Synthese (2023) 201:85
https://doi.org/10.1007/s11229-023-04083-9

ORIGINAL RESEARCH

Introducing a four-fold way to conceptualize artificial agency

Maud van Lier[1]

Received: 29 October 2021 / Accepted: 8 February 2023

Abstract
Recent developments in AI-research suggest that an AI-driven science might not be that far off. The research of for Melnikov et al. (2018) and that of Evans et al. (2018) show that automated systems can already have a distinctive role in the design of experiments and in directing future research. Common practice in many of the papers devoted to the automation of basic research is to refer to these automated systems as 'agents'. What is this attribution of agency based on and to what extent is this an important notion in the broader context of an AI-driven science? In an attempt to answer these questions, this paper proposes a new methodological framework, introduced as the Four-Fold Framework, that can be used to conceptualize artificial agency in basic research. It consists of four modeling strategies, three of which were already identified and used by Sarkia (2021) to conceptualize 'intentional agency'. The novelty of the framework is the inclusion of a fourth strategy, introduced as conceptual modeling, that adds a semantic dimension to the overall conceptualization. The strategy connects to the other strategies by modeling both the actual use of 'artificial agency' in basic research as well as what is meant by it in each of the other three strategies. This enables researchers to bridge the gap between theory and practice by comparing the meaning of artificial agency in both an academic as well as in a practical context.

Keywords Artificial agency · AI-driven science · Intentional agency · Modeling · Basic research

✉ Maud van Lier
maud.van-lier@uni-konstanz.de

[1] Department of Philosophy, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Baden-Württemberg, Germany

Published online: 23 February 2023

1 Introduction

In the roughly 70 years since the field has emerged, artificial intelligence research has attempted to automate a broad range of human activities. One recent domain that AI-research is focusing on is the use of automated systems in basic research. Automated systems are being developed to take over specific processing and data-analysis tasks (Du Sautoy, 2019; Evans et al., 2018, pp. 150–168), while at the same time there is also a more general aim of designing "general purpose-tools for scientific research" (Nautrup et al. 2020, p. 1). In Evans et al. (2018), for example, the researchers show how AlphaFold, a neural network developed for predicting protein shapes, can greatly improve medical research and waste disposal research since it is able to predict the shapes of proteins much faster than human researchers can. Where AlphaFold is meant to take over a specific part of the research process, studies like those of Melnikov et al. (2018), Iten et al. (2020) and Ried et al. (2019) investigate the possibility of automating scientific tasks that are more conceptual in nature. Even though the actual tasks performed by the automated systems in these studies are still rather simple, the success in performing them does suggest that automated systems will eventually be able to perform (almost) all of the tasks typical of a research process—an AI-driven science might not be that far off.

This paper forms an initial contribution to the formulation of the conceptual foundations of such an AI-driven science. Exploring these conceptual foundations does not only show us what an AI-driven science might entail, but it might inform other debates as well. Already the conceptualization of its three most obvious components—'AI', 'driven' and 'science'—raises interesting questions for scientists, action theorists, and philosophers of science alike. For instance, in what way does AI-driven science differ from data-driven science or theory-driven science? And what do we then mean by AI? To what extent can we automate the research process and what skills or abilities are required for this automation? Is the notion of 'science' to be taken as broad or narrow? Depending on whether our notion of science includes disciplines that make use of qualitative methodologies, the specific tasks of, and use for, automated systems might differ greatly. Exploring these conceptual foundations of an AI-driven science will thus not only contribute to current developments in AI-research, but as well to those of other academic fields, by challenging accepted views and pointing out potentially new and interesting research topics.

A first step in this overall attempt to lay the conceptual foundations for an AI-driven science is to choose a methodological framework. The scope of this paper is rather narrow. It focuses on finding a suitable methodological framework for the conceptualization of *one* of the concepts that the author supposes to be foundational for an AI-driven science: that of 'artificial agency' in basic research.[1,2] The notion

[1] This does not, of course, exclude the possibility that the chosen framework will be useful for the conceptualization of the other basic notions as well.

[2] 'Basic research' is here understood as research that is "performed to further scientific knowledge without an obvious or immediate benefit" (Council, 2004 p. 20). A first consideration for this focus on basic research is that this field is at the forefront of developing minimally autonomous automated systems in order to advance their research (see also Sect. 2). A further consideration for this focus is to avoid (for now) the more far-reaching discussions about the moral applications of artificial agency.

of 'agent' is already in use in basic research interested in the automation of science and generally refers to the automated systems studied there (e.g. Iten et al., 2020; Melnikov et al., 2018; Nautrup et al., 2020; Scholkopf et al., 2021; Wu & Tegmark, 2019). However, at the moment, one could argue that these 'agents' have little in common with the kind of paradigmatic entities that philosophers usually attribute agency to—humans.

Roughly put, human agents seem to be capable of a certain form of self-movement: agency. Agency as a capacity manifests itself when the agent acts, where acting is the ability of an entity to engage directly with its surroundings in such a way that this engagement originates with the agent and can only indirectly be attributed to external factors. Even though the exact nature of agency has remained controversial up to today, many philosophers have related agency to other capacities that often accompany it, such as setting goals (Bratman, 1987; Ferrero, 2022) , mental causation (Davidson, 1980) , intentional reasoning (Anscombe, 2000) , the ability to settle things (Steward, 2012; Aguilar & Buckareff, 2022) , or a combination of these. Up to now, almost no one has been prepared to attribute any of these abilities to current AI-systems. The question can thus be raised why artificial agency is likely to form a foundational concept of an AI-driven science?

Even though the use of the term 'agent' for AI-systems in basic research seems currently to be no more than a conventional expression, scientists are working hard to develop AI-systems that are more worthy of that label (Sect. 2.1). In quantum computing, for example, studies have shown the advantages of using different types of machine learning in quantum error correction (Nautrup et al., 2020; Convy et al., 2022) . Both of the studies emphasize the 'active' component of these AI-systems. Not only are the AI-systems able to detect possible errors, but they also *correct* the errors as soon as they find them. These AI-systems are thus able to autonomously decide to intervene on the qubits they analyze and to execute this intervention accordingly. Even though the systems have only yet been tested in simulated environments, the success of these tests suggests that the implementation of such systems in laboratory settings might soon become reality. The aforementioned systems display a couple of features that we usually associate with agents: the ability to intervene, to do this autonomously, and to base this intervention on their own decisions.[3] Even so, one could still wonder what the added value is of referring to these systems as agents (in the philosophical sense), when one can just state that they are systems that display agential features. In Sect. 2.2, I will give a number of arguments for the use of this term, emphasizing especially its explanatory power, the fact that it still allows for a lot of adjustment in meaning, and its connection to the vast literature in philosophy. What is more, I argue that any objections against the use of the term should not be seen as defeating, but rather should be taken as standards for any proposed conceptualization.

For this conceptualization, though, we first need to adopt a methodological framework. Sarkia's (2021) *neo-Gricean framework* is the most recent, pragmatic and pluralistic framework that has been proposed for the conceptualization of a paradigmatic form of agency: intentional agency. It combines three modeling strategies—Gricean modeling, analogical modeling and theoretical modeling—that

[3] I leave aside for now that these decision-options might already be in some sense hard-coded in the systems.

have all been used in the philosophy of mind and action to conceptualize intentional agency. According to the framework, researchers should not give precedence to one of these strategies over the others, but rather model complex concepts such as 'intentional agency' in a pluralistic and pragmatic way. In Sect. 3, I discuss the merits of this framework for a conceptualization of artificial agency and then proceed to discuss how each of Sarkia's modeling strategies can be implemented to conceptualize artificial agency and what we can gain from each of them (Sects. 3.1, 3.2, 3.3). However, rather than arguing for the one-to-one adoption of the neo-Gricean framework for the conceptualization of artificial agency, I will argue in Sect. 3.4 that we can gain even more understanding of what we mean by artificial agency when we add a fourth modeling strategy to the other three.

This fourth strategy I will introduce as *conceptual modeling*, and it is meant to add a semantic dimension to the overall conceptualization. After setting out what this kind of modeling entails in Sect. 4.1, I will show in Sect. 4.2 how this modeling strategy can be implemented to conceptualize artificial agency and how the strategy complements the other three strategies. Finally, in Sect. 4.3, I introduce *the Four-Fold Framework*, a framework that includes all four modeling strategies, and argue that we should adopt this framework if we want to obtain a broad and comprehensive account of what we might mean by artificial agency in an AI-driven science.

2 Why artificial 'agency'?

In the introduction, I claimed that some of the current AI-systems in quantum error correction display agential features such as the ability to intervene, to do this autonomously, and to base this intervention on their own decisions. What make these features agential? And what is the added value of calling these systems 'agents', rather than systems capable of displaying these features? In light of these questions, I will defend two claims in this section. After comparing a number of technologies currently in use in basic research, I claim in Sect. 2.1 that some state-of-the-art AI-systems are increasingly taking on a more active role in this field of research. Subsequently, in Sect. 2.2, I discuss a number of arguments for and against attributing the term 'agent'—in the philosophical sense - to these active AI-systems. I conclude by claiming that the term 'agent' has the most explanatory power to describe the particular role that we want these systems to play in basic research. Furthermore, rather than speaking against the use of the term 'agent', I argue that the objections to the use of the term show the need for, and focus points of, a conceptualization of artificial agency in basic research.

