Speculative Exploration on the Concept of Artificial Agents Conducting Autonomous Research

Shiro Takagi
Independent Researcher*

Abstract

This paper engages in a speculative exploration of the concept of an artificial agent capable of conducting research. Initially, it examines how the act of research can be conceptually characterized, aiming to provide a starting point for discussions about what it means to create such agents. The focus then shifts to the core components of research: question formulation, hypothesis generation, and hypothesis verification. This discussion includes a consideration of the potential and challenges associated with enabling machines to autonomously perform these tasks. Subsequently, this paper briefly considers the overlapping themes and interconnections that underlie them. Finally, the paper presents preliminary thoughts on prototyping as an initial step towards uncovering the challenges involved in developing these research-capable agents.

Contents

*https://t46.github.io/

The LaTeX file for this paper is currently hosted on GitHub. As the paper is still a work in progress, it may contain errors. Should you have any suggestions for improvement, or if you identify any mistakes or misunderstandings, please do not hesitate to submit a pull request to the repository. Your contributions towards refining this paper are greatly appreciated: https://github.com/t46/research-automation-perspective-paper

1 Introduction

Research has been the foundation of human progress. Through research, humans have deepened their understanding of the world and created unprecedented innovations, leading to groundbreaking advancements. It would be not an exaggeration to say that the future development of humanity heavily depends on the evolution and progress of research endeavors.

Since the inception of artificial intelligence (AI) research, a key goal has been to develop AI capable of conducting research [1]. AI-led research not only accelerates existing research and development but also offers significant potential for improving research practice and methodologies themselves, unencumbered by human cognitive limitations, research conventions, or unnecessary social constraints [1, 2].

Researchers have developed systems that automate scientific decision makings [3], infer natural laws [4], autonomously cycle through hypothesis and experimentation [5], and many more [1, 6]. With advancements in machine learning, there have been remarkable successes in using it for scientific discovery [7, 8, 9, 10]. These efforts by humanity have significantly advanced the potential of machines as valuable assistants in research.

On the other hand, the journey toward fully autonomous intelligent agents[1] capable of conducting research is still ongoing [1, 11]. "Autonomous" here implies functioning without human intervention, prior design, or preparation. An agent is considered more autonomous if it can conduct research with less human involvement. By "agents that conduct research," I refer to a single agent capable of performing research activities in various fields, like history, mathematics, or physics. The wider the range of research a single agent can handle, the more general it is considered. In this paper, when referring to an agent capable of conducting research, it denotes a general and autonomous artificial researcher. The realization of such agents has been a deeply held aspiration in the quest for advancing human research capabilities.

While there are excellent papers presenting perspectives on AI capable of conducting research [1, 12, 11, 2, 7, 6, 9, 13, 14], there is still much to discuss about the nature of such agents and what it means to create an agent capable of doing research. Therefore, this paper presents a speculative thought around the concept of artificial intelligent agents capable of conducting research. This discussion aims

[1] In this paper, the terms AI, machine, and agent are used interchangeably.

to provide an opportunity to consider what future discussions are needed as we work towards realizing agents capable of conducting research

First, I will explore conceptually how to characterize the research activities. This preliminary discussion is intended to help us consider what it would mean to create an agent capable of conducting research and what discussions would be necessary for its development. Next, I will discuss elements widely considered essential in research, specifically exploring the nature of question construction, hypothesis generation, and hypothesis verification, along with the potential challenges in autonomously performing these tasks. Subsequently, I will consider topics that combine these elements, topics common to them, and points that could not be discussed previously. Finally, as a reference, I will share some simple, preliminary ideas for prototyping aimed at identifying challenges in developing an agent capable of conducting research.

The speculative discussion in this paper is still in its early stages and is provisional. Given the breadth of the subject matter and my limited capabilities, each point of discussion may be somewhat superficial or not entirely accurate. I plan to update these discussions continuously. Therefore, if anyone notices any points that should be improved, errors, or topics that would be beneficial to discuss, I would greatly appreciate your feedback.

2 Conceptual Characterization of Research

Understanding the fundamental nature of research is crucial for creating an agent capable of autonomous research. This section will therefore speculatively consider how the act of research can be characterized.

The aim of this section is not to establish a universal and singular definition of research, a task that exceeds the scope of this paper. Rather, it explores some characteristics of research to provide a provisional basis for discussions on the development of an artificial researcher.

As such, the definition of research presented here should be regarded as tentative and operational, and the ensuing discussion is just one example of an endeavor to characterize research. Refining our understanding of research through in-depth discussions in the future is essential for the development of agents capable of conducting research.

2.1 Research as Knowledge Production

While finding a unified, all-encompassing definition of research or science remains infeasible [15, 16], various interpretations exist. For instance, one view posits that research occurs "whenever we gather information to answer a question that solves a problem" [17], while another describes research (and development) as comprising "creative and systematic work undertaken in order to increase the stock of knowledge" [18]. Additionally, some perceive the science as "processes that maximize the evidence for a generative model of the sensed and measured world" [19]. These descriptions highlight different crucial aspects of research, with none being entirely incorrect or absolutely definitive.

Among these, a broadly recognized definition can be that **research is an endeavor to generate new knowledge**. Since this characterization seems to align with our research practices, regardless of the field, this definition serves as a suitable starting point for our discussion. In this paper, I adopt this interpretation as a provisional working definition. Specifically, I will regard research as the attempts to produce new knowledge for certain society. The inclusion of "for certain society" acknowledges the societal relativity of knowledge, a point I will elaborate on in subsequent sections.

2.2 Knowledge Production as Belief Revision

Having defined research as the endeavor to generate new knowledge, it becomes important to consider what "knowledge" itself entails, and what constitutes its production. This section aims to explore these concepts.

Defining knowledge and the process of knowledge production rigorously remains an unsettled philosophical debate [20]. Given that providing a precise definition of knowledge is beyond the scope of this paper, an in-depth exploration of these debates will not be undertaken here. Instead, this section aims to present a basic and preliminary conception of what knowledge might entail so that it serves as a starting point for further discussion.

2.2.1 Knowledge as Belief

The concept of knowledge has long been a subject of debate within *epistemology*, a branch of philosophy. This paper will reference some basic and introductory concepts in this field as an exemplar to explore how research might be characterized.

Within epistemology, knowledge has traditionally been viewed as *justified true belief (JTB)* [20]. The term "true" is challenging to define rigorously; however, for the purposes of this discussion, it can be understood as something that corresponds with fact. "Belief" is tentatively defined as an individual's thought or conviction about a subject. "Justified" implies that it is reasonable to hold such a belief. The nature of justification has been a focal point of debate in epistemology, particularly following criticisms that JTB may not adequately define knowledge [21]. Consequently, the refinement or expansion of the JTB has become a significant topic in epistemological discourse [20].

Although many philosophers contend that the JTB properties alone are insufficient for defining knowledge, there is some consensus that they might be necessary components [20]. In epistemological debates, rather than discarding JTB entirely, many theorists use it as a foundational concept. Hence, this paper will tentatively adopt JTB as a preliminary basis for discussion.

The subsequent sections will explore how research is conceptualized under this definition. Analyzing the congruences and discrepancies between the implications drawn from this characterization and our expectations of a research-capable agent will generate insights for refining the definition of research and hypothesizing the capabilities such agents should possess.

2.2.2 Knowledge Production as Belief Updating

In the current framework, knowledge is equated with belief. Therefore, knowledge production can be reinterpreted as the process of adopting the belief that a particular proposition is true and subsequently revising this belief based on justification. Essentially, in this framework, knowledge production equates to the updating of beliefs.

While it might initially seem counterintuitive to view research as a process of updating beliefs, this perspective gains plausibility when considering several factors. Research involves the continual revision of hypotheses and theories; inductive reasoning, unlike deduction, does not conclusively prove propositions; and, as will be discussed, knowledge is essentially subjective. The characterization of research as belief aligns well with these aspects of research. Therefore, this characterization seems reasonably valid.

The primary aim of research can be conceptualized as uncovering the unknown truths of the world. Consequently, the justification employed in research must be capable of accurately discerning the truth or falsity of propositions. This form of justification, known for its capacity to lead to truth, is termed "truth-conducive." While there are varied debates on the nature of justification, it is widely accepted that justification in research should indeed be truth-conducive. Thus, knowledge production can be conceptualized as the construction of belief in new propositions about the world and ascertaining their veracity through truth-conducive justification.

2.3 To Know Depends on Knowing Subjects

Fundamentally, the concept of knowing presupposes not only the existence of the object being known but also of the subject doing the knowing. This interplay explains why the definition of knowledge incorporates the subjective element of belief. Consequently, while the notion that knowledge is a form of belief might initially seem counterintuitive, it holds validity in this context.

Moreover, conceptualizing research as an updating of beliefs aligns closely with actual research practices. For instance, the experimental validation of a hypothesis reinforces our belief in its truth or falsity. Our confidence in a hypothesis increases as it withstands various rounds of verification. This process of iterative validation and belief reinforcement mirrors the concept of research as a continual renewal of beliefs.

Finally, since the justification in research is expected to be truth-conducive, the knowledge thus produced would have an objective quality. Therefore, in conjunction with the discussion in the previous section, it seems that the use of the subjective concept of belief is not that problematic.

2.3.1 Knowledge for Humans

As previously discussed, research is a pursuit dedicated to uncovering the unknown truths of the world, necessitating that the knowledge it generates be novel. This raises the question: what constitutes new or unknown knowledge?

Under the current definition, knowledge is a justified belief regarding the truth or falsity of a certain hypothesis. Thus, unknown knowledge could be a state where such a belief is either non-existent or, if it exists, lacks justification. In simpler terms, within this framework, the state of certain knowledge being unknown is essentially a state of belief.

Given that the concept of knowing is contingent on the knowing subject, the notion of the unknown is also inherently subject-dependent. In the realm of research, the term "subject" has seemingly encompassed humanity at large. Researchers do not deem knowledge as unknown simply because it eludes an individual; it is regarded as truly unknown only when it is beyond the collective understanding of humanity. Consequently, the knowledge produced through research is expected to contribute to the collective understanding of human society. This is the reason why the term "for society" is included in the definition of research.

2.3.2 Knowledge for Non-Humans

Since the act of knowing is subject-dependent, it is theoretically feasible to conceive of non-human knowledge and research by considering non-human agents as knowing subjects. Naturally, there is also the unknown for these non-human agents. Such knowledge and unknowns for non-humans can naturally differ from those for humans. Therefore, when referring to "knowledge," "unknown," or "novel," it is necessary to specify for whom these concepts apply.

While this discussion might seem like mere speculation, the idea that machines might have a different scope of the unknown compared to humans has implications for realizing an artificial researcher. This is because it suggests that merely replicating current research methodologies in AI might not necessarily yield new knowledge for humans.

Research methods essentially have been developed to uncover truths unknown to knowing subjects. Therefore, an AI mimicking these methods might only reveal truths unknown to itself, which may not align with human knowledge gaps. If we want AI to conduct research autonomously, we must find a way to ensure it understands what is unknown to humans, not just to itself, and guide it to discover knowledge that is truly unknown in the human context.

This is just a preliminary discussion. It is hoped that further discussion will continue on how to realize them for developing research-capable AI

2.3.3 Can Truth-Conducive Justification Be Developed by Non-Human Agents?

I acknowledge that we cannot definitively "prove" our empirical verification methods to be entirely truth-conducive. Yet, given the myriad discoveries achieved through these methods, questioning their legitimacy seems to have little merit. If agents can thoroughly grasp and effectively apply these human-developed justifications, it is poised to uncover numerous unknown truths, a prospect few researchers would dispute.

Then, what about when the agent constructs its own methods of justification? Is it possible for an agent to construct new truth-conducive justification methods on its own, instead of just mastering human-developed methods? Humans have devised tools like statistical hypothesis testing to evaluate hypotheses; can AI similarly innovate unique methodologies? Even if it were possible, how significant would those be?

Our justification methods are founded on various premises, implying diverse interpretations of "what constitutes a truth-conducive justification" or "what justification entails." These interpretations lead to multiple methods of justification even among humans [22]. Consequently, when allowing artificial agents to autonomously develop justification methods, these agents must consider what can serve as justification, navigate value judgments regarding the nature and effectiveness of justification, and thereby conceive and select optimal methodologies.

This inquiry goes beyond mere philosophical speculation. Since the current justification methodologies have not been proved to be the optimal, the potential for superior methodologies exists in theory. Machines, unfettered by human cognitive limitations, theoretically possess the possibility to discover such methods. Importantly, truth-conducive verification methods does not essentially require

human value judgments to evaluate their quality, thus, there is ample potential for machines to "autonomously" devise them. The feasibility, realization, and significance of these possibilities remain open for exploration and debate.

2.4 Conclusion

In this section, I have discussed a provisional working definition of research. My initial premise was the intuitive belief that research is an endeavor aimed at generating new knowledge for a particular society. The discussion subsequently delved into a speculative inquiry into the idea that that knowledge is fundamentally a form of belief and that the production of knowledge is the updating of beliefs. Building on these ideas, I presented some conjectural insights of non-human agents doing knowledge production.

It is important to note that the definition proposed here serves merely as a starting point. A more comprehensive and nuanced understanding of research can be cultivated through the collective insights of philosophers, scientists, and practitioners across various fields. This collaborative approach will enable us to delve deeper into the definition of research and develop more robust and effective guidelines to realize an artificial researcher.

3 Question, Hypothesis, and Verification

In the preceding section, I briefly examined a preliminary conceptual definition of research and its implications. This section shifts focus to the widely acknowledged fundamental components of research: question construction, hypothesis generation, and hypothesis verification.

The objective here is to advance the discussion beyond the somewhat too abstract considerations in the previous section. By dissecting these core elements, this section seeks to offer a more concrete exploration of what constitutes research and the prospects of machines engaging in research activities.

3.1 Question Construction

The first essential element in research is *question construction*. To produce new knowledge, it is imperative to recognize what is unknown and strive to generate that elusive knowledge. This act of identifying the unknown for investigation can be considered as the process of questioning. Subsequently, formulating potential answers for these questions constitutes hypothesis generation. Essentially, research can be reinterpreted as the act of posing and answering to questions. Furthermore, since research inherently involves lots of uncertainty, the generation of multiple questions, beyond the initial research question, is a natural part of the process. That is, in the pursuit of confronting the unknown, question construction is an inevitable aspect of research.

There have been the studies to find research questions and challenges from academic literature [23, 24, 25], generate ideas for future work [26], and identify research trends [27, 28]. However, enabling machines to autonomously generate research questions is less common. In the domain of question answering from natural language processing (NLP) research, tasks exist for generating questions [29, 30], but these are motivated differently from generating research questions. Research on artificial curiosity for generating non-textual questions has been conducted [31], yet it doesn't generate research questions akin to a human researcher, as far as I know. Recent advancements in large language models (LLMs) have led to initial attempts at generating research questions [32, 33], but this area remains nascent.

While there are efforts towards automation as shown above, the number of these attempts are relatively limited compared to those for hypothesis generation and verification. The automation of question construction, or determining the underlying goals of such automation, is recognized as a key challenge in the field of research automation [11, 6, 2].

In this section, I start with a speculative exploration of the nature of questioning. This will be followed by a discussion of the open challenges in enabling an artificial agent to effectively pose research questions.

3.1.1 What is Questioning?

Asking questions is often characterized as an information-seeking behavior [34, 35]. This behavior typically involves two distinct steps: firstly, recognizing an *information need*, and secondly, undertaking actions to acquire the desired information [36, 37]. While not all information-seeking behaviors necessitate linguistic expressions [34], in the context of research, queries are typically formulated in text. This textual formulation occurs between the stages of information need recognition and the initiation of information-seeking behavior. Specifically, in research, the process of question construction is generally understood as the journey leading to the formulation of such queries. Thus, for the purposes of this paper, question construction is regarded as the process culminating in the formulation of a query. The subsequent steps of information seeking are considered part of hypothesis generation and verification.

Recognizing an information need seems to involve at least two sub-processes: identifying the knowledge gap and deciding to address it (judging that the missing information is a "need").[2] Therefore, to enable an artificial agent to autonomously construct questions, it is necessary to consider how to imbue it with these capabilities. The following sections will delve into a speculative consideration and exploration of these steps.[3]

3.1.2 Recognizing Unknowns

Recognizing that certain knowledge is unknown typically involves an initial attempt to access that knowledge. This process usually entails referring to our personal knowledge base and, upon not finding the information, deeming it as unknown. For an individual, this knowledge base is essentially the memory stored within the brain. However, in the realm of research, the unknowns that researchers aim to elucidate are those unknown to a specific society, not just to an individual researcher. That is, a research-capable agent does not need to judge whether it is unknown to itself, but rather it can directly determine whether it is unknown to certain society. In this context, the knowledge base extends beyond the agent's memory to encompass societal knowledge sources, such as a collection of research papers.[4]

As outlined in Section 2, for machines to generate new knowledge beneficial to humans, they must be capable of identifying what is unknown to humans, not to themselves. While this task might initially seem as straightforward as conducting a literature survey as humans do, the reality may necessitate more complex approaches. The specific requirements for achieving this goal merit further discussion.

An additional consideration arises regarding the reliance on a machine's judgment to determine what is unknown to us. If an AI has been pre-trained on an extensive corpus of scholarly papers, its judgment might appear credible. However, as previously mentioned, an AI's determination of unknowns does not necessarily coincide with human unknowns. Therefore, it is not certain whether it is appropriate to unconditionally believe that what AI has determined to be unknown is indeed unknown. This issue becomes increasingly pertinent and significant as AI amass more knowledge and enhance their capabilities.

3.1.3 Deciding What Knowledge to Seek

While we encounter numerous unknowns, we do not formulate questions for each one, as not all unknowns hold equal "importance" or "interest." Instead, we construct questions for matters we are eager to understand. This process involves assessing the "value" of questions based on certain criteria to determine their worthiness of pursuit. For individuals, this can be a largely subconscious process. However, in research, this need not be an internal process, as long as that is the value judgments of questions.

[2] In this discussion, the process of recognizing an information need is described as initially identifying something as unknown and then deciding whether to formulate a question about it. However, the sequence of these steps is not fixed. For instance, one might first have a desire to know something and only afterward ascertain that it is indeed unknown. The critical aspect is that the process encompasses these two elements.

[3] It is important to note that questions in research are not personal but societal in nature. This societal aspect may introduce slight variations in the question construction process. This point will be revisited later in the paper.

[4] As mentioned earlier, something being unknown implies either the absence of a proposition or the presence of an unverified belief. Therefore, accurately determining the unknown from academic papers requires an assessment of whether each paper has been appropriately justified, meaning whether its verification processes are sound.

Knowledge, in itself, is value-neutral. The "value," "significance," or "goodness" of knowledge is ascribed by its users. A noteworthy aspect here is that the criteria for determining "value" are subjective and arbitrary. Hence, if we aim for artificial agents to autonomously pose questions meaningful to humanity, it is crucial to identify what constitutes "good" or "significant" questions for us and instill these values in the agents.

On the other hand, it is also vital to recognize that certain questions deemed "unimportant" by us may actually hold importance under different criteria. Fundamental research, for example, often yields knowledge initially perceived as "useless" but later proves pivotal for innovations. Human cognitive limitations may sometimes hinder our ability to fully appreciate the potential utility of such knowledge. Moreover, social factors unrelated to the initial purpose of knowledge production can influence our value judgments, implying that these human judgments are not always optimal.

Given that machines are not inherently limited by such constraints, they could theoretically make more effective value judgments. Therefore, while providing some guidance to ensure that the generated question is relevant to humans is crucial, developing agents capable of autonomously constructing these value criteria themselves may also be fruitful. How to achieve this forms a significant open challenge in developing research-capable agents.[5]

3.1.4 Origin of Information Need

I previously outlined that question formation begins with the recognition of an information need. This leads to the question: what triggers the recognition of an information need?

The initiation of this process can be attributed to various factors. Some researchers may generate questions through logical contemplation aimed at achieving a specific goal. Others might identify questions upon noticing anomalies in experimental data or inconsistencies between theoretical assumptions and actual observations. Researcher also sometimes search questions that can be answered by techniques that you have. Furthermore, humans typically do not rely on a singular criterion for value judgment. Instead, multiple criteria are often intricately combined and weighted according to the context, culminating in a complex value assessment process. To develop an agent capable of autonomously constructing research questions as humans do, therefore, it appears necessary to create a system with a general methodology for questioning applicable across these diverse scenarios.

The process in humans that connects these various triggers to an information need, and the development of an agent capable of emulating this process, remains an open question. Researchers in the field of curiosity, which is broadly conceptualized as a "drive state for information" [38], have been investigating this challenge. Curiosity is often characterized as a precursor to information need in information-seeking processes [37]. In reinforcement learning, efforts to instill curiosity or knowledge-based intrinsic motivation in AI have been explored. Here, curiosity is defined in terms of novelty, information gain, or prediction error, and is considered a catalyst for exploration [39].

These efforts provide insights into implementing mechanisms that drive AI towards question formation. However, we are still distant from realizing a system that autonomously constructs research questions under complex value judgments, as humans do. A significant challenge lies in identifying the minimal input required for question generation; namely, while the minimal input is clear for hypothesis generation and hypothesis verification, it remains unclear for the construction of questions. Designing a complex, contextually adaptive internal driving force for questioning remains a significant hurdle. Identifying the prerequisites for an AI system with such a mechanism is an ongoing challenge.

3.1.5 Examples of Criteria for Evaluating Research Questions

Thus far, I have discussed abstract concepts related to questioning in general. Now, I will move on to focusing specifically on the characteristics pertinent to research questions.

The "quality" of a research question can be evaluated against various criteria. Here, I will briefly explore some examples to illustrate how humans seemingly appraise the value of a research question. It is important to note that these examples are only a few of the many criteria utilized and do not represent a comprehensive list.

[5] Kitano has described the approach where humans apply their value judgment criteria to determine questions and hypotheses as *value-driven science* [2]. He advocates for the advancement of *exploration-driven science*, which prioritizes extensive and comprehensive exploration. While a completely value-neutral exploration is unattainable, the notion of employing diverse and extensive criteria is indeed significant for the future of research. By embracing a wider range of criteria, we can expand the exploration space of knowledge.

One widely accepted criterion within the research community is that a question is important if it offers new perspectives, understandings, or conceptual advances, particularly those that challenge our common assumptions. For example, Alvesson and Sandberg emphasize the significance of such questions and discuss strategies for their construction [40]. This criterion rests on the idea that a valuable question is one that significantly impacts our current knowledge. This seems to be a value that aligns with the highest-level objectives of the endeavor of research.

No matter how significant a question may be, if it is nearly impossible to address with current technology, deriving meaningful research outcomes from it may be unfeasible. Consequently, the feasibility of answering a question is considered a vital aspect of its quality [41, 42, 43]. Assessing feasibility involves complex decision-making, taking into account factors like available resources, researcher capabilities, deadlines, and technological constraints. An agent engaged in research would need the capacity for such multifaceted evaluations.

Another prevalent view is that research questions should stem from individual intellectual curiosity. Given that curiosity drives exploration [44], curiosity-driven research can foster exploration in the knowledge space. Research can be seen as an exploration of the world's truths, making this value standard important. However, curiosity is not the sole criterion for exploration; there may be better criteria for uncovering unknown truths. If agents can adopt such criteria, it might surpass human efficiency in uncovering truths.

In contrast to bottom-up curiosity-driven research, questions that contribute to achieving specific top-down goals are also considered valuable. For example, in corporate or government-led research, questions aligned with predetermined objectives are prioritized. Since we expect that agents capable of autonomous research will contribute to human-set goals, ability to make such value judgments deemed important.

In practice, the value of a research question is determined by integrating multiple criteria. Hulley et al. suggest that questions which are feasible, interesting, novel, ethical, and relevant (FINER) are considered valuable [41]. Huntington-Klein argues that a good research question is one that is answerable and whose answer enhances our understanding of the world [43]. As mentioned above, autonomous agents are also expected to determine the questions they should pursue based on such complex value judgments.

As emphasized, these criteria represent only a portion of the value judgments humans make in question formulation. Future discussions should further investigate the nature of these judgments, their role in scientific discovery, and how they can be replicated in artificial researcher.

3.2 Hypothesis Generation

The second integral element of research is hypothesis generation. Research inherently involves posing questions and endeavoring to answer them. Typically, researchers bifurcate the answering process into two phases: generating hypotheses and verifying them. Hypothesis generation entails predicting the answer to a posed question, while hypothesis verification involves examining the plausibility of that prediction. This two-stage approach is adopted because researchers address questions to which no one in this world knows the answer, making it challenging to immediately ascertain definitive answers. Therefore, the separation of hypothesis generation and verification represents a human-developed methodology for uncovering truths in a context of high uncertainty.

Hypothesis generation is often seen as a showcase of human creativity in research. The long-standing belief that human creativity defies analysis has led to the assumption that both question construction and hypothesis generation are inherently unanalyzable [45]. However, efforts to characterize this creative process began to emerge in the mid-20th century. Notable concepts include the role of abduction in generating hypotheses for "why" questions [46, 47], the significance of analogical reasoning [48], the interpretation of scientific discovery as a form of search problem [4], and the conceptualization of hypothesis generation as probabilistic sampling [49].

The potential for machines to generate hypotheses has been a focal point in artificial intelligence research. Pioneering attempts to develop machines capable of hypothesis generation date back to the early 20th century [4, 3]. By the mid-2000s, advancements led to the creation of machines capable of making autonomous scientific discoveries [5].

3.2.1 Hypothesis Generation and Machine Learning

Generating hypotheses is predicting answers to questions from existing knowledge. This process essentially aligns closely with machine learning, particularly question-answering. As it involves predicting answers that even nobody knows, it can also be viewed as a prediction under significant distribution shifts.

Indeed, machine learning have become increasingly prominent in scientific hypothesis generation [8, 9, 7]. Attempts such as predicting protein structures [50] and new materials [51] are all examples of hypothesis generation.[6]

The sources for hypotheses, the nature of the hypothesis space, and the representation of hypotheses differ across research fields. For instance, hypotheses can be represented combinatorially, and machines can be employed to explore these spaces to find hypotheses [12]. Some studies represent hypotheses as symbolic equations and try to discover them from scientific data [52], while others endeavor to generate or extract textual hypotheses from academic papers [53, 54, 55, 56, 57]. The advent of LLMs has spurred efforts to generate hypotheses from the models' internal knowledge, without relying on direct information from academic papers [58, 10].

While the specifics of hypothesis generation vary, a unified description can be drawn from certain perspectives. Viewing the hypothesis space as a human-defined and fixed entity, scientific discoveries can often be framed as search problems [12]. Since the hypothesis space is often combinatorially vast, strategies for efficient exploration in the space deemed necessary [11, 1], and efforts have been made to optimize exploration with techniques such as active learning. As another perspective, Wang et al. provide a categorization of how AI is utilized in scientific hypothesis generation, highlighting its applications in black box prediction, aiding hypothesis space exploration, and finding solutions within a differentiable hypothesis space [7].

As such, integration of machine learning to hypothesis generation has progressed significantly compared to question formulation and hypothesis testing. The applications in this field are vast and diverse, and detailed evaluation of individual cases goes beyond the author's capacity. Therefore, this paper does not provide introductions to specific cases. Those interested are encouraged to refer to survey papers in their respective fields.

While machine learning's application in hypothesis generation is notable, the development of AI capable of generating complex hypotheses in response to varying questions remains a challenge. Achieving such capability may require abilities to generate hypothesis in versatile and flexible manner and to construct hypothesis spaces themselves. Future discussions are expected to further explore how to realize these capabilities in AI.

3.2.2 Speculation on Key Aspects of Hypothesis Generation by Machines

It can be said that autonomous hypothesis generation in general manner by AI has already gained considerable attention, compared to question generation and hypothesis verification. That's largely because, as previously mentioned, predicting answers to questions is a problem already central to many machine learning researchers. Therefore, many challenges in aiming for AI that generates hypotheses as flexibly as humans overlap with the challenges of pursuing an artificial general intelligence (AGI). These include systematic thinking such as deduction, out-of-distribution generalization, causal inference, efficient exploration, and problem decomposition, all crucial for autonomous flexible hypothesis generation and considered fundamental in the pursuit of AGI as well.

In this section, I will preliminary explore elements deemed important for AI's ability to generate hypotheses. However, due to the circumstances mentioned in the previous paragraph, this discussion might intersect with existing debates in the realm of AGI, potentially lacking novelty. Nonetheless, I will explore two aspects that appear vital for the development of an artificial hypothesis generator.

Firstly, it's important to note that even AI might not know the answers to the research questions. It's not always true that a question unknown to humans is also unknown to machines, as has been repeatedly emphasized. However, once the answer is unknown to both humans and the machine, hypothesis generation can be a challenging task even for AGI. I acknowledge that this essentially reduces to a problem of out-of-distribution generalization, but it's particularly challenging because no agents in this world know the answer. To solve such problems, machines, like humans, may need to recognize their ignorance of the answer, reduce uncertainty step-by-step, and gradually approach the

[6] Since there are a vast number of studies, I will skip the introduction of individual studies here.

answer. Current AI still does not even understand what it doesn't know [59, 60]. How AI capable of reasoning under such high uncertainty can be realized remains an open question.

Secondly, the role of mathematics in hypothesis generation cannot be overstated. The first point to note is that the power of mathematics in hypothesis generation lies significantly in its deductive nature. Deduction ensures that if the premises are true, the resulting conclusions are also true, even if they may seem counterintuitive. This aspect gives AI, which largely depends on experiential inferences, a substantial advantage. Furthermore, as humans do by the hypothetico-deductive method, deduction enables the evaluation of hypotheses that are not directly testable. If deductive results are rejected, the hypothesis is deemed false; acceptance, conversely, strengthens its plausibility. This plays a crucial role in expanding the empirical knowledge boundaries in research.

The abstract nature of mathematics is also important. Since ancient times, even before the formalization of deductive methods, mathematics engaged with concepts such as numbers, which are fundamentally abstract and have long captivated human interest [61]. The introduction of symbolic representation and manipulation has further amplified its abstract nature. Significantly, mathematics not only abstracts real-world objects but also engages in a cycle of further abstraction. By abstracting already abstracted concepts, it has developed highly sophisticated systems [62]. This level of abstraction allows for the reference to subjects not directly experienced, facilitating the progress in science [63].

These characteristics render mathematics an indispensable tool in the process of hypothesis generation. AI capable of doing mathematics has not yet been realized, but related research attempts have made steady progress [64, 65, 66].

While these two elements are discussed separately, systematic thinking seems necessary for both, reinforcing the widely acknowledged importance of systematic or high-level thought in AI development. However, due to the extensive existing discourse on this topic [67], I will not delve deeper into it here.

Due to my limitations, this paper only scratches the surface of this topic. I would appreciate any feedback from those with insights into elements not widely recognized in the machine learning community but deemed essential for autonomous hypothesis generation.

3.3 Hypothesis Verification

The final critical element in the research process is the verification of hypotheses. We justify our belief in the truth or falsehood of a hypothesis by confirming the plausibility of our prediction in response to a question through verification. Thus, verification is essential for generating knowledge.

Verification hinges on the nature of the question and hypothesis posed. For instance, a "why" question demands verification methods that elucidate causal relationships. Questions about the physical world require interaction with physical world for verification. In cases where hypotheses are amenable to mathematical proof, such proof constitutes verification. This necessitates an agent capable of verification to possess an understanding of what constitutes verification and to develop suitable verification methods tailored to the specific question and hypothesis.

While there has been extensive discussion on AI in hypothesis generation, its involvement in verification is less explored. Certainly, some studies have utilized AI in aspects of verification, such as experimental design [68] and scientific simulations [69]. However, initiatives enabling AI to fully comprehend and independently execute verification processes akin to human scientific research are still limited.

In machine learning, research focusing on the validation of scientific claims [70], factual accuracy of predictions [71], evidence search to support hypotheses [72], and self-verification of machine responses [73] aligns with aspects of verification. Nevertheless, none of them aim to construct and execute verification as humans do in a scientific research. The automation of peer review [74, 75] is also related to verification in the sense that it demands judgment on the validity of the verification, but it does not generate verification.

In this section, I aim to delve into the concept of verification to stimulate further contemplation. Having already addressed the nature of verification, or justification, in Section 2, I will omit that discussion here. Instead, I will focus on experimentation, an essential aspect of human-like verification in scientific inquiry.

3.3.1 Experimentation

No researcher would deny the importance of experiments. An experiment involves the planning and execution of a series of procedures to empirically test a hypothesis, essentially constituting the process of verification in empirical science. Therefore, any agent capable of verification must necessarily possess the ability to conduct experiments.

In experiments, phenomena that are difficult to observe, or the effects of various conditions, are precisely investigated. This is achieved by artificially generating phenomena in a controlled manner and actively intervening in them [76]. Such interventions create the different observations of interest. These observations are recorded as experimental data, and subsequent analysis of the data determine the validity of the hypothesis.[7]

To conduct an experiment, one must first design it, document the procedures, and plan its execution. Planning requires an understanding of what constitutes a successful hypothesis test and the ability to devise methods to realize this using existing technology. For example, if the hypothesis pertains to the causes of a particular phenomenon, one must understand what causality is in the first place and how it can be identified in order to plan the verification process.

Preparations for the experiment are also essential. These preparations can include purchasing chemicals, preparing flasks, training animals, applying to ethics committees, constructing necessary equipment, and sometimes even building large apparatus like accelerators from scratch. Unfortunately, since research aims to uncover the unknown, constructing equipment from scratch for experiments is not uncommon in research. The autonomous execution of these preparations by a non-human agent from scratch seems almost infeasible.

After the preparation for the experiment is complete, the experiment is conducted according to the experimental protocol. This task also presents considerable challenges for autonomous machines. The reason is that even a single experiment requires a myriad of low-level operations such as grasping, cutting, carrying, mixing, moving, pouring, dispensing, washing, and opening lids. These operations need to be flexibly combined and executed according to the self-generated experimental protocol. An autonomous machine capable of conducting experiments must possess the ability to generate these operations flexibly in response to the experimental protocol.

3.3.2 Automating Experimentation

As we observe, the challenge of making a machine fully autonomous in planning, preparing, and executing experiments is considerable. Particularly, since the specific experiments to be conducted cannot be determined until questions and hypotheses are formulated, enabling a machine to autonomously conduct research from question construction demands the capability to accommodate many possible experimental scenarios. This, I believe, represents one of the greatest barriers to creating machines capable of autonomously conducting research.

Automating experiments is a daunting task, yet humanity has made steady progress in this area. In relation to the planning stage of experiments, the automation of exploring experimental conditions have a long-standing history, for example. Wang et al. have summarized these studies, which utilize AI to assist in experiment planning, research guidance, and generating observational data through numerical simulations [7].

Furthermore, there is an initiative known as *laboratory automation* or *self-driving lab* that aims to automate experiments, including their execution – an aspect previously mentioned as challenging [77, 78]. A notable example is the research in genetics by King et al., who fully automated the cycle of hypothesis generation, verification, and the discovery of new hypotheses [5]. Another example is the work of A.I. Cooper, which facilitated the use of experimental equipment by autonomous robots, similar to human researchers [79]. These are just a few examples, and there is a vast number of studies in this field.

These examples illustrate efforts to autonomously drive the research cycle, encompassing hypothesis generation, planning and execution of experiments, and generation of new hypotheses based on exper-

[7]Experiments are not conducted solely during the verification phase but also when generating hypotheses. Furthermore, new questions and hypotheses are often formulated based on the results obtained from experiments. In these instances, the process leading up to data generation, or conducting experiments not solely for verification but for data generation and some form of data analysis, seems to be what is referred to as an experiment. This paper defines an experiment as the planning, preparation, data generation, analysis, and determination of verification results. However, be aware that this definition may not always reflect actual practice.

imental results. Such endeavors are referred to as the *closed-loop* automation of scientific discovery [1], representing a significant milestone in achieving high autonomy in research automation. Additionally, there are efforts to develop humanoid robots capable of conducting multiple different experiments with a single robot, considered a foundational step towards more generalized research automation [80].

In recent years, there have been efforts to explore possibilities of autonomous experiment using LLMs [81, 82, 83, 84]. For instance, Boiko et al. developed an autonomous agent comprising multiple LLMs that successfully designed and executed complex scientific experiments [81]. Also, Huang et al. have developed an LLM agent that autonomously designs, executes, and interprets experiments, incorporating the results into machine learning engineering tasks [84].

While we have primarily discussed experimental data generation process, validation also requires interpretation of the data. Observation inherently involves theoretical underpinnings [46]. Hence, interpreting experimental data necessitates adequate prior knowledge. Some studies are focused on enabling machine learning models to interpret scientific data by embedding physical prior knowledge, like symmetries, differential equations, and intuitive physics, into them [85, 86].[8]

Various research efforts have significantly advanced the automation of experiments. However, it is also true that numerous challenges remain in realizing machines capable of autonomously conducting experiments. Coley et al. delve into these challenges in the automation of experimental and computational validations and the selection of experiments, while referring to studies on automated verification [11]. They also point out the significance of removing hardware constraints to increase the automatability of research projects and reducing the costs associated with research automation [11]. Zenil et al. also discuss the challenges of automating experiments and propose specific action plans [6].

Among these challenges, the development of robots capable of manipulating low-level actions as humans do, which is necessary to achieve a versatile automated experimental machine adaptable to diverse research tasks, seems to be exceedingly challenging. How to address these challenges will require further discussion.

4 Additional Topics

4.1 Combining Question Formulation, Hypothesis Generation, and Hypothesis Verification

Reflecting retrospectively on completed studies, it becomes evident that each study possesses its unique question, an accompanying hypothesis, and a process for verifying that hypothesis. Viewed from this perspective, research can be considered an endeavor that involves a sequence of constructing questions, generating hypotheses, and verifying these hypotheses, as classically described.

However, as we know, actual research is a highly complex, cyclical process of trial and error. Rarely do these tasks unfold as initially planned or occur just once in a single study. In practice, for instance, numerous questions and hypotheses might be generated even when formulating a single hypothesis, and not all of these lead to the final research outcome.[9] You will notice that even some major scientific discoveries throughout history were also made through these trials and errors [46, 88, 89].

Therefore, it is more accurate to view the construction of questions, the generation of hypotheses, and the verification of hypotheses as fundamental units for reducing uncertainty. In the research process, they are combined to gradually reduce the vast uncertainty inherent in the research endeavor. AI capable of conducting research is expected to master these flexible and complex operations. In this section, I will speculatively explore these characteristics of real research practice that have not been previously discussed

4.1.1 Countless Questions, Hypotheses, and Verifications in a Single Research Process

In the course of developing a single question or hypothesis, or in planning and preparing for a single verification, we generate countless questions and hypotheses, including implicit ones. Whether it's searching for problems, contemplating why a problem hasn't been solved, considering possible hypoth-

[8] Interpretation of scientific data is not solely for validation purposes. Thus, these technologies extend beyond just automating validation processes.

[9] The trial-and-error nature of activities is particularly significant in the context of discovery [87].

esis candidates, planning verification, or doing anything else, we always pose questions and formulate hypotheses whenever dealing with unknowns or uncertainties.

We also conduct a form of verification, whether implicit or explicit, and with varying degrees of simplicity, to generate plausible hypotheses. Generating plausible hypotheses requires having sufficient grounds to believe in their validity. These grounds could include knowledge from our memory, insights from recently researched literature, opinions from other researchers, or a belief in the simplicity of natural laws. Furthermore, we might conduct simple tests or even preliminary experiments to assess their plausibility. All these function as verification for researchers to be convinced.

Agents capable of conducting research should autonomously generate numerous questions and hypotheses as needed, and select the more plausible hypotheses through simple verifications during the knowledge production process. These are inevitable as long as uncertainties exist. How to realize such flexible agents remains an open question.

4.1.2 Operations Apparently Unrelated to Knowledge Production

Research comprises numerous operations that may initially seem unrelated to knowledge production.[10] In Section 3.3.1, I argue that such tasks are essential in the context of experimentation. Most researchers would concur that daily academic activities are primarily characterized by these operations.

The construction of questions, generation of hypotheses, and verification of these hypotheses represent the core aims and functions in the knowledge production process. To implement these functions, performing operations like those mentioned above and combining them effectively to achieve desired objectives is crucial. Appropriately integrating these varied operations poses a significant challenge, even when tailored to a specific research question [11]. To develop an agent capable of autonomously executing the entire research process, starting from question generation, replicating the flexibility of human action is indispensable.

4.1.3 Discovering New Questions

Researchers often begin with a specific question, only to discover an entirely unrelated question during their investigation. This new question, divergent from the original and its underlying purpose, can lead to a shift in research focus and potentially significant scientific breakthroughs. Given the inherent unpredictability in research, such discovery and redirection of focus are not rare phenomena.

If an agent designed for conducting research were tasked with a singular objective, such serendipitous discoveries might be overlooked. This is because the agent, focused on its predefined goal, may disregard new questions not aligned directly with its initial objective, regardless of their scientific value. To facilitate the agent's identification of such unrelated questions, it may be necessary to assign multiple objectives or a broader, overarching goal that accommodates both the original and emergent questions.

On the other hand, having a common high-level goal alone is not sufficient. For an agent to transition from its current question to a newly discovered one, it must be capable of evaluating which question is more valuable. The decision-making process regarding the value of a question, as discussed in Section 3.1.3, should encompass comparing multiple questions that share higher-order objectives. The development of such evaluative capabilities remains a challenging and open question.

4.1.4 Incorporating Feedback from Verification Result

In research, it is uncommon that the initial hypothesis is the answer of the posed question. Typically, research entails revising the hypothesis based on verification results and conducting subsequent rounds of testing. This iterative cycle of hypothesis revision and retesting is critical for scientific discovery. Therefore, an agent designed for conducting research should possess the ability to revise its hypotheses based on these outcomes of verification.

Efforts to automate incorporating feedback from verification results include studies in closed-cycle laboratory automation, as discussed in Section 3.3 and automation with scientific workflow [91]. These have significantly contributed to automating the hypothesis revision cycle.

Despite these advancements, challenges persist in developing machines that autonomously analyze and respond to verification results like humans. When verification results are negative, pinpointing

[10] Latour's anthropological study of daily practices in laboratories aptly illustrates these realities [90].

the exact cause is complex. This complexity arises because verification relies on a web of implicit and explicit hypotheses, any of which could contribute to the result [92]. The cause might be the primary hypothesis, underlying premises, auxiliary hypotheses, observations, experimental instruments, or a combination of these. A research-conducting agent must be capable of discerning the likely cause among these numerous candidates. Although it seems that humans do this well [93], it is a challenging task for machines to do this autonomously.

Moreover, appropriate interpretation of experimental data is essential. Interpretations can vary based on the researcher's beliefs, prior knowledge, theoretical framework, and expectations [46]. Thus, verification results may undergo multiple reinterpretations, each potentially altering the hypothesis in need of revision. Additionally, as previously mentioned, researchers sometimes derive entirely different questions from these results and may temporarily halt their research. Current machines have yet to match this level of complex interpretation and adaptability in handling verification results that humans exhibit. An ideal autonomous research agent would be expected to possess these capabilities.

4.2 Common Topics

In the preceding sections, I have speculatively examined various key elements integral to research and their interplay. In this section, I aim to delve into topics that are universally relevant to these elements or that have not yet been explored in this paper.

4.2.1 Language Models

Recent years have witnessed rapid advancements in language models [94], opening new frontiers in research. The future of research agents is undoubtedly intertwined with the insights provided by these models. This section explores various initiatives investigating the potential of language models in research.

Beginning with the transformative impact of the Transformer [95] and BERT [96], the concept of "scaling law" [97] and "foundation model" [98] has catalyzed the development of large scale models pre-trained on extensive corpora. This has led to the creation of scientific language models like SciBERT [99] and others [100, 101, 102, 103, 104, 105, 106, 107, 108, 109]. Given that scientific data is multimodal, attempts are emerging to construct general-purpose models using multimodal data as well [110, 111, 112, 113].

The development of GPTs, including GPT-3 [114], InstructGPT [115], and GPT-4 [116] and the advent of the web application called ChatGPT [117] marked significant milestones. Their ability to perform various intellectual tasks has spurred research into their scientific applications. Emerging research examines the potential of LLMs, especially these GPTs, in various research fields, including the natural sciences [10, 81, 82, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128], mathematics [129, 65, 130, 66], engineering [131, 132, 84, 133, 134, 135, 136], and social sciences [72, 137, 138, 139, 140, 141, 142, 143].

Additionally, there are efforts applicable to all research fields, which involve using GPTs for processing academic documents [144]. Some of them include paper search and reading [145, 146], paper writing [147], abstract generation [148], literature review generation [149], and peer review [150, 151, 152].[11]

The scope of applications of LLMs ranges from generating hypotheses to generating research questions finding research challenges [32, 24, 33]. Some studies even attempted to automate experimentation [81, 82].

It is important to recognize that research and development in the field of language models are advancing rapidly. The potential applications for automating research are vast and constantly evolving, warranting further exploration and critical assessment of their impact on the scientific method.

4.2.2 Incorporating Scientific Knowledge

While discussing the challenges of enabling machines to conduct research, it's pertinent to acknowledge that even humans do not embark on research from a completely zero starting point. Firstly, humans are inherently equipped with brains and bodies evolved and developed for interpreting the world. Moreover, before engaging in research, we study the fundamental knowledge of the research fields. Thus, it would be reasonable to assume that agents should also possess basic knowledge before they

[11] Hosseini and Horbach discuss the influence of LLMs on the role of peer review [153].

begin conducting research. The advancement of language models has brought a wealth of knowledge to machines, but it is still considered necessary for them to acquire the knowledge required for research.

This is embodied in the concept of *physics-informed machine learning* [86], where biases and scientific knowledge are integrated into AI to process scientific data. Karniadakis et al. [86] and Hao et al. [85] provide a systematic overview of research in this domain. As previously discussed, imparting scientific knowledge through training on textual and multi-modal data is also a prevalent strategy.

Furthermore, just as humans continuously update their scientific knowledge, it's imperative for machines to not only embed knowledge through pre-training and inductive biases or retrieval during inference but also to continually update this knowledge. Specifically, it is crucial to acknowledge that knowledge produced by research is perpetually evolving. Thus, an agent must assimilate new knowledge, retain and update existing knowledge, and adapt to revisions in previously learned concepts. While the methodology for achieving this remains an open question, it is a subject of increasing debate [2, 1].

4.2.3 Autonomy, Generality, and Open-Endedness

As previously emphasized, ongoing efforts are being made to enable machines to autonomously generate research questions, formulate hypotheses, and validate these hypotheses. However, the significant challenge remains in achieving this autonomy with minimal human intervention. Even in the realm of closed-loop research automation, which represents a substantial stride towards autonomy, full automation of all research processes is still an unrealized goal [6, 12, 11].

This problem becomes particularly serious when attempting to enable machines to independently formulate research objectives, problems, and questions. If machines are to autonomously generate goals and questions, they must also be capable of independently generating and validating corresponding hypotheses. This necessitates a versatile approach to hypothesis generation and validation, adaptable to a wide range of questions.

In such scenarios, humans cannot provide predefined methods, potential hypotheses, or necessary information. Consequently, machines must be equipped to extract pertinent information from an open-ended environment, mirroring the human approach. Research, in essence, is a process of seeking information from the vast outer world of scientific data and processing it within the agent's cognitive framework, as highlighted in [13]. Assuming an open-ended environment means minimizing human-imposed constraints on this outer world and allowing the agent maximum freedom in selecting and processing information from the outer world.

Given these considerations, the extent to which autonomy should be expected from machines and the level of constraints that can be imposed without stifling their potential for autonomous hypothesis generation and validation, remains a critical and open question.

4.2.4 Scientific Understanding

While the primary focus of this paper has been on knowledge discovery, it's crucial to also consider another important goal of research: understanding. As Krenn et al. highlight, scientific discoveries can be made without understanding [154], suggesting that facilitating scientific understanding in humans by automated machines demands more than the things discussed so far.

Scientific understanding in humans involves comprehending theories or hypotheses – both their nature and their underlying rationale. Therefore, an additional requirement seems necessary in the context of hypothesis generation. It remains unclear whether this pertains to the representation of generated hypotheses, the description of their generation process, or anything else. Identifying what is needed to be added in this process to foster scientific understanding, and how to implement it, is a significant issue.

Krenn et al. propose two conditions for an AI to achieve scientific understanding: 1. the ability to "recognize qualitatively characteristic consequences of a theory without performing exact computations and use them in a new context", and 2. the capacity to "transfer its understanding to a human expert" [154]. As previously mentioned, the belief systems of humans and AI may differ. Thus whether AI having scientific understanding is necessary for bringing new scientific understanding to humans is unclear. Nonetheless, the capability to communicate understanding to humans is essential for bringing scientific understanding to humans. The explanation of machine prediction results has already been

extensively researched and discussed as explainable AI [155], and its importance has already been widely pointed out in AI for Science research, so I will not delve further into it here.[12]

4.2.5 Alignment

Alignment is a critical concern in the development of autonomous AI researchers, akin to other areas of AI research. The primary concern is ensuring that these autonomous agents do not harm humans, a priority that becomes increasingly significant as we seek greater autonomy in machines. Given that knowledge is inherently value-neutral and can be used for benevolent or malevolent purposes, addressing this challenge is complex and necessitates ongoing discourse.

As previously discussed in Section 3.1, alignment with human values and worldviews is crucial not only for safety but also for the relevance and effectiveness of AI-generated knowledge for humans. AI agents need to make value judgments aligned with human assessments of question quality and discern what is unknown and comprehensible from a human perspective, not just from their own standpoint.

These value judgments are often not explicitly stated in human-generated texts, indicating a need for proactive teaching of these values to AI. The methodology for implementing such teaching and ensuring alignment in a broader sense remains a complex issue that warrants further discussion and exploration.

This topic's complexity is amplified by the varying and evolving nature of human values and worldviews. As AI continues to advance, the challenge of continuously adapting these systems to align with human ethics and understanding becomes more pronounced. Future discussions should focus on developing robust frameworks and methodologies to achieve and maintain this alignment, ensuring that autonomous AI researchers contribute positively and safely to the scientific community.

5 Ideas for Prototyping

Realizing an autonomous intelligent agent capable of conducting research is an exceptionally challenging goal, one that will likely require a significant amount of time to achieve. The challenges discussed so far represent merely the tip of the iceberg found in speculative discussions; undoubtedly, many more critical issues remain unidentified. Therefore, it is crucial to begin by identifying these unknown challenges. A practical starting point might be the development of a simplified prototype of a research-capable agent. Such a prototype would allow us to explore and understand the challenges inherent to our goal during the prototyping process. In this section, I aim to discuss, in a speculative and brief manner, what might constitute such prototyping.

5.1 Prototyping Agents that Conduct Research

5.1.1 Requirements for Prototype

As discussed in Section 3, Research seems to involve constructing questions, generating hypotheses, and verifying these hypotheses. Therefore, it appears appropriate for this prototype to incorporate these functions as distinct modules. The "question construction" module should take any input and formulate a question. The "hypothesis generation" module would then take this question as input and generate a hypothesis. Subsequently, the "hypothesis verification" module would take the hypothesis and provide verification results. By flexibly combining these modules at various levels, the prototype could mimic the research process.

For the prototype agent to function autonomously, human involvement in its design, implementation, and intervention should be minimized. Consequently, each module should autonomously gather information from the open-ended world, similar to how humans acquire information for research, while requiring minimal inputs. This means the agent should interact with the physical or digital realms to gather necessary information for research.

Moreover, to ensure the system's generality, the internal workings of each module should not overly rely on specific research topics. For instance, a verification method like experimentation tailored

[12] In considering how machine-driven scientific discoveries can facilitate human understanding, it might be worth exploring not only the enhancement of machine capabilities but also the expansion of human cognitive boundaries. While delving into this might border on speculative, it's a discussion that could yield valuable insights in the scientific community.

for specific physics research would not be applicable to psychological research. The human-designed elements of each module should be minimal, confined to what is essential for the module's function.

Creating a system that simultaneously meets the criteria of autonomy and generality, while effectively constructing questions, generating hypotheses, and verifying them within this abstract framework, is impractical, even in simpler scenarios. Thus, it may be necessary to introduce some constraints to this abstract framework. Discussing the nature, necessity, and potential relaxation of these constraints could shed light on the challenges in realizing an autonomous research agent. To initiate this discussion, I will present some candidates for potential constraints.

5.1.2 Candidate Constraints in Prototyping

In Section 2, I discussed the perspective of research as a process of updating beliefs and the potential for autonomously constructing verification from foundational concepts, as well as autonomously assessing the value of questions. However, these concepts are visionary and present significant challenges, making it unrealistic to expect immediate, meaningful outcomes for humans through prototyping. Therefore, it would be beneficial to start by prototyping agents that can master existing human values and research methods.

As mentioned in Section 3.1, formulating questions from open-ended situations is an exceptionally challenging task, often with even no clear starting point. A pragmatic approach would be to predetermine the inputs for question construction, rather than relying on unrestricted information sources. A viable input could be a high-level goal, commonly assumed in many studies. Specifically, it would be beneficial to start with high-level goals recognized as research objectives in specific research fields.

A significant obstacle in developing a fully autonomous research agent, as discussed in Section 4.1.2, is the need for expertise in complex low-level actions. Developing a robot capable of free physical world interaction like humans remains a formidable challenge. Hence, for prototyping, focusing initially on research confined to computational environments appears more manageable. While creating an agent capable of operating freely within a computer environment is also challenging, it is arguably more feasible than one operating in the physical world. Indeed, there have been efforts to enable language models to perform various computer operations [156, 157], and to operate web browsers [158, 159].

The primary objective of prototyping is to concretize a concept, however rudimentary, and to identify challenges. Thus, it seems prudent to initially limit the prototype's environment to the digital realm, while waiting for advancements in foundational research that could enable free activity in the physical world.

The ideas presented here are merely initial suggestions and are neither definitive nor exhaustive. In the prototyping phase, it is crucial to discuss the extent and nature of the constraints to be applied. More suitable constraints are likely to emerge as these discussions evolve.

5.1.3 Implementing Each Module with Large Language Models

Given the need for generality and considering the remarkable capabilities of LLMs, it seems inevitable that each module in prototypes would be instantiated as an LLM. As highlighted in previous sections, there are emerging studies focused on constructing automated research pipelines using LLMs. I propose that our initial prototyping efforts should focus on creating autonomous research agents modeled on these LLM pipelines, in alignment with current endeavors to develop autonomous agents utilizing language models [137, 160].

Here is a provisional concept, modeled after a typical autonomous agent. The research agent begins by formulating a question based on a high-level goal provided by a human. Following the posing of this question, the agent autonomously generates hypotheses to address it, and then proceeds to verify these hypotheses. After obtaining the final verification results, they are analyzed in relation to the initial objective and research question, prompting the generation of subsequent questions. This process – comprising question formulation, hypothesis generation, and hypothesis verification – is iteratively and hierarchically repeated to conduct research.

The agent is envisioned to perform four fundamental actions: 1) Formulating questions, 2) Determining task completion, 3) Verifying hypotheses, and 4) Executing low-level computer operations. The processes of question formulation, hypothesis generation, and verification primarily involve executing these low-level computer operations.

When the agent opts to generate a question, it temporarily pauses its current task, such as hypothesis verification, and initiates hypothesis generation for the new question. Upon completing hypothesis generation, the agent decides whether to proceed with verification. Following verification, it updates the hypotheses based on the results. Whether or not the hypotheses are verified, the agent then resumes the higher-level process that was previously paused, incorporating the results of the low-level process. In this way, the agent continuously cycles through lower-level tasks of question construction, hypothesis generation, and verification until the highest-level hypothesis is formulated. When the highest-level hypothesis – the response to the original question – is ready, the agent always proceeds to its verification.

To ensure the system's adaptability to a wide range of research questions, the prompts given to the LLMs should be composed of only general instructions. For example, an instruction like "generate a hypothesis for the following question" is sufficiently generic to apply to any research question. However, providing such instructions alone is unlikely to spontaneously yield research outcomes, so there may be a need for additional auxiliary instructions that are as general as possible; identifying these is one of the main goals of prototyping.

For open-ended operations within a computer environment, ideally, the LLMs should have access to nearly all operations on the computer. As previously mentioned, initiatives to develop language models capable of executing any action in such environments are underway [157, 156]. Minimal access to web browsers, search engines, or shells may be permissible, but reliance on custom corpora or predefined hypothesis spaces should be avoided. Successful autonomous research under these conditions would indeed demonstrate the system's capacity for independent research.

5.1.4 Agents that Conduct Machine Learning Research

To effectively provide a high-level goal for the prototype, it is necessary to select objectives from a specific research field that align with the constraints previously outlined and are conducive to prototyping.

I propose that machine learning research is an ideal candidate for such prototyping. First, many aspects of machine learning research, including verification, can be conducted entirely on a computer, thus meeting the constraints we have established. Second, the field typically features shorter research cycles compared to other disciplines, which allows for more rapid feedback for the prototype. Third, machine learning not only forms a foundational technology across various research fields but is crucial for developing a research-capable agent itself. Automating machine learning research would thus not only contribute to the automation of research processes in numerous other areas but also advance our primary objective. Finally, there already have been significant efforts towards automation in machine learning, such as AutoML [161, 162, 163, 164] and MLOps [165]. Particularly in recent years, there have been attempts to utilize language models for these tasks [134, 131, 135, 136]. These existing efforts are likely to provide valuable support in developing the prototype.

An emerging noteworthy study in this direction of research is that by Huang et al [84]. This study attempts to have LLM agents autonomously perform machine learning engineering tasks, including those on Kaggle. Particularly noteworthy is that they have the agents autonomously plan, execute, and interpret experiments, and connect them to subsequent experiments. They accomplish this by having the agents choose actions that are versatile, such as reading and writing files, and executing tasks. This research will likely become a foundational study for developing agents capable of autonomously conducting a wide range of machine learning research in the future.

In conclusion, initiating the prototyping of autonomous agents, composed of language models and given the most general instructions possible, and focusing on specific types of machine learning research appears to be a strategic choice. While such efforts are already underway, I anticipate that increased participation in this area will significantly accelerate this movement.

5.2 Prototyping Agents that Conduct Peer Review

To identify challenges associated with creating agents capable of conducting research, another promising initial step could be to target the automation of the academic peer review process. This approach presents several advantages, which I will detail in the following section.

5.2.1 Why Aim for Agents Capable of Conducting Peer Review?

Firstly, the competencies necessary for peer review closely align with those required by a research-capable agent. This similarity arises because peer review fundamentally involves evaluating critical aspects of research, such as the soundness of the verification.

Secondly, automating peer review might present fewer challenges compared to developing a fully autonomous research agent. The distinction lies in the scope of tasks: peer review primarily entails assessing whether research incorporates the necessary elements, whereas a research-capable agent must not only evaluate but also synthesize these elements. As a preliminary step in prototyping, addressing simpler problems like peer review could effectively highlight key challenges.

Thirdly, peer review is a universal practice across various research fields. Insights gained from automating this process can thus contribute significantly to the development of a general research agent, applicable in multiple disciplines.

Fourthly, peer review predominantly involves textual analysis and does not require physical or extensive digital interactions, unlike conducting research autonomously. Although it may involve searches to review existing literature, tasks such as performing experiments are typically not necessary. Given the advancements in LLMs, we are now better equipped to handle complex textual tasks. This focus on text-based evaluation is beneficial for pinpointing specific challenges in achieving our broader objective.

Finally, peer review encompasses the evaluation of subjective aspects like the "significance" of a research question. As discussed in Section 3.1, understanding how humans assess such value in research is crucial, especially considering that alignment with human values is a significant challenge. Peer review offers a unique opportunity to observe and analyze these value judgments explicitly. Therefore, beginning with the automation of peer reviews could provide valuable insights into human evaluative processes in research.

5.2.2 Peer Review Automation

A considerable body of research has been devoted to automating various aspects of the peer review process. Efforts have included automating the generation of reviews [166, 167, 168], screening papers [169], assessing research papers [74], and assigning reviewers [170], among other tasks. In line with trends across other fields, recent years have witnessed a surge in studies exploring the use of LLMs for automating peer review [150, 151, 152, 153]. For a more comprehensive understanding of traditional research in this area, Kousha et al. [74] and Lin et al. [75] have conducted extensive literature reviews.

Considering the goals of prototyping, it is desirable that such efforts already exist. The insights and findings from these prior attempts would help further discussions on developing AI capable of conducting peer review.

6 Conclusion

In this paper, I have undertaken a speculative exploration of the concept of an artificial agent capable of conducting research. The initial discussion centered on characterizing what constitutes research, tentatively framing it as the process of updating beliefs in hypotheses. Subsequently, I delved into the critical elements of research: the construction of questions, generation of hypotheses, and their verification. This paper then briefly looks at the common themes to these elements. Following the discussions, I highlighted the significance of identifying challenges in realizing such agents and proposed preliminary ideas for prototyping.

It is important to acknowledge that the discussions in this paper are purely speculative. The definition of research provided is provisional, the challenges and implications discussed represent only a fraction of the myriad possibilities, and the ideas for prototyping are rudimentary, akin to early-stage experiments. Furthermore, the literature referenced is not exhaustive, omitting many pivotal works. My ability to evaluate each reference thoroughly may have been limited, potentially leading to partial perspectives or inaccuracies. I plan to update this paper in the future to address these limitations. I greatly value feedback and corrections from readers, as they will be crucial in improving this work.

The primary motivation for publishing this paper in its current nascent form, despite its numerous limitations, is to serve as an initial step towards future exploration into the concept of a research-capable artificial agent. In order to realize agents capable of conducting research, there must still be

many issues that need to be discussed. I hope that the discussion surrounding this concept will become more active to accelerate the development of research-capable agents.

References

[1] Hector Zenil, Jesper Tegnér, Felipe S Abrahão, Alexander Lavin, Vipin Kumar, Jeremy G Frey, Adrian Weller, Larisa Soldatova, Alan R Bundy, Nicholas R Jennings, et al. The future of fundamental science led by generative closed-loop artificial intelligence. *arXiv preprint arXiv:2307.07522*, 2023.

[2] Hiroaki Kitano. Nobel turing challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications*, 7(1):29, 2021.

[3] Robert K Lindsay, Bruce G Buchanan, Edward A Feigenbaum, and Joshua Lederberg. Dendral: a case study of the first expert system for scientific hypothesis formation. *Artificial intelligence*, 61(2):209–261, 1993.

[4] Pat Langley. *Scientific discovery: Computational explorations of the creative processes*. MIT press, 1987.

[5] Ross D King, Kenneth E Whelan, Ffion M Jones, Philip GK Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.

[6] Hector Zenil and Ross D. King. *The Automated AI-driven Future of Scientific Discovery*, pages 679–691. World Scientific, 2023.

[7] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[8] Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, et al. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4):100179, 2021.

[9] Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.

[10] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*, 2023.

[11] Connor W Coley, Natalie S Eyke, and Klavs F Jensen. Autonomous discovery in the chemical sciences part ii: outlook. *Angewandte Chemie International Edition*, 59(52):23414–23436, 2020.

[12] Connor W Coley, Natalie S Eyke, and Klavs F Jensen. Autonomous discovery in the chemical sciences part i: Progress. *Angewandte Chemie International Edition*, 59(51):22858–22893, 2020.

[13] Tom Hope, Doug Downey, Oren Etzioni, Daniel S Weld, and Eric Horvitz. A computational inflection for scientific discovery. *arXiv preprint arXiv:2205.02007*, 2022.

[14] National Academies of Sciences Engineering, Medicine, et al. *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop*. The National Academies Press, 2022.

[15] Alan F Chalmers. *What is this thing called science?* Hackett Publishing, 2013.

[16] Brian Hepburn and Hanne Andersen. Scientific Method. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

[17] Wayne C Booth, Gregory G Colomb, and Joseph M Williams. *The craft of research*. University of Chicago press, 2003.

[18] Frascati Manual et al. Guidelines for collecting and reporting data on research and experimental development. *URL: http://www. oecd. org/sti/frascati-manual-2015-9789264239012-en. htm*, 2015.

[19] Francesco Balzan-francesco, John Campbell, Karl Friston, Maxwell James Ramstead, Daniel Friedman, and Axel Constant. Distributed science-the scientific process as multi-scale active inference. *OSF Preprints*, 2023.

[20] Matthias Steup and Ram Neta. Epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

[21] Edmund L Gettier. Is justified true belief knowledge? *analysis*, 23(6):121–123, 1963.

[22] Jun Otsuka. *Thinking About Statistics: The Philosophical Foundations*. Taylor & Francis, 2022.

[23] Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S Weld, et al. A search engine for discovery of scientific challenges and directions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11982–11990, 2022.

[24] Jonas Oppenlaender and Joonas Hämäläinen. Mapping the challenges of hci: An application and evaluation of chatgpt and gpt-4 for cost-efficient question answering. *arXiv preprint arXiv:2306.05036*, 2023.

[25] Gabriela Surita, Rodrigo Nogueira, and Roberto Lotufo. Can questions summarize a corpus? using question generation for characterizing covid-19 research. *arXiv preprint arXiv:2009.09290*, 2020.

[26] Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. Paperrobot: Incremental draft generation of scientific ideas. *arXiv preprint arXiv:1905.07870*, 2019.

[27] Mario Krenn and Anton Zeilinger. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences*, 117(4):1910–1916, 2020.

[28] Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, Joao P Moutinho, Nima Sanjabi, et al. Predicting the future of ai with ai: High-quality link prediction in an exponentially growing knowledge network. *arXiv preprint arXiv:2210.00881*, 2022.

[29] Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*, 2019.

[30] Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43, 2021.

[31] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

[32] Yiren Liu, Mengxia Yu, Meng Jiang, and Yun Huang. Creative research question generation for human-computer interaction research. In *Joint Proceedings of the ACM IUI Workshop*, 2023.

[33] Adi Lahat, Eyal Shachar, Benjamin Avidan, Zina Shatz, Benjamin S Glicksberg, and Eyal Klang. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Scientific reports*, 13(1):4164, 2023.

[34] Lani Watson. What is a question. *Royal Institute of Philosophy Supplements*, 89:273–297, 2021.

[35] Robert S Taylor. The process of asking questions. *American documentation*, 13(4):391–396, 1962.

[36] Tom D Wilson. Information behaviour: an interdisciplinary perspective. *Information processing & management*, 33(4):551–572, 1997.

[37] Donald O Case and Lisa M Given. *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald Group Publishing, 2016.

[38] Celeste Kidd and Benjamin Y Hayden. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460, 2015.

[39] Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

[40] Mats Alvesson and Jorgen Sandberg. *Constructing research questions: Doing interesting research*. Sage, 2013.

[41] Stephen B Hulley. *Designing clinical research*. Lippincott Williams & Wilkins, 2007.

[42] Uri Alon. How to choose a good scientific problem. *Molecular cell*, 35(6):726–728, 2009.

[43] Nick Huntington-Klein. *The effect: An introduction to research design and causality*. CRC Press, 2021.

[44] Pierre-Yves Oudeyer. Computational theories of curiosity-driven learning. *arXiv preprint arXiv:1802.10546*, 2018.

[45] Jutta Schickore. Scientific Discovery. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

[46] Norwood Russell Hanson. *Patterns of discovery: An inquiry into the conceptual foundations of science*. CUP Archive, 1965.

[47] Lorenzo Magnani. *Abduction, reason and science: Processes of discovery and explanation*. Springer Science & Business Media, 2011.

[48] Dedre Gentner. Analogy in scientific discovery: The case of johannes kepler. In *Model-based reasoning*, pages 21–39. Springer, 2002.

[49] Ishita Dasgupta, Eric Schulz, and Samuel J Gershman. Where do hypotheses come from? *Cognitive psychology*, 96:1–25, 2017.

[50] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[51] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 2023.

[52] Stefan Kramer, Mattia Cerrato, Sašo Džeroski, and Ross King. Automated scientific discovery: From equation discovery to autonomous discovery systems. *arXiv preprint arXiv:2305.02251*, 2023.

[53] Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction*, 2022.

[54] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21, 2018.

[55] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Learning to generate novel scientific directions with contextualized literature-based discovery. *arXiv preprint arXiv:2305.14259*, 2023.

[56] Yi Xu, Shuqian Sheng, Bo Xue, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. Exploring and verbalizing academic ideas by concept co-occurrence. *arXiv preprint arXiv:2306.02282*, 2023.

[57] Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*, 2023.

[58] Yang Jeong Park, Daniel Kaplan, Zhichu Ren, Chia-Wei Hsu, Changhao Li, Haowei Xu, Sipei Li, and Ju Li. Can chatgpt be used to generate scientific hypotheses? *arXiv preprint arXiv:2304.12208*, 2023.

[59] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[60] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[61] M Burton David. *The history of mathematics an introduction*. McGraw-Hill Professional, 2010.

[62] Salomon Bochner and Banesh Hoffmann. The role of mathematics in the rise of science. *American Journal of Physics*, 36(6):564–565, 1968.

[63] Werner Heisenberg. Abstraction in modern science. *Nishina Memorial Lectures*, pages 1–16, 2008.

[64] Markus N Rabe and Christian Szegedy. Towards the automatic mathematician. In *Automated Deduction–CADE 28: 28th International Conference on Automated Deduction, Virtual Event, July 12–15, 2021, Proceedings 28*, pages 25–37. Springer International Publishing, 2021.

[65] Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, Aimin Zhou, and Liang He. Mathematical language models: A survey. *arXiv preprint arXiv:2312.07622*, 2023.

[66] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexey Novikov, et al. Mathematical discoveries from program search with large language models. *Nature*, 2023.

[67] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

[68] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

[69] N Baker et al. Basic research needs workshop for scientific machine learning: Core technologies for artificial intelligence. *Document prepared for Department of Energy Advanced Scientific Computing Research, USA*, 10, 2019.

[70] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.

[71] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[72] Sai Koneru, Jian Wu, and Sarah Rajtmajer. Can large language models discern evidence for scientific hypotheses? case studies in the social sciences. *arXiv preprint arXiv:2309.06578*, 2023.

[73] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

[74] Kayvan Kousha and Mike Thelwall. Artificial intelligence technologies to support research assessment: A review. *arXiv preprint arXiv:2212.06574*, 2022.

[75] Jialiang Lin, Jiaxin Song, Zhangping Zhou, and Xiaodong Shi. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98, 2023.

[76] Hans Radder. The philosophy of scientific experimentation: a review. *Automated experimentation*, 1(1):1–8, 2009.

[77] Ian Holland and Jamie A Davies. Automation in the life science research laboratory. *Frontiers in Bioengineering and Biotechnology*, 8:571777, 2020.

[78] Milad Abolhasani and Eugenia Kumacheva. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, pages 1–10, 2023.

[79] Benjamin Burger, Phillip M Maffettone, Vladimir V Gusev, Catherine M Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M Alston, Buyi Li, Rob Clowes, et al. A mobile robotic chemist. *Nature*, 583(7815):237–241, 2020.

[80] Nozomu Yachie and Tohru Natsume. Robotic crowd biology with maholo labdroids. *Nature biotechnology*, 35(4):310–312, 2017.

[81] Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

[82] Xiaokai Qin, Mingda Song, Yangguan Chen, Zhehong Ai, and Jing Jiang. Gpt-lab: Next generation of optimal chemistry discovery by gpt driven robotic lab. *arXiv preprint arXiv:2309.16721*, 2023.

[83] Gary Charness, Brian Jabarian, and John A List. Generation next: Experimentation with ai. Technical report, National Bureau of Economic Research, 2023.

[84] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Benchmarking large language models as ai research agents. *arXiv preprint arXiv:2310.03302*, 2023.

[85] Zhongkai Hao, Songming Liu, Yichi Zhang, Chengyang Ying, Yao Feng, Hang Su, and Jun Zhu. Physics-informed machine learning: A survey on problems, methods and applications. *arXiv preprint arXiv:2211.08064*, 2022.

[86] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[87] Itai Yanai and Martin Lercher. A hypothesis is a liability, 2020.

[88] John Gribbin and Mary Gribbin. *On The Origin of Evolution: Tracing 'Darwin' s Dangerous Idea' from Aristotle to DNA*. Rowman & Littlefield, 2022.

[89] Derek Thomas Whiteside. Before the principia: The maturing of newton's thoughts on dynamical astronomy, 1664–1684. *Journal for the History of Astronomy*, 1(1):5–19, 1970.

[90] Bruno Latour. *Science in action: How to follow scientists and engineers through society*. Harvard university press, 1987.

[91] Yolanda Gil. Will ai write scientific papers in the future? *AI Magazine*, 42(4):3–15, 2022.

[92] Kyle Stanford. Underdetermination of Scientific Theory. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition, 2023.

[93] Zhichu Ren, Zekun Ren, Zhen Zhang, Tonio Buonassisi, and Ju Li. Autonomous experiments using active learning and ai. *Nature Reviews Materials*, 8(9):563–564, 2023.

[94] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[95] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[96] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[97] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[98] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[99] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[100] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020.

[101] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*, 2022.

[102] Rahul Nadkarni, David Wadden, Iz Beltagy, Noah A Smith, Hannaneh Hajishirzi, and Tom Hope. Scientific language models for biomedical knowledge base completion: an empirical study. *arXiv preprint arXiv:2106.09700*, 2021.

[103] Tanishq Gupta, Mohd Zaki, and NM Anoop Krishnan. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.

[104] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[105] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.06786*, 2023.

[106] Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, et al. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*, 2023.

[107] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

[108] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

[109] Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Le Zhou, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. Learning a foundation language model for geoscience knowledge understanding and utilization. *arXiv preprint arXiv:2306.05064*, 2023.

[110] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[111] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023.

[112] Seiji Takeda, Akihiro Kishimoto, Lisa Hamada, Daiju Nakano, and John R Smith. Foundation model for material science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15376–15383, 2023.

[113] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[114] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[115] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[116] OpenAI. Gpt-4. https://openai.com/research/gpt-4, 2023. Version of the Generative Pre-trained Transformer.

[117] OpenAI. Chatgpt. https://openai.com/chatgpt, 2023. Software available from https://openai.com/chatgpt.

[118] Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[119] Andrