continue to say so when we compare their functioning to our own way of doing things.

As said before, the term agent in philosophy generally refers to a system that things can be 'up to'. What does this mean in the case of the self-driving laboratory? The answer is not as clear. Self-driving laboratories function independently within a highly specialized context, where the number of experiments that can be executed is limited by, for example, the degrees of freedom of a robotic arm, or the overall technologies included in the platform. What is more, there is often an already pre-determined way to go about the design of experiments and the generation of hypotheses. Self-driving laboratories can take over routinized scientific practices. Where before only the experiment itself was automated, now the entire *practice* is automated, so both the technological and the human part, allowing for automation on a higher level. On the one hand, then, these systems function autonomously in the sense of executing a practice independently, while on the other hand they are limited in how they execute this practice given their training and the narrowness of the scope in which they function. Given these latter features, one could also make a case for a description of their functioning as one of high-level automation. The question then becomes whether being capable of the former kind of autonomy — independent functioning — is sufficient for being called an artificial agent in the narrow sense since it could be described as high-level automation too.

My aim here is not to speak in favor of either of the two options. What I want to point out, though, is that this is a fruitful discussion. One that is the result of adopting a more narrow understanding of agent, and then discussing whether a system could appropriately be called such an agent or not. It is a discussion that forced us to think about what we mean by automated and autonomous functioning in artificial systems, a domain of systems in which this distinction is less intuitively obvious than when talking about living systems for example. It suggested as well that functioning might be autonomous *up to a degree*. The question is now what we mean concretely when we say that things can be up to an agent. In the following, I will provide a more specific reading of this notion and reflect on how it can be used to think about LLMs and other artificial systems.

### *B. Autonomy as Indeterminate Determinability*

Thus far, I have worked with an understanding of agent that, even though it is more narrow than Floridi's use of the term, is still rather abstract. At the beginning of section III-A, I defined agents as entities that things can be 'up to', that are autonomous because it is the agent and nothing else to which certain things are up. This notion of agent is derived from the account of Steward [8][6] and has served me well until now because it is broad enough to potentially include artificial agents,[7] but is not so broad that the term no longer carries any explanatory power of its own. Of course, by saying that things can be up to the agent, nothing much has been said yet. Where we might have an intuitive sense of what this means, we need a more workable definition to determine what artificial systems can or could count as agents — we need to know what the kind of 'things' are that can be up to agents, and how such things can be up to them.

Agents, according to Steward, are able to 'settle things' in the sense that "any exercise of agency is always such that it does not have to have happened" [8, p. 104]. She claims that (much of) the animal kingdom can already count as such agents. There are many parts of Steward's account that

[4]Wu [13, p. 203] states that "to understand agency, we must see that every process can have automatic and controlled features". Even though I do not necessarily see in Wu's account how one can make a clear distinction between the two features, I think that it is important to recognize that every agent can have more or less control, but minimally has to have some control.

[5]One could refer to things as 'autonomous behavior is the behavior one is aware and/or conscious of', or compare autonomous behavior with mere reflexes.

[6]"Agents are entities that things can be up to" [8, p. 25].

[7]Even though Steward is sceptical about the possibility of us being able to create artificial agents, her position does not reject it as an option [8, See for example p. 15 & footnote 40].

capture what we intuitively associate with agents, and that should therefore be included in an agent account. However, it also makes an objection clear that any such account faces, including my own: the question of whether there can even be something like agency in a world that appears to be mostly deterministic. Even though answering this question exceeds what is possible in this paper, Steward has made a nice attempt to do so, and I will shortly touch upon it here.

Steward's overall aim is to defend an Agent Incompatibilist position. Agent Incompatibilists hold that agency itself is incompatible with universal determinism. Where most Agent Incompatibilists have argued that *human* agency is incompatible with universal determinism, Steward argues that *animal agency* makes already trouble for universal determinism. Her argument is that the overwhelming number of examples of entities that things seem to be up to — that are capable *making themselves move* rather than move by themselves — make it very likely that agents exist and that therefore universal determinism is not true [8, see pp. 12-15]. Even though I find Steward's arguments convincing, they are based on intuitive and logical reasoning, and not on any conclusive empirical proof. I will therefore not make any claim about whether things that appear to be up to certain systems are truly 'up to' them. I will merely hold that if it appears like they do, then it makes sense to refer to these systems as agents.

So what does Steward mean when she says that things can be 'up to' agents? As stated before, she holds that things can be up to the agent in the sense that it is able to settle things. This does not mean however that the agent is free to do whatever it wants. Its ability to act is constrained both by its nature as well as by its environment. According to Steward, "it is utterly undeniable that all animal agency takes place within a framework which constrains, sometimes very tightly, what can be conceived of as a real option for that animal" [8, p. 104]. Within these constraints, though, agents have a lot of flexibility. Even though an animal has to eat within a certain time frame, for example, it can still decide when to eat, what route to take to get there, to eat slow or fast, to grasp its food this way and not that way, etc. The agent thus continuously settles *how* it does the things that it does. More advanced agents like us might have the ability to also settle more of *what* we do, but at the minimum for Steward an entity has to be able to settle how it does things for it to be an agent.

I think that there are two important things to take away from this account.[8] A first thing is that both the nature of the agent and its environment constrain what can be 'up to' it. Even though Steward focuses mostly on constraints that have to do with the particular embodiment of the agent and the natural laws that it and its environment are subject too, I think that a third factor that should be included is personal history. We evolve and learn over time, and each learning trajectory is different. The three factors that influence the kind of things that the agent can settle are thus the kind of agent that it is, its personal history, and its environment.

A second important aspect of agents is that they are *unpredictable* in a sense: even though the 'real options' available to them might be limited because of the above-mentioned factors, it is still the agent that settles what it will do in the end. Given that it is the agent that settles, we cannot exactly predict what it will do. This does not mean that their behavior is *random*, though. The options available to them depend on in part on the kind of agent they are, their personal history and the environment in which they move. So how can the autonomy of an agent be understood if its behavior is both predictable and unpredictable?

I think that Husserl's notion of *indeterminate determinability* [16, p. 283] can be quite illuminating here. Even though Husserl talks about this notion in the context of our Ego and what makes us a person, I find the notion helpful to spell out exactly what kind of unpredictability agents exhibit. Husserl uses the notion of style to indicate the kind of stereotypical behavior that we can expect from other people. Through our lived experiences, we develop certain styles of behavior that each present us with a number of options to behave in particular situations. However, since each of us has our own lived experience, we all have our own personal mixture of styles, we are each our own "individual kind" [16, p. 286].

The fact that we are our own individual kind makes our behavior unpredictable up to an extent:

> One can to a certain extent expect how a man will behave in a given case if one has correctly apperceived him in person, in his style. The expectation is generally not plain and clear; it has its apperceptive horizon of *indeterminate determinability* within an intentional framework that circumscribes it, and it concerns precisely one of the modes of behavior which corresponds to the style [16, p. 283, italics are mine].

The autonomy of an agent can thus be understood as indeterminate determinability: the agent will behave according to stable patterns, but can always diverge from them. These stable patterns depend in part on the agent's personal history, in part on the kind of system it is, and in part on the world that it moves in. The influence of each of these three factors on the action courses that are available to the agent, make each agent into its own 'individual kind'. Given that we all move through the same world and share certain characteristics with each other, we develop stable patterns (or styles of behavior) that resemble those of others. These patterns make it easier for other agents to predict what we will do, but since we are agents, we are also unpredictable in that in the end *the agent* settles what it will do.

In this more narrow understanding of agent, can we say that current LLMs can be seen as agents? I think that arguments can be made both for the affirmative, and the negative. Given that the probability calculations on which the text generation of LLMs is based contain stochastic elements, this

[8] Actually I would argue that there is also a third point, namely that both the agent as well as its environment should contain stochastic elements. Even when it cannot be proven beyond doubt that our world is (in part) indeterministic, we respond to it, and the entities in it, as if it is.

production could be described as indeterminate (stochastic elements) determinability (probability calculations). A first question that can be raised is whether what we are judging to be predictable is our own behavior or that of the LLM. Is it just learning to simulate us as well as possible, or is it learning to generate texts on its own? So, is it the LLM that is predictable or are we?

As a personal history, one could argue that the LLMs gain experience through the training on a particular text-corpus.[9] The question is of course whether this suffices for a 'personal' history, or whether we need something more, like interaction with other agents or interaction with a physical world. A question is as well what kind of system an LLM is. What role does the hardware form in its functioning? In what way does it provide or constrain the options of the LLM? As for the software, is the LLM always only able to choose from the same number of 'options' for the next word? Or does this depend on the context? And, does it always need to choose? It seems that agents, to have the choice in how they respond, need to be able as well to not act — to not generate a text. Can LLMs choose not to respond? Should this be necessary for them to count as agents? Can they settle things on their own without being prompted? And is the choice process (even though complicated) always the same for the LLM? Or will it depend on the interaction that is subjected to?

A final question is what counts as an environment for these LLMs. One could say that their environment consists of the prompts of their users and maybe even any texts that they have access too. If they are connected to the internet, for example, is then the whole of the internet their environment? And do we include the hardware in the environment? Why yes or no? Is it important to limit this environment? What then count as the agent? Only the software, software and hardware, or the environment as well?

These are relevant and important questions, all prompted by adopting a more narrow understanding of agent and using it to study current artificial systems. Given that already this still rather crude notion of agency as indeterminate determinability can raise and guide many interesting questions that we might have, I think it is important that we should make use of a more narrow understanding of the term agent in AI-research.

## IV. CONCLUSION

The successes of LLMs like ChatGPT are very impressive and foreshadow great changes in how we do things in our society. At the same time they also foreshadow a change in our interaction with AI-systems that we are not yet conceptually ready for. Whether we would call such LLMs agents or not influences namely the way we interact with them. If the artificial system is exactly that — a system — then my behavior will, and should, be different towards it than when it is an agent. A tool can break or malfunction and its use should therefore be regulated and learned.

[9]I want to thank my reviewer for pointing this out.

Where we use tools, we *inter*act with agents, and this interaction is shaped by the kind of agent that we interact with. Think for example of a cat, that cannot be held responsible for damaging the couch with its scratches, because we do not expect it to be able to reason why what it did is making us agitated. A twelve-year old child, though, can be held responsible for writing on the wall, because we expect it to be able to understand why this is not okay. Where current LLMs are very impressive, we are not yet sure if they are agents, and if they are, what kind of agents. To make this clearer, we need to develop the notion of 'artificial agent' further. This will not only protect the users by telling them what kind of interaction they can expect, but also provide us with some conceptual tools for Explainable AI and AI-ethics. The concept of 'artificial agent' can do a lot of work, if we give it the attention it deserves.

## REFERENCES


[1] Floridi, L. (2023). Ai as agency without intelligence: On chatgpt, large language models, and other generative models. *Philosophy & Technology*, 36(1):15.

[2] Russo, F. (2022). *Techno-scientific practices: an informational approach*. Rowman & Littlefield.

[3] Sparrow R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1):62–77.

[4] Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology*, 34(3):589–607.

[5] Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, 24(4):1201–1219.

[6] Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, 14:349–379.

[7] Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research*, 79(2):251–278.

[8] Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.

[9] List, C. and Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

[10] Gunkel, D. J. (2012). The machine question. *Critical perspectives on ai, robots, and ethics*, page 5.

[11] Ferrero, L. (2022). *The Routledge Handbook of Philosophy of Agency*. Routledge.

[12] Paul, S. (2020). *Philosophy of action: A contemporary introduction*. Routledge.

[13] Wu, W. (2022). Agency, consciousness and attention. In Ferrero, L., editor, *The Routledge Handbook of Philosophy of Agency*, chapter 18, pages 201–210. Routledge, Taylor & Francis Group, first edition.

[14] Seifrid, M., Pollice, R., Aguilar-Granda, A., Morgan Chan, Z., Hotta, K., Ser, C. T., Vestfrid, J., Wu, T. C., and Aspuru-Guzik, A. (2022). Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, 55(17):2454–2466.

[15] Abolhasani, M. and Kumacheva, E. (2023). The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, pages 1–10.

[16] Husserl, E. (1989). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy: Second book studies in the phenomenology of constitution*, volume 3. Springer Science & Business Media.