2.1 The role of AI-systems in basic research

By now, it is generally accepted that technologies[4] play an important role in the production of knowledge in scientific research. What kind of role, though, and to what extent

[4] Here I understand technologies in the broadest sense possible, to denote any kind of artificial (in the sense of 'human-made') instrument in use by scientists.

these technologies can influence knowledge production remains a topic of debate. Some of the rather new fields, like Technoscience and Science and Technology Studies, have focused on the overall contribution of technologies to knowledge production, arguing that this contribution has been more foundational than has been assumed in more conventional views (Boon, 2011; Bensaude & Loeve, 2018) . However, the contribution of these technologies to knowledge production has been, albeit foundational, rather passive: technologies are still directly guided in what they do, and to what extent they do so, by scientists. With the growing integration of AI-systems in basic research, this is slowly changing. In this section, I will first show how AI-systems are starting to take over the guiding role of scientists in routine scientific practices. Subsequently, I will argue that the most recent advances in quantum error correction suggest that AI-systems will increasingly be able to take over other roles of scientists as well. These two developments indicate already that AI-technologies are starting to play a more active role in knowledge production. Given the fast pace in the developments in AI-research, it is likely that the level of activity and versatility of these systems will only expand further.

First, though, I have to make clear what I mean by a technology playing a passive or active role in science. To do so, I will take an experiment as a classical example of a scientific research practice. Following Beisbart (2018, p. 176), one could say that:

> "A person A is experimenting on system X only if
> 1. A intervenes on X;
> 2. A takes observations about X; and
> 3. A does 1 and 2 with the superordinate aim to obtain information about the way X behaves and reacts to the intervention."

Let us see person A as a (human) scientist. A technology now plays only a *passive* role when the scientist with the help of this technology intervenes in (condition 1), or observes (condition 2), a particular (controlled) environment. I will further hold that a technology plays an *active* role in scientific research when it can, by itself, intervene (condition 1) and observe (condition 2) in such an environment.[5]

The passive use of technologies ranges from more classic examples like the use of a microscope or a thermometer for observation or the use of a centrifuge or robot arm for intervention, to fully automated experiments where an automated system performs the entire experiment (see for example Renner et al. 2020; Majdpour et al. 2021). The automation of laboratory devices and experiments is not a new phenomenon. Already around the 1950s, the first automated devices for laboratories were developed (Olsen, 2012) . Automating devices and scientific procedures like experiments brings with it many advantages,[6] but such automated systems still only passively contribute to scientific research as it is the scientist that decides the parameters of the experiment, how the experiment should be executed, and what conclusions to draw from it.

[5] I think that an AI-system can play an active role in science, even when its overall role in the research is instrumental. This is why I focus only on the first two conditions, holding that a system is able to play an active role when it is able to meet the first *and* the second condition without human intervention.

[6] Besides speeding up procedures that would take up a lot of time when done manually, automating devices and experiments has further advantages like allowing for a more precise repetition of the procedure and the protection of scientists working with dangerous materials.

In the last few years, great successes have been booked with self-driving laboratories (see Melnikov et al. 2018; Du et al. 2021; Erps et al. 2021; Rooney et al. 2022; Gongora et al. 2020; Bennett and Abolhasani 2022). These laboratories differ from automated experimental set-ups in that they form a closed-loop: they can often run by themselves for days—if not weeks—without any human intervention (Soldatov et al., 2021, p. 619). This is because a self-driving laboratory usually consists of both a robotic platform that performs experiments and a machine learning algorithm, such as Bayesian optimization, that optimizes and designs the ensuing experiments that the robot should perform. A scientist thus no longer has to plan, optimize or design new experiments—this is done by the AI-system itself (condition 2). At the moment, these autonomous laboratories are mainly used to perform routine operations, but these operations are performed faster, continuously, and with less human error (see Soldatov et al., 2021).[7] Even though the autonomous laboratories and experiments, as a configuration of technologies, already perform a more active role in research, and new and usable materials and experimental set-ups have been discovered because of them, the AI-systems that are part of this configuration do not actively take part in the experiments. In line with our definition of an experiment, these systems would at most fulfill condition 2.

Recent developments in quantum error correction suggests that it will now not be long before such AI-systems can play an active role in real-life experiments. In quantum computing, it is of the utmost importance that one can optimally control the experimental setting since any disturbance in the form of noise can destroy the coherence of the qubits. To avoid this disturbance, quantum scientists implement quantum error correction (QEC) methods. This entails checking each data qubit separately for noise and correcting the errors where necessary. Outsourcing this task to AI-systems like reinforcement learning systems and neural networks brings with it many advantages. In continuous quantum error correction (CQEC), for example, researchers have mostly focused on signal measurements that behave in an idealized manner. However, as Convy et al. (2022) point out, "in real dispersive readout signals we observe a wide variety of 'imperfections' caused by hardware limitations and post-processing effects, which can lead to more complicated syndrome dynamics or significant alterations to the noise distribution" (p. 2). One of the difficulties of designing a well-calibrated CQEC protocol is "to generate a precise mathematical description of the imperfections present in real measurement signals" (p. 2). Convy et al. now propose to let a recurrent neural network function as a CQEC algorithm and train it on non-ideal signals. This network should be able as well to correct *actively*, meaning correcting the errors "during the experiment as soon as they are observed" (p. 2). Here then, the CQEC algorithm is meant to take an active part in the experiment, as it observes and intervenes by itself (conditions 1 and 2).

Convy et al. are not the only ones who want their system to take an active stance. In Nautrup et al. (2020), "a reinforcement learning (RL) framework for adapting and optimizing QEC codes" is presented, that should be able to change "the structure of

[7] What is more, because an optimization algorithm like Bayesian optimization can learn from the data that is produced by the experiment, it can with only small data-sets improve on the experimental set-up rapidly—something that is very useful in materials science research where experiments are often complex and for which only little data is available (Pruksawan et al., 2019; Gongora et al., 2021; Coley et al., 2019).

the quantum mechanical system itself, (...) both a priori and during a computation" (p. 2). It is the AI-system itself, then, that is supposed to adapt the computation based on its own judgment of whether this adaptation is necessary. Here, again, both condition 1 and 2 are met. Even though both of these studies have only been proven successful in simulations, they do suggest that soon similar AI-systems might be able to work in real-life labs successfully. Such technologies thus have the potential to take an active part in science and are developed with the intention to do so.

2.2 Pros and cons of using the term 'agent'

What should such AI-systems be called? The self-driving laboratories are, *as a configuration of technologies*, able to do some of the routine tasks that are usually performed by scientists themselves. And in simulations, the AI-systems currently under development in quantum error correction can learn to detect noise when they come across it by themselves, recognize it, and correct it. Both of these kind of systems are supposed to take on a more active role in scientific research, but does that mean as well that we should start calling them agents?

As stated in the introduction, the paradigm case for an agent has been the 'human being' and the capacity of agency has been associated with such characteristics as 'decision making', 'intentionality', 'autonomy', but also with 'cognition', 'mental states', and 'morality' (Ferrero, 2022, pp. 7–8). Especially based on these latter characteristics, we are probably not willing to attribute agency to the just-discussed AI-systems. However, over the last few decades, an increasing number of scholars have argued that the attribution of agency should not be limited to humans alone, since it can be attributed as well to most animals and maybe even to artificial entities like groups, legal persons and robots. This meant, though, that a more simplified notion of agency was needed that could be attributed to non-anthropomorphic agents too. New characteristics associated with agency have been offered as a result, such as adaptivity, flexibility, autonomy, the ability to settle, autopoiesis, etc (Müller & Briegel, 2018; Steward, 2012; Aguilar & Buckareff, 2022; Burge, 2009; Arnellos & Moreno, 2015) . At least some of these characteristics we can attribute to the above mentioned systems. Still, why would we be willing to do so? What is the added value of calling these active systems agents? And what are some reasons not to do so?

Let us start with some of the arguments for attributing the term 'agent' to these active systems. First, the term has a lot of explanatory power. What I mean by this is that it can better describe what kind of systems scientists are trying to develop than terms like 'instrument' or 'technology' can. With current non-autonomous technologies we are starting to reach the limits of what we, as humans, can achieve. To get further, to understand the world better, and to find solutions to dangerous threats like climate change, we need to transcend what we can do (see for example Schneider et al. 2021). For this, we need systems that can take on an active role in science, that can work alongside scientists and make contributions of their own. Such systems have to be able to learn for themselves, to judge by themselves, and to act by themselves. Even when the systems at the moment do not live up to this aim yet, using the term 'agent' shows where we aim to go. In line with this is the fact that the term still allows for a

lot of adjustment in meaning. Agency seems to be something that comes in degrees. As was just discussed, agency has been attributed to a variety of entities and has been associated with a number of capacities. Where we would now be dealing with a rather basic artificial agent, this does not mean that we are only limited to this particular understanding of agent. The advancements in basic research have only just started.

A further advantage is that there is a vast literature on the notion of agency. By using the term 'agent', we can make use of this literature to hone in on what it is that we exactly mean with the term in a basic research context, where it clashes with other uses of the term, and where we can find similarities with other agents. Related to this is the fact that the term is used in many different disciplines—not only in philosophy. Linking and interacting with these other fields might open connections that we were not aware of yet or which can provide new insights, on a mutual basis. A final point of interest is our social interactions with these AI-systems. We interact differently with the world around us based on the kind of entities we encounter. Referring to active AI-systems as agents rather than as technologies changes our interaction with them, and based on the kind of agency we attribute to them, this interaction and especially what we expect from this interaction changes again. Just as we do not blame dogs for what they do but still put them on a leash since we expect them to act in certain situations in particular ways, so does attributing agency to AI-systems integrated in society at large allow us to set certain expectations and standards for our interaction with them.

As one can see, there are many arguments for the use of the term agent. Even so, we can name a number of objections as well. First of all, one can either fundamentally disagree with the view that agency can be attributed to a being that is not alive, or one can even think that there is no such a phenomenon as agency. The latter viewpoint is a matter of opinion and stops all discussion. So I will let it be for what it is. However, I think the former point is interesting when rephrased as a point of attention: we often see agency as something more than just the ability to 'intervene' or to 'act'. Agency is often equated with 'true autonomy', whatever that might be. We have to consider whether our conceptualization can and should reflect this. A second objection can be that current AI-systems are still very far removed from acquiring any of the capacities often associated with human 'agents', like those of conscious reflection and reasoning. If we want to develop AI-systems that can perform more advanced tasks in science, would such capacities as conscious reflection and reasoning be necessary? Or do we require something else from them? Lastly, one could argue that agents need to have some physicality to act, and that that is hard to align with current AI-systems. How do they interact with their environment? And could they even do so? Do we require them to be embodied? And then what kind of embodiment? I think that all of these objections are relevant and require an answer. However, rather than forming fundamental reasons not to use the term 'agent' to refer to active AI-systems, I would say that they set the standards for any conceptualization of artificial agency in basic research that we come up with. Either we can produce a conceptualization that makes our form of embodiment, cognition, and aliveness obsolete for 'artificial agency', or we have to set a higher standard for the conditions under which we accept artificial systems as 'agents'. For this, though, we first need a conceptualization of artificial agency. We

thus return to the main question of this paper: what kind of methodological framework should we adopt for this conceptualization?

3 The neo-Gricean framework

In the search for a methodological framework, it seems only natural to first take stock of what kinds of attempts have already been made to conceptualize (forms of) agency. Sarkia (2021)'s *neo-Gricean framework*[8] forms the most recent, pragmatic and pluralistic framework that has been proposed for the conceptualization of a paradigmatic form of agency: intentional agency. It combines three "contrasting methodological strategies for modeling intentional agency in contemporary analytic philosophy of mind and action" (§1): Gricean modeling, analogical modeling and theoretical modeling. In the framework, these contrasting methodological strategies are interpreted as alternative modeling strategies that complement and inform each other during conceptualization. Sarkia places himself in the tradition of naturalistic philosophers of science like Giere (1988), Godfrey-Smith (2006) and Weisberg (2012), and holds, just like them, that a model is 'an indirect representation of the world'. In this paper, however, I will subscribe to a broader notion of a model in the tradition of Bailer-Jones (2009), understanding a model to be 'an interpretative description of a concept, entity, or phenomenon that facilitates access to the concept, entity or phenomenon, where this access can be perceptual as well as intellectual'.[9]

Making use of this pluralistic and pragmatic framework to conceptualize artificial agency has three benefits. The first benefit is simple: Sarkia has already shown that the framework is useful for the conceptualization of another kind of agency. Another point is that, the notion of artificial agency is relatively novel and still fairly controversial. Limiting ourselves to the implementation of just one strategy might exclude potentially fruitful approaches. Sarkia's framework combines three methodological strategies that have been used to model intentional agency in the philosophy of mind and action. By adopting his framework, we avoid excluding any of these approaches. Lastly, (artificial) agency is a rather complex phenomenon that is not easily definable since we cannot empirically observe it in a direct way. Sarkia (2021), inspired by scientific modeling, points out that "the indirect nature of model-based science makes it possible for scientists to study phenomena, which could not easily be studied empirically because of causal complexity, epistemic opacity or rarity" (§1). Even though I am interested in model-based philosophical analysis rather than model-based science, I do find the argument convincing that model-based strategies, because of the mediating nature of models, might make it easier for us to study such a complex phenomenon like (artificial) agency. Given these three benefits, I will describe in the following each

[8] The reason that Sarkia (2021) gives for calling it a neo-Gricean framework is simply that "Paul Grice (1974–1975) was arguably the first contemporary philosopher to suggest a model-based approach to the philosophy of mind and action" (§2).

[9] Bailer-Jones (2009)'s definition applies to scientific models that describe phenomena. I am not just interested in scientific models, however, but models of any kind. This is why I have broadened her definition of a model (pp. 1–2) so as to include concepts and entities as well. Other than that it is the same definition.

of the three modeling strategies in more detail, and make some suggestions as to how they can be implemented to conceptualize artificial agency.

3.1 Gricean modeling and its implementation

The first modeling strategy that Sarkia (2021) identifies is Gricean modeling, which he describes as a "'bottom-up' construction of the phenomenon of intentional agency from its constituent, primitive parts" (§1). The entity or phenomenon that is being modeled is thus 'reconstructed' as it were from its constitutive parts. Sarkia (2021) links this kind of modeling to a certain type of knowledge, called *engineer's knowledge*, which can be gained by "opening the 'black box' of the mind to investigate how each primitive building block of intentional agency contributes to the goal-directed behaviors of an agential system" (§5).

This modeling approach is based on Grice (1974)'s account of how a psychological theory should be developed. In his account, he proposes to model imaginary types of creatures (pirots) after actual ones so as to understand the psychological make-up of these actual creatures better. The idea is that each pirot is exposed to an increasingly complex living condition, for which it requires more (complex) psychological capabilities to continue to survive. The researcher hypothesizes what these additional capabilities might be and is thereby limited by two provisions. For each living condition, we can endow the pirot only with those psychological capabilities that (1) are manifested by some behavior and, (2) are minimally necessary for the pirot to have the optimal chance of surviving in that particular living condition (see Grice 1974, pp. 36–40). Even though Grice himself focused mainly on psychological capabilities, I believe that the strength of Gricean modeling lies in that it allows the researcher to step-wise posit a number of related capabilities that the entity could possess that would explain why the entity can behave as it does in the end, where these capabilities do not necessarily need to be psychological.

There are several advantages to the use of this kind of modeling for the conceptualization of (artificial) agency. First, it allows the researcher to assume only a *gradual* difference between less and more advanced creatures. As stated at the end of Sect. 2.2, I see agency as a capacity that comes in degrees—agents often display several agential features to a higher or lesser degree, or display only some and not other features. Gricean modeling allows the researcher to capture these differences. Another advantage is that because only the minimally necessary is added to each new living condition and both the pirot and the living condition need to resemble actual physical reality, the researcher will not assume more agential features than there actually are, than might be physically possible or that can be perceived. Finally, Gricean modeling lends itself well for computer simulation. So, one can simulate the different living conditions and let agents with various combinations of capacities 'live' in these conditions for a while, to see what combinations of capacities provide the agent with the greatest chance of survival.

How could this method be implemented to study artificial agency? I think there are two ways to go here. First, we can model artificial agency in a way where we start from very simplistic forms of artificial agency—so those automated systems that are only

able to perform actions like moving an object from one place to another - and then build up to ever more advanced versions until we reach the kind of agent that would be able to carry out scientific tasks like performing an experiment.[10] Nyholm (2018) has done something similar in his functional approach towards artificial agency. His attribution of agency is based on the level of functionality of an automated system. In his paper, he makes a distinction between four levels of agency: domain-specific basic agency, domain-specific principled agency, domain-specific supervised and deferential principled agency, and domain-specific responsible agency (pp. 1207–1208). At each new level the agent can perform the same functions as the agents at previous levels, plus one more thing.

As in any kind of modeling, the goals and motivations of the researcher influence what kind of model is being built. In the case of Nyholm's model, his interests are in tackling the responsibility question in a context where a human is harmed or killed by an automated system. So the 'end-point' in his model is a morally responsible artificial agent. In the context of this paper, our 'end-point' would be an artificial agent capable of carrying out scientific tasks. The levels of agency that we construct or are interested in will thus slightly diverge from those of Nyholm. To give an example: in order for an entity to be able to bear responsibility for its actions, it has to have a certain degree of control over its actions. This is why in Nyholm's model, each difference between the levels of agency centers around the degree of supervision that the artificial agent is exposed to.[11] I take it that if we were to construct such a model for an artificial agent capable of carrying out scientific tasks, other qualities would be the main focus, like 'reasoning', 'creativity', or 'communication'.

Rather than modeling the artificial agent, we could also decide to model 'the scientist' to find out at what level current AI-systems would fit the model. So we could construct a model of ever more advanced scientists. Here I will make some suggestions as to what the outline of such a model would look like. Just as with the functional model of Nyholm, we first have to establish what our 'end-point' is. Based on this decision, it might also become clear what we see as influential factors on the level of skill of the scientist.[12] For now, I will hold that the highest level of the model is where the scientist is able to make valid contributions to the scientific field that s/he is working in. I assume that a contribution is only valid when it is accepted as such by

[10] Here I leave open whether the model is based on actual systems that have been built or whether it is based on hypothetical systems.

[11] As an example, look at the final two levels of agency that Nyholm (2018) proposes (italics are mine): Domain-specific supervised and deferential principled agency: pursuing a goal on the basis of representations in a way that is regulated by certain rules or principles, *while being supervised by some authority who can stop us* [*sic*] *or to whom control can be ceded, at least within certain limited domains* (p. 1208). Domain-specific responsible agency: pursuing goals in a way that is sensitive to representations of the environment and regulated by certain rules/principles for what to do/not to do (within certain limited domains), while having the ability to understand criticism of one's agency, *along with the ability to defend or alter one's actions based on one's principles or principled criticism of one's agency* (p. 1208). The difference between the two levels lies in the degree of control that the agent has over its actions. So in domain-specific supervised and deferential principled agency, the agent is still supervised and can be stopped or lose the control over its actions if the supervisor deems it unfit. Only at the last level would Nyholm say that the agent is responsible. At this level there is no longer someone who can intervene in the behavior of the agent. Nyholm thus links control over one's own actions to responsibility.

[12] Here I want to note that a model can thus differ depending on the kind of end-point that is chosen.

the scientific community that the scientist is part of. Based on this end-level, we can now reconstruct the previous ones:

Level 1: The Juvenile Scientist A scientist that has no background knowledge yet and which sort of 'stumbles around', learning the structure and regularities of its environment.
Level 2: The Hypothesizing Scientist A scientist that is still working on its own, but that has already built up a considerable body of hypotheses concerning the world around it.
Level 3: The Collaborating Scientist A scientist that communicates with other scientists and who compares their work to its own. The body of knowledge that the scientist has is build up out of its own observations and its acquired knowledge about the work of others.
Level 4: The Contributing Scientist A scientist that is able to make valid contributions to the scientific field that it is working in.

As we can see, there are three variables in this model. A first determining factor is whether or not the scientist is part of a community, sharing knowledge with them (levels 1 and 2 vs. levels 3 and 4). Secondly, there is the determining factor of whether the scientist contributes itself or not (levels 1 and 3 vs. levels 2 and 4). Lastly, there is a difference in whether the contribution is accepted by others or not (level 2 vs. level 4). This initial outline is of course still very basic and only one example of how such a model can be constructed. More elaborate models might be specifically of use to particular fields of research and might include additional levels. A further point of difference might be the end-point that is chosen. One could imagine an end-level where the scientist can make scientific discoveries. This would then change the levels based on what one thinks that a scientific discovery entails.

Now that we have constructed this model, a following step can be to see whether current research on the automation of science can be fitted into the model. I would say that most current research is hovering between level 1 and level 2. The AI-system in Ried et al. (2019), for example, is a simple learning agent, that "autonomously identifies abstract variables in the process of learning about its environment" (pp. 1–2). The agent learns new concepts by moving through its environment. This would be something for a scientist at the first level. However, it also learns to make correct predictions with regard to the experiments that it performs. This would place the agent in level 2. It does not communicate with other agents yet, so level 3 and 4 are not relevant here.[13] In the research of Nautrup et al. (2020), the different neural networks 'communicate' with each other based on a principle of 'most optimal communication' regarding the observations that are made and the question that each agent needs to answer. This would be a (very simplistic) example of communicating agents and the

[13] One can still wonder whether it is the AI-system itself that 'discovers' the hidden variables or that the researchers who interpret the output of the system do so. In future research it would be interesting to analyze what we mean by concepts like 'experimenting', 'designing' and 'formulating' when attributing them to an automated system compared to an autonomous system. For now, I will assume that the automated systems can perform actions that are similar enough to 'experimenting', 'formulating' etc. that we recognize their actions as instances of these scientific tasks.

research of Nautrup et al. (2020) could thus be a prototype for what kind of automated systems belong to level 3.

Research like that of Ried et al. (2019); Nautrup et al. (2020); Melnikov et al. (2018); Wu and Tegmark (2019) and Iten et al. (2020) all focus on the automation of specific parts of the research process. Sparkes et al. (2010) attempt instead to design a robot that can carry out the entire research process: "forming hypotheses, devising and carrying out experiments to test those hypotheses, interpreting the results and repeating the cycle until new knowledge is found" (p. 73). The question is how autonomous their robot is, however, since it still needs a lot of intervention from technicians and it is very specialized. What is more, the researchers understand by science 'the formulating and testing of hypotheses through the performance of physical experiments'. As soon as one has a more inclusive understanding of science, by for example including the social sciences, then these kind of experiments would no longer be sufficient to test the hypotheses.[14] A further interesting question that these robot scientists raise is what we understand by reasoning and background knowledge. In one of Sparkes et al. (2010)'s prototypes, the one named 'Adam', a lot of facts have been programmed into the software. Sparkes et al. raise themselves the question whether this would count as 'knowledge' or 'information'. One of the benefits of trying to fit current research into the model is thus that model constructors become aware of what parts of the model need to be specified more.

There are also broader benefits to implementing this Gricean modeling method to the conceptualization of artificial agency in a basic research context. Implementing the modeling method can provide researchers with some very interesting insights in multiple domains. It can direct future research by showing researchers the level that they are at now as well as by showing them what features they should focus on to reach the next level. At the same time, current research can influence the construction of the model too, as we saw just now. Besides this insight in what parts of the model need to be specified more, the research objectives and motivations of these researchers could influence the specific 'end-point' that is chosen for the Gricean model, which subsequently has an effect on how the other levels are formulated.

3.2 Analogical modeling and its implementation

The second strategy that Sarkia mentions is analogical modeling, where "a 'horizontal' shift" is made "between two or more categories of presumptive agents" (§1). The idea here is that a comparison is made between two manifestations of agency, of which there is one that the researcher is more acquainted with. The modeling follows the formula "(A is to B like C is to D)", where the relation between A and B is better known, but similar in some way to the relation between C and D (§3).[15] A great advantage of this kind of modeling is that the researcher can work from a particular framework that

[14] Such a more inclusive view of scientific practice necessarily puts into question the standard definitions of 'model', 'experiment', and 'objectivity' in the natural sciences. Whether the robot thus qualifies as a 'scientist' depends largely on one's notion of science. See also Russo (2017) and Montuschi (2014).

[15] There exists a rich literature on what principles or reasoning process an analogy must pertain to. Gentner (1983), for example, lists preservation of relationships, the systematicity principle and one-to-one correspondence as the inferential principles that analogical reasoning is based on. However, scholars like Dunbar

s/he is already familiar with. This kind of modeling generates, according to Sarkia, *acquaintance knowledge*, which is knowledge of "familiar kinds of intentional agents and their possible structural similarities with other agential types" (§1).

Once a great analogy is found, analogical modeling can be a very fruitful endeavor and inspire a broad range of fields. Think for example of famous analogies like seeing nature as a mechanistic system, or likening evolution to a tree. In each of these cases, the analog forms a source of inspiration, given that it invites the researcher to think about, and provide reasons for, the (dis)similarities between the analog and the target. Here it is important to notice that the researcher can learn a lot about the *differences* between the analog and the target as well. However, first there needs to be enough of a resemblance, so that the dissimilarities can be properly understood within the broader framework.[16]

In this section, I want to propose three potentially fruitful analogies for understanding artificial agency in the context of basic science. The first analogy is an analogy between artificial agency and other forms of agency, where we consider the scientist as a human agent and then compare it to the 'agency' of the active AI-system.[17] As the literature that I refer to in footnote 15 states, the kind of analogy depends for a large part on the motivations of the researcher and the context in which s/he writes. My specific motivation here is to see at what point we would perceive of the autonomous system as an actual member of the research team. The context that I work in is basic research. The analogy that I would like to draw is thus one between a scientist and an active AI-system, where I see the scientist as an agent that can perform actions like carrying out scientific tasks. The question then becomes what qualities we would normally attribute to a non-artificial scientist. What comes to mind are abilities like reasoning, creative thinking, and maybe even imagination. In making this analogy, we would not only make connections between the automation of science and disciplines in psychology, epistemology, and philosophy of science, but we would also be forced to revisit our own notion of 'scientist'. Constructing this analogy would thus require interdisciplinary research that would benefit all disciplines involved by providing them with new insights. These insights could then afterwards direct future research in these disciplines.

A second form of analogy that is interesting to think about is one between active AI-systems and technological artifacts/tools/instruments. The field of Technoscience

(2001), Holyoak and Thagard (1997), Markman and Gentner (1993), and Spellman and Holyoak (1992) also consider the goals of the researcher and the context as important influential features of analogical reasoning. Another interesting angle is to see analogical reasoning as proceeding in stages (e.g. Gentner and Maravilla 2018; Gentner and Smith 2012; Holyoak 2012).

[16] An example might be helpful here. In the analogy between a clockwork and nature, one difference that we perceive is that we experience some form of (mental) freedom. In mechanics, however, everything is determined—there is no room for such freedom. This analogy gives direction to new research, in that it challenges researchers to find ways in which this feeling of freedom and determinism can be combined (e.g. Müller and Briegel 2018; Steward 2012).

[17] As stated in the introduction, my interests are in conceptualizing an artificial agent in a research context. The form of agency required for this is of a fairly high level involving reasoning and creative thinking. Drawing an analogy between artificial agency and for example animal agency, which is a kind of agency not capable of this kind of reasoning, is therefore not relevant here.

draws awareness to the fact that technological tools or instruments often play a fundamental role in a scientific knowledge production process (Boon, 2011; Lacey, 2012) . Hacking (1983), for example, mentions the way in which the microscope is essential to the investigation of microscopic organisms. We can only see them with the microscope, never without. Even though these technological artefacts have this fundamental role, they are still 'instruments': agency is only attributed to the researcher that makes use of the instrument.[18] The role of active AI-systems in basic research is still instrumental in nature.[19] However, more than with technological artefacts, there is the expectation with AI-systems that they can play an active role in science, even when still instrumental. What is this difference based on? Drawing an analogy between technological artefacts and active AI-systems might show us this difference. The analogy is further helpful because the agency of an entity is almost always compared to the agency of other kinds of entities. To compare artificial agency with the non-agency of the technological artifacts might provide us with some new insights.

A final interesting analogy would be between the artificial agency of states and corporations and that of active AI-systems. States and corporations have a kind of independent power that goes beyond the individual decisions and actions of its members. Rather, it seems to emerge from the collective actions of a group of individuals. List and Pettit (2011) make an interesting case for the existence of group agency. Leaving aside here whether I agree with List and Pettit, it would be interesting to look for any similarities between the agency of such groups and for example the way that autonomous laboratories function. If we think back of our discussion in Sect. 2.1, then these laboratories, as a configuration of technologies, can perform the same task as a scientist. Does such a laboratory exhibit something like group agency? And if it does not, but we do attribute agency to states and corporations, what is this distinction then based on?

Each of these three analogies can, I think, improve our understanding of what we mean by artificial agency. The strength of analogical modeling is that it allows us to focus on concrete and specific features that are shared between the analog and its target. What is more, it provides the researcher with the freedom to draw on a wide variety of analogies, as long as there are enough points of resemblance between the analog and its target. This means as well that this strategy does not limit the researcher to just one academic discipline. Already we saw that we can look at technoscience, psychology and political science for potential analogies. Analogical modeling thus differs from Gricean modeling in its focus and the knowledge that it provides. Even so both can give us interesting insights about artificial agency, so there is no reason to choose one over the other in our conceptualization.

[18] There are some theories on collaborative agency (e.g. Nyholm 2018) and performative agency (e.g. Young 2021), as well as accounts on the way that artefacts influence our way of acting (e.g. Verbeek 2005, pp. 147–172), but in all these cases it is still the human that is the agent able to act.

[19] Take for example the way in which AI-systems are now used in medical research (Richens et al., 2020) or algorithms like AlphaFold which are used to make the process of predicting the shape of proteins go faster (Senior et al., 2020) .

3.3 Theoretical modeling and its implementation

The final modeling strategy that Sarkia (2021) mentions is theoretical modeling, which generates "*theoretician's* knowledge of abstract theoretical generalizations that are applicable to an open-ended domain of agential phenomena" (§1). Theoretical modeling is a top-down approach that involves:

> reasoning about the laws and regularities that are associated with a particular domain of phenomena without detailed reference to either particular entities that populate that domain (as in analogical modeling) or particular mechanisms that maintain those laws and regularities (as in Gricean modeling). (§4)

Thus, the focus lies on the laws and regularities that a particular group or domain of phenomena is subject to, without referring to specific descriptions of the concrete phenomena that fall under those laws and regularities and without referring to the particular realizations of them. An advantage of this kind of modeling is that it is rather abstract—one could potentially use the same model for a broad range of different phenomena as long as they belong to the same domain that the model represents.

According to Sarkia, one example of theoretical modeling in the philosophy of mind and action is common sense or 'analytical' functionalism. This approach can be described as theoretical modeling, because it "involves *indirect* representation of agents through general law-like regularities connecting mental states to behavior" (§4). This kind of functionalism was initially meant as a response to the Psycho-Physical Identity Theory (see e.g. Place 1956; Feigl 1958; Smart 1959; Hill 1991; Polger 2011), which holds that mental states correspond to particular physical brain states. In functionalism, these mental states instead correspond to the causal role that they play in the cognitive system that they are part of, which enables researchers to abstract away from particular brain states. By focusing on this causal role that they play in a cognitive system, mental states become multiple realizable, meaning that one particular mental state can be realized by different (brain) states as long as the mental state still plays the same causal role in the cognitive system. This multiple realizability is what makes functionalism in the philosophy of mind and action interesting for accounts of artificial agency as well.

Artificial agents are 'artificial' in the sense of being created and up to now 'mechanical' in that they are not build from living tissue. If we would ever claim that AI-systems have some form of mental life, then this would in all probability not be because they have the exact same brain state as ours. To account for any mental life similar to ours, we would therefore need the condition of multiple realizability. Another argument for functionalism is that even in the case of humans, there is the epistemological problem of how to ever *know (for certain)* that someone else is in a particular mental state (or even that others have minds), given that we have only indirect access to the thoughts/sensations of others.[20] Given this controversial nature of mental states, it seems sensible to avoid trying to attribute mental states to automated systems and to instead aim at making claims like 'if an automated system functions in this particular

[20] See for example the debate on whether there are criteria for attributing mental states to others and what such criteria might be (e.g. McDowell 1982; Wright 1984; Albritton 1959; Witherspoon 2011).

way, then it functions in a way that we ordinarily associate with someone being in a particular frame of mind'.

Functionalism in the philosophy of mind and action forms one instance of how we can implement theoretical modeling to the conceptualization of artificial agency. Can we think of more examples? Let us first return to how Sarkia (2021) described theoretical modeling: "theoretical modeling involves reasoning about the laws and regularities that are associated with a particular domain of phenomena" (§4). To use this modeling for conceptualizing artificial agency, it seems that we first have to think of the different kinds of domains in which we would place these artificial agents (step 1). After establishing the domain, we have to determine the particular laws and regularities that are associated with that domain if artificial agents are included in it as well, and then we have to build a model based on these laws and regularities (step 2). A last interesting step is to see whether current active AI-systems can be described by these models (step 3). With the help of List and Pettit (2011)'s basic account of agency, I will now show how each step can be interpreted.

In step 1, we select the domain. Since we are already interested in artificial agency, we choose the domain of 'agents' as the domain that at least some active AI-systems can belong to. In step 2, then, we have to establish some laws and regularities that are broad enough to apply to both paradigmatic agents such as humans and animals as well as active AI-systems such as those developed in quantum error correction. List and Pettit (2011) have gone about this in the following way. They propose to "imagine a simple system and to ask what we expect of it when we think of it as an agent" (p. 20). What List and Pettit expect of this system they have captured in the following features:

> **first feature.** It has representational states that depict how things are in the environment.
> **second feature.** It has motivational states that specify how it requires things to be in the environment.
> **third feature.** It has the capacity to process its representational and motivational states, leading it to intervene suitably in the environment whenever that environment fails to match a motivating specification. (p. 20)

Humans can perform each of these features and so can animals. Can the active AI-systems in quantum error correction? Yes, or, at least in simulations. They first are able to establish whether a qubit is influenced by noise (feature 1). They have learned that a qubit that is influenced by noise should be corrected by executing particular operations. When they recognize that a particular error occurs at a qubit, they will know what operation to execute to correct it (feature 2). Lastly, they will execute this operation (feature 3). Of course, these systems have only been shown to work in simulations, but since the model in principle applies to them as well, we can accept it as an abstract model of an agent (step 2). Lastly, in step 3, we could look whether other AI-systems fit this model. So with regard to autonomous labs, we could say that the Bayesian optimization algorithm is able to depict how things are (feature 1) and can also specify how it requires things to be (feature 2). However, it would not be able to intervene (feature 3). Still, it could trigger the robotic platform to intervene. The self-driving laboratory as a whole could thus then count as an agent.

Now many might say that if on the basis of this model an autonomous laboratory could count as an agent, the model is too broad. They could base this reasoning on some of the objections mentioned at the end of Sect. 2.2. Or they could base this view on a Gricean model or an analogical model that attributed other features to agents. They could also simply disagree with this particular model and propose different regularities and laws that are abstract enough to in principle be applicable to artificial beings, just not to the currently existing ones. A theoretical model could therefore also direct future research in what features to aim for in the development of active AI-systems. Theoretical modeling can thus lead to new and interesting insights about artificial agency in a way that is again different from Gricean and analogical modeling.

3.4 Some remarks on the neo-Gricean framework 'as is'

As I have shown in each of the previous sections, the modeling strategies included in the neo-Gricean framework lend themselves well for the conceptualization of artificial agency in basic research. One reason that they are useful for this conceptualization is based, I would say, on the fact that they are 'modeling' strategies. (Artificial) agency is a phenomenon that is difficult to observe and study empirically since its existence can only be inferred indirectly through the observation of an entity performing particular behavior—that is, when it 'acts'. Using models as mediators or instruments enables researchers to study the phenomenon of (artificial) agency in a more simple and abstract manner. What is more, models are useful for trying out, and making explicit, one's assumptions regarding the phenomenon or entity or concept[21] that is studied. This explicitness, in turn, creates a research context that lends itself well for revision, criticism, and collaboration—a kind of research context that we should wish for in the study of a relatively new concept like artificial agency.

Another reason why the strategies included in the framework are useful for the conceptualization of artificial agency is that they each, as Sarkia (2021) puts it, "serve different theoretical goals to varying degrees of satisfaction, they may be more or less amenable to different types of targets, and they may complement (or sometimes, compete with) one another" (§1). Gricean modeling, for instance, lends itself well for the analysis of what happens to an ability or characteristic of the entity that is being modeled when one changes certain variables. Rather than studying the entity that one is interested in in isolation, analogical modeling makes a comparative study of a target phenomenon, entity or concept through the use of an analog. Lastly, theoretical modeling looks at the domain of the target, extending the comparison between the target and other phenomena, entities or concepts to a domain-scale. Each modeling strategy thus serves different theoretical goals, can have different types of targets (as in the aspects of artificial agency that we are interested in) and is complementary to the other strategies in that it serves other goals and targets.

Given the framework's plurality of methods and its focus on models, should we adopt the neo-Gricean framework 'as is' for the conceptualization of artificial agency in basic research? I would argue against this, for the following two reasons. First, each of the modeling strategies focuses on modeling the 'phenomenon' of artificial

[21] See again my definition of model as discussed at the beginning of this section.

agency. However, as I mentioned before, the term 'artificial agency' is already in use in basic research interested in the automation of science. At the moment, this term is often used to refer to automated systems that are highly specialized, that can perform rather simple tasks, and whose behavior is still highly influenced by the motivations and directions of the programmer. Researchers interested in the automation of science thus seem to refer to a rather different phenomenon than for example philosophers do. To capture these differences in the meaning of artificial agency and to bring theory (the three strategies) and practice (basic research) closer together, we need to add a fourth strategy that is able to capture this semantic difference and that enables communication about the specific (dis)similarities between the theory and practice.[22] This communication might ensure that the models constructed in the strategies are 'pragmatically plausible' and that current researchers in the automation of science are aware of the possible directions that their research can take.

A further motivation for adding this fourth strategy to the previous three is that it provides a medium to make it possible that the alternative modeling strategies "complement (or sometimes, compete with) one another" (Sarkia, 2021, §1). Adding a semantic dimension, in which one can properly compare the meaning of the concepts used in each of the strategies, enables one to see where the model of each strategy complements, overlaps or contrasts with the others. The fourth strategy would therefore not only bridge the gap between theory and practice, but also provide a medium to properly compare the models that have been constructed through the implementation of the different strategies. So what kind of strategy could function as such a fourth strategy?

4 The fourth strategy: conceptual modeling

The aim of this paper is to find a methodological framework suitable for the conceptualization of artificial agency in basic research. Sarkia's neo-Gricean framework seemed like a favorable option because of its focus on models, its plurality of methods and its usefulness for the conceptualization of complex phenomena. However, as was argued in Sect. 3.4, rather than adopt the framework 'as is', we can enrich it by adding a fourth strategy to the previous three, one that adds a semantic dimension to our conceptualization. The purpose of this additional strategy is to provide the medium in which one can analyse and compare the differences in the interpretation of artificial agency in for example theory vs practice (to obtain pragmatically plausible models) or in model vs model (to see where the strategies complement, or compete with, one another).

In this section, I will propose a modeling strategy that can add this semantic dimension to the framework. This modeling strategy I will call *conceptual modeling* and it is based on the model approach as introduced by Betti and van den Berg (2014). Section 4.1 describes this fourth strategy in more detail and Sect. 4.2 shows how

[22] One might argue that semantics does not let us know anything about the 'phenomenon' artificial agency itself. However, I would argue that our whole scientific teaching about phenomena has taken place in language. Focusing on what we mean by the concepts that we use to refer to phenomena will show us to which phenomena we are actually referring.

conceptual modeling can be implemented to conceptualize artificial agency in basic research. Lastly, in Sect. 4.3, I introduce a new methodological framework, which I call *the Four-Fold framework*, and I argue for its adoption for the conceptualization of artificial agency in basic research.

4.1 Conceptual modeling

Each of the modeling strategies up to now aimed at modeling a phenomenon in the world. Conceptual modeling differs from these strategies in that it does not aim at modeling the phenomena themselves, but rather the concepts that refer to these phenomena. This focus on the meaning of the concepts themselves is important, because even when the same concept—like artificial agency—is used by each of the strategies and in basic research, this concept might have different connotations in the various contexts in which it is used. These connotations might vary in such a way, that the domain of phenomena that is referred to by a particular interpretation of the concept is broader, more narrow, or even completely different from the domains that the other interpretations refer to. To give an example of the latter case, the bigram 'artificial agency' has been used as well to describe states and corporations (see e.g. Runciman 2021; Foisneau 2021). Even though the same term is used as in our case, the connotation, and subsequently the domain of phenomena that is referred to, is completely different—one being automated systems, the other being social entities.

Of course, this is a rather extreme case, but one can imagine that there exist slight differences between the various interpretations of artificial agency depending on the motivations of the researcher, the discipline, etc. Making these differences explicit allows for the possibility of (interdisciplinary) collaboration and communication, revision and an increase in overall understanding. More specific to our case, such communication and collaboration between the modelers and researchers interested in the automation of science will ensure, on the one hand, that the models constructed are pragmatically plausible, and, on the other hand, that researchers become aware of other directions that their research could take. Besides this exchange between the researchers and the modelers, conceptual modeling is helpful for an exchange between the modelers themselves as well. This is because, if we truly want the modeling strategies to complement or even compete with one another, we first have to bring them on equal footing. By specifying what each of the modelers means by artificial agency and what kind of aspects or abilities they associate with this agency, we are able to see where the different models overlap and can complement, or compete with, one another. How, however, can we construct models of concepts?

A relatively new development in the History of Ideas is to make use of computational tools in text-based research. In studying the way that certain ideas have spread over time, historians of ideas have to search through massive bodies of texts that are often divided over different disciplines, span multiple decades and are written in more than one language. Outsourcing this search to computational tools does not only save time, but provides the historian of ideas as well with far more source-material than what they could have obtained by themselves. Still, as Betti and van den Berg (2013) show, without a sound method, the use of these tools will just make the research more

complex, more biased, and the corpus search more challenging. This is, firstly, because "the quantity of source-material a researcher following the traditional method could possibly process is rather small" (Betti et al. 2019, p. 298). Having access to more texts does not change this fact. So, to return to our example, simply searching for texts that contain the bigram 'artificial agency' might produce both texts that mean by artificial agents 'automated systems' and texts that mean by it 'states and corporations'. Historians of ideas thus need a way to specify the connotations related to the concept or idea that they are interested in.

As such a method, Betti and van den Berg (2014) have introduced *the model approach* in the History of Ideas, in which "ideas or concepts are construed [by the historian of ideas] as (parts of) models, that is, complex conceptual frameworks" (pp. 818–819). A model is then "a complex relational structure of subconcepts of both stable (core) and variable (margins) elements the historian ascribes to that idea" (Betti & van den Berg, 2013, footnote 5). Examples of such models can be found in De Jong and Betti (2010) and Betti et al. (2019), but I will discuss here only the latter one. Betti et al. (2019) are interested in tracing the history of the notion of 'conceptual scheme' as it is used in the tradition of Henderson (1932) and Quine (1960). The two conditions necessary for this notion are:

> (1) a science must have a conceptual scheme;
> (2) a conceptual scheme is a/the (or a part of the) theoretical framework of a science in terms of which empirical phenomena, facts or data are interpreted. (§2)

Here the stable elements, or, the core, are the two conditions and how they relate to one another. The margins—also called the *determinables*—are 'empirical phenomena, facts, or data'. These can be differently interpreted or expressed by scholars, while the overall conditions, and the way that they structurally relate to each other, stay in tact.

Let us try this out for our example—the two interpretations of artificial agency. The core of an artificial agent would—very roughly put!—be that an artificial agent is 'constructed' and 'that it can act for itself' (here there is a conjunctive relation between constructed and acting for itself). The margins, which would be defined in more detail, can change from one interpretation to the next. So in the case of states and corporations, this 'constructedness' means in more detail that the artificial agent emerges when a group of people works together. 'That it can act for itself', would, in more detail, mean that it can make (rational) decisions, but not, for example, reflect on these actions.[23] Contrarily, in the case of artificial agents as active AI-systems, this 'constructedness' refers to it being a technology that has been created or programmed by a computer expert. 'That it can act for itself' depends on the level of autonomy and on what the active AI-system can do. Again, this was a very rough and simple example of such a model in the case of artificial agency. When we would actually construct such models of the concepts used in basic research and in each of the models, the model would, depending on how broad our notion of artificial agency is, contain more conditions and be more detailed.

[23] Here discussions about whether or not states and corporations can take moral responsibility are interesting. See for example Hormio (2017).

To summarize, conceptual modeling should thus be understood as an approach in which the modeler construes a concept as (a part of) a model, where a model is a complex relational structure of subconcepts of both stable (core) and variable (margins) elements which the modeler ascribes to the concept. The benefits of the modeling strategy is that it makes the assumptions of the modeler regarding the concept explicit, communicable, and open to revision and improvement. A further benefit of this strategy is that these models lend themselves well for computational searches. The next section shows why we can benefit from a modeling strategy that lends itself well for computational searches.

4.2 Implementing conceptual modeling

Up to now, I have not paid much attention yet to the actual use of artificial agency in basic research. In this section, therefore, I will make a suggestion[24] as to how one can approach constructing a conceptual model of artificial agency in basic research and I subsequently show how one can execute the different steps in this construction process.

Qua suggestion for the construction process, I propose the following steps: (*1*) Take an interpretation of artificial agency from one of the papers on the automation of science; (*2*) Rewrite this interpretation as a model; (*3*) Use this model to do a computational search in a corpus of papers devoted to the automation of science; (*4*) Based on the result of this search, revise the model, and repeat the process.

I will now proceed by showing how each step can be executed.
(*1*) In Ried et al. (2019), a definition of a deliberating agent is put forward which is based on the framework of projective simulation as introduced by Briegel and De las Cuevas (2012):

> entities that can *act* on their environment, thereby generally changing its state, and, more importantly, that make their *own decisions* in the sense that they are not pre-programmed to take particular actions under given circumstances, but instead are flexible and develop their own action and response patterns. (Ried et al., 2019, §3)

(*2*) Based on this definition, we could construct a model of an (artificial) deliberating agent that has the following conditions:

> (1) an *entity* that can act on their *environment*, thereby generally changing its state.
> (2) an *entity* than can make its own decisions in the sense that it is not pre-programmed to perform particular actions under given circumstances, but instead is flexible and can develop its own action and response patterns.

Here the italicized concepts are determinables and might be differently expressed by different papers. Acting could be such a determinable as well, although I would

[24] Please keep in mind that this is still a rather simple suggestion. I am looking forward to comments on how to improve on it.

say that acting itself is one of the necessary conditions of agency, where one can still discuss what one exactly understands by 'acting'.
(*3*) To make this model usable for computational searches, we can use the procedure that Betti and van den Berg (2013) propose:

> (1) humanities experts provide computational experts with a model, i.e. explicitly structured, shared (or shareable)—though not formal—semantic framing of domain knowledge about a certain concept; (2) the computational experts turn the core (stable parts) of the model into an ontology (initial ontology), and adapt techniques of ontology extraction to the domain and the corpus, all of this in close collaboration with the humanities experts ('co-makers', cf. [McCarty (2013, p. 255)]); (3) ontology extraction is applied to the corpus. (§5)

(*4*) Lastly, we can then revise the model based on the close-reading of the texts that were produced by the computational search in step (*3*) and then repeat the process, starting from step (*2*).[25,26]

Once the model is constructed, we can then proceed to make similar conceptual models for the concepts used in the other modeling strategies (and maybe use these for textual searches as well), and see where they overlap, complement, or compete with, this model of artificial agency in basic research and with the conceptual models of the concepts used in the models.

4.3 Introducing the four-fold framework

As argued in Sect. 3.4, even though the modeling strategies included in Sarkia's neo-Gricean framework are truly valuable to our conceptualization of artificial agency, I would not adopt this framework 'as is'. This is because for an account of artificial agency that can be truly useful for basic research, we have to reflect on, and take into account, the different interpretations of artificial agency both in theory (the modeling strategies) and in practice (basic research). In short, we have to add a semantic dimension to the framework. This semantic dimension has the further advantage of enabling one to compare the models produced by the different strategies on a semantic level. By adding a fourth (semantic) strategy to the other three, one can thus:

> (1) Bridge the gap between theory and practice.
> (2) Obtain a medium to properly analyse where the different models overlap, complement, or compete with, one another.
> (3) Create an overall account of artificial agency that is as broad and comprehensive as possible.

[25] To me it seems important to mention the sources on which you base your interpretation of each of the conditions of your model as well. This enables better communication and the possibility for revision, since other researchers can become acquainted with your sources, or suggest better/other ones that might enrich the model.

[26] To ensure that the model is up-to-date and reflects the interpretation of the researchers, an additional step could be to send around questionnaires, in which one asks researchers about their thoughts on how to define artificial agency and to ask for responses to the model.

Since one cannot achieve this by adopting the neo-Gricean framework 'as is', I propose that, for the conceptualization of artificial agency in basic research, we should adopt a new pluralistic and pragmatic methodological framework, one which I will call *the Four-Fold Framework*. Within this framework, the researcher can make use of four strategies in complementary fashion. These four strategies are Gricean modeling, analogical modeling, theoretical modeling and conceptual modeling. By adopting this new methodological framework, we can expect to obtain a broad and comprehensive account of artificial agency in a basic research context.

5 Conclusion

Over the years, AI-research has attempted to automate a broad range of tasks. More recently, AI-researchers have focused their attention on tasks performed in scientific research. Successes in this automation of science suggest that an AI-driven science might not be that far off. This paper forms an initial contribution to the formulation of the conceptual foundations of such an AI-driven science, by proposing a methodological framework that can be used for the conceptualization of (at least one) of the conceptual foundations of AI-driven science—artificial agency. In this paper, I have introduced and proposed the pluralistic and pragmatic Four-Fold Framework as a potential methodological framework. I have motivated this claim by arguing that the four modeling strategies included in the framework can together provide us with a broad and comprehensive account of what we might mean by artificial agency in a basic-research context.

To support my proposal to adopt this methodological framework, I have made and defended the following claims in this paper. In Sect. 2.1, I argued that in basic research scientists are currently developing AI-technologies that are meant to take on a more active role in scientific research. In Sect. 2.2, I provided a number of arguments as to why it would be fitting to call these active AI-systems 'agents'. I further pointed to a number of objections that can still be raised against using this term. However, rather than seeing these objections as defeating, I argued that they should be used to set the standard for any conceptualization of artificial agency that we come up with. For such a conceptualization, though, we need a methodological framework.

In Sect. 3, I introduced Sarkia's neo-Gricean framework as such a possible methodological framework and suggested how each of the modeling strategies included in this framework could be implemented to conceptualize artificial agency. In our search for a methodological framework for the conceptualization of artificial agency in basic research, it has become clear that we can profit a lot from a pluralistic and pragmatic framework that allows us to explore the various ways in which we can approach and think about a relatively new concept like artificial agency. However, since we are interested in a methodological framework for conceptualizing artificial agency *in basic research*, and the term 'artificial agency' is already in use in this field, we need an additional strategy that is able to capture the semantic (dis)similarities between the notion of artificial agency in theory (the models) and in practice (basic research). This is so as to ensure that our eventual account of artificial agency is relevant and pragmatically plausible. In Sect. 4, therefore, I introduced conceptual modeling as a

strategy that could add this semantic dimension to the framework and I concluded by arguing that we should adopt a new framework for the conceptualization of artificial agency in basic research. I called this framework *the Four-Fold Framework*, which includes all four modeling strategies that were discussed in the previous sections.

An AI-driven science might not be that far off and conceptualizing its foundations is therefore urgently needed. One of these conceptual foundations is the notion of 'artificial agency'. However, before one can even begin to conceptualize such a notion, one first has to choose and adopt a methodological framework. By introducing a new methodological framework and by making suggestions as to how this framework can be implemented, I hope to have made this choice easier.

Acknowledgements Support by VolkswagenStiftung Grant Az:97721 is gratefully acknowledged.
A big thank you as well to the anonymous reviewers for their helpful remarks and to those who have taken the time to read (some version of) the paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Aguilar, J. H., & Buckareff, A. A. (2022). Agency and causation. *The Routledge handbook of philosophy of agency* (pp. 27–36). Routledge.
Albritton, R. (1959). On Wittgenstein's use of the term "criterion". *The Journal of Philosophy, 56*(22), 845–857.
Anscombe, G. E. M. (2000). *Intention*. Harvard University Press.
Arnellos, A., & Moreno, A. (2015). Multicellular agency: An organizational view. *Biology & Philosophy, 30*(3), 333–357.
Bailer-Jones, D. M. (2009). *Scientific models in philosophy of science*. University of Pittsburgh Press.
Beisbart, C. (2018). Are computer simulations experiments? And if not, how are they related to each other? *European Journal for Philosophy of Science, 8*(2), 171–204.
Bennett, J. A., & Abolhasani, M. (2022). Autonomous chemical science and engineering enabled by self-driving laboratories. *Current Opinion in Chemical Engineering, 36*, 100831.
Bensaude Vincent, B., & Loeve, S. (2018). Toward a philosophy of technosciences. *French philosophy of technology* (pp. 169–186). Springer.
Betti, A., & van den Berg, H. (2013). Towards a computational history of ideas. *DHLU*.
Betti, A., & van den Berg, H. (2014). Modelling the history of ideas. *British Journal for the History of Philosophy, 22*(4), 812–835.

Betti, A., van den Berg, H., Oortwijn, Y., & Treijtel, C. (2019). History of philosophy in ones and zeros. In M. Curtis & E. Fischer (Eds.), *Methodological advances in experimental philosophy.* Pittsburgh University Press.

Boon, M. (2011). In defense of engineering sciences: On the epistemological relations between science and technology. *Techné: Research in Philosophy and Technology, 15*(1), 49–71.

Bratman, M. (1987). *Intention, plans, and practical reason*. CSLI Publications.

Briegel, H. J., & De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports, 2*(1), 1–16.

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

Coley, C. W., Thomas, D. A., III., Lummiss, J. A., Jaworski, J. N., Breen, C. P., & Schultz, V. (2019). A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science, 365*(6453), eaax1566.

Convy, I., Liao, H., Zhang, S., Patel, S., Livingston, W. P., Nguyen, H. N., & Whaley, K. B. (2022). Machine learning for continuous quantum error correction on superconducting qubits. *New Journal of Physics, 24*(6), 063019.

Council, N. R. (2004). *Science, medicine, and animals*. The National Academies Press. https://doi.org/10.17226/10733

Davidson, D. (1980). *Essays on actions and events*. Clarendon Press.

De Jong, W. R., & Betti, A. (2010). The classical model of science: A millenniaold model of scientific rationality. *Synthese, 174*(2), 185–203.

Du, X., Lüer, L., Heumueller, T., Wagner, J., Berger, C., Osterrieder, T., et al. (2021). Elucidating the full potential of opv materials utilizing a high-throughput robot-based platform and machine learning. *Joule, 5*(2), 495–506.

Dunbar, K. (2001). *The analogical paradox: Why analogy is so easy in naturalistic settings yet so difficult in the psychological laboratory* (pp. 313–334). The analogical mind: Perspectives from cognitive science.

Du Sautoy, M. (2019). *The creativity code*. Harvard University Press.

Erps, T., Foshey, M., Luković, M. K., Shou, W., Goetzke, H. H., Dietsch, H., & Matusik, W. (2021). Accelerated discovery of 3d printing materials using data-driven multiobjective optimization. *Science Advances, 7*(42), eabf7435.

Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., & Senior, A. (2018). De novo structure prediction with deeplearning based scoring. *Annual Review of Biochemistry, 77*(363–382), 6.

Feigl, H. (1958). The 'mental' and the 'physical'. *Minnesota Studies in the Philosophy of Science, 2*(2), 370–497.

Ferrero, L. (2022). An introduction to the philosophy of agency. *The Routledge handbook of philosophy of agency* (pp. 1–18). Routledge.

Foisneau, L. (2021). An answer to David Runciman, <<artificial agency vs artificial intelligence>>. Retrieved from https://hal.archives-ouvertes.fr/hal-03350846/

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170.

Gentner, D., & Maravilla, F. (2018). Analogical reasoning. In J. Ball & V. Thompson (Eds.), *International handbook of thinking and reasoning* (pp. 186–203). Psychology Press.

Gentner, D., & Smith, L. (2012). Analogical reasoning. Encyclopedia of human behavior. *2*, 130–136.

Giere, R. (1988). *Explaining science: A cognitive approach*. University of Chicago Press.

Godfrey-Smith, P. (2006). Theories and models in metaphysics. *The Harvard Review of Philosophy, 14*(1), 4–19.

Gongora, A. E., Snapp, K. L., Whiting, E., Riley, P., Reyes, K. G., Morgan, E. F., & Brown, K. A. (2021). Using simulation to accelerate autonomous experimentation: A case study using mechanics. *Iscience, 24*(4), 102262.

Gongora, A. E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K. G., & Brown, K. A. (2020). A bayesian experimental autonomous researcher for mechanical design. *Science Advances, 6*(15), eaaz1708.

Grice, P. (1974). Method in philosophical psychology (from the banal to the Bizarre). *Proceedings and Addresses of the American Philosophical Association, 48*, 23–53.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Henderson, L. J. (1932). *An approximate definition of fact*. Johnson Reprint Corporation.

Hill, C. S. (1991). *Sensations: A defense of type materialism*. Cambridge University Press.
Holyoak, K. J. (2012). Analogy and relational reasoning. In R. Morrison (Ed.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). Oxford University Press.
Holyoak, K. J., & Thagard, P. (1997). The analogical mind. *American Psychologist, 52*(1), 35.
Hormio, S. (2017). Can corporations have (moral) responsibility regarding climate change mitigation? *Ethics, Policy & Environment, 20*(3), 314–332.
Iten, R., Metger, T., Wilming, H., del Rio, L., & Renner, R. (2020). Discovering physical concepts with neural networks. *Physical Review Letters, 124*, 010508.
Lacey, H. (2012). Reflections on science and technoscience. *Scientiae Studia, 10*, 103–128.
List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.
Majdpour, D., Tsoukas, M. A., Yale, J.-F., El Fathi, A., Rutkowski, J., Rene, J., & Haidar, A. (2021). Fully automated artificial pancreas for adults with type 1 diabetes using multiple hormones: exploratory experiments. *Canadian Journal of Diabetes, 45*(8), 734–742.
Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*(4), 431–467.
McCarty, W. (2013). Knowing. . . : Modeling in literary studies1. *A Companion to Digital Literary Studies*, 389–401.
McDowell, J. (1982). Criteria, defeasibility, and knowledge. *Proceedings of the British Academy London, 68*, 455–479.
Melnikov, A. A., Nautrup, H. P., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A., & Briegel, H. J. (2018). Active learning machine learns to create new quantum experiments. *Proceedings of the National Academy of Sciences, 115*(6), 1221–1226.
Montuschi, E. (2014). Scientific objectivity. In E. M. Nancy Cartwright (Ed.), *Philosophy of social science. A new introduction* (pp. 123–144). Oxford University Press.
Müller, T., & Briegel, H. J. (2018). A stochastic process model for free agency under indeterminism. *Dialectica, 72*(2), 219–252.
Nautrup, H. P., Metger, T., Iten, R., Jerbi, S., Trenkwalder, L. M., Wilming, H., & Renner, R. (2020). Operationally meaningful representations of physical systems in neural networks. *arXiv preprint* arXiv:2001.00593.
Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. *Science and Engineering Ethics, 24*(4), 1201–1219.
Olsen, K. (2012). The first 110 years of laboratory automation: Technologies, applications, and the creative scientist. *SLAS Technology, 17*(6), 469–480. https://doi.org/10.1177/2211068212455631
Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology, 47*(1), 44–50.
Polger, T. W. (2011). Are sensations still brain processes? *Philosophical Psychology, 24*(1), 1–21.
Pruksawan, S., Lambard, G., Samitsu, S., Sodeyama, K., & Naito, M. (2019). Prediction and optimization of epoxy adhesive strength from a small dataset through active learning. *Science and Technology of Advanced Materials, 20*(1), 1010–1021.
Quine, W. V. O. (1960). *Word and object* (new). MIT Press.
Renner, H., Grabos, M., Becker, K. J., Kagermeier, T. E., Wu, J., Otto, M., et al. (2020). A fully automated high-throughput workflow for 3d-based chemical screening in human midbrain organoids. *elife, 9*, e52904.
Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications, 11*(1), 1–9.
Ried, K., Eva, B., Müller, T., & Briegel, H. J. (2019). How a minimal learning agent can infer the existence of unobserved variables in a complex environment. *CoRR, abs/1910.06985*. Retrieved from arXiv:1910.06985.
Rooney, M. B., MacLeod, B. P., Oldford, R., Thompson, Z. J., White, K. L., Tungjunyatham, J., & Berlinguette, C. P. (2022). A self-driving laboratory designed to accelerate the discovery of adhesive materials. *Digital Discovery, 1*(4), 382–389.
Runciman, D. (2021). Artificial agency vs. artificial intelligence. *Handout of a talk given online at the séminaire de philosophie politique normative, 8 juin 2021, 17–19h, hosted by the University of Paris*.
Russo, F. (2017). Model-based reasoning in the social sciences. *Springer handbook of model-based science* (pp. 953–970). Springer.
Sarkia, M. (2021). Modeling intentional agency: A neo-Gricean framework. *Synthese*, 1–28.

Schneider, T., Jeevanjee, N., & Socolow, R. (2021). Accelerating progress in climate science. *Physics Today, 74*(6), 44–51.
Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE, 109*(5), 612–634.
Senior, A., Jumper, J., Hassabis, D., & Kohli, P. (2020). *Alphafold: Using AI for scientific discovery*. Blog Post. https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientificdiscovery
Smart, J. J. (1959). Sensations and brain processes. *The Philosophical Review, 68*(2), 141–156.
Soldatov, M. A., Butova, V. V., Pashkov, D., Butakova, M. A., Medvedev, P. V., Chernov, A. V., & Soldatov, A. V. (2021). Self-driving laboratories for development of new functional materials and optimizing known reactions. *Nanomaterials, 11*(3), 619.
Sparkes, A., Aubrey, W., Byrne, E., Clare, A., Khan, M. N., Liakata, M., & King, R. D. (2010). Towards robot scientists for autonomous scientific discovery. *Automated Experimentation, 2*(1), 1–11.
Spellman, B. A., & Holyoak, K. J. (1992). If saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology, 62*(6), 913.
Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.
Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Penn State Press.
Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.
Witherspoon, E. (2011). *Wittgenstein on criteria and the problem of other minds*. The Oxford handbook of Wittgenstein.
Wright, C. (1984). Second thoughts about criteria. *Synthese*, 383–405.
Wu, T., & Tegmark, M. (2019). Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E, 100*, 033311.
Young, M. (2021). *How artifacts acquire agency*. Retrieved from https://www.2021spt.com/ (SPT conference 2021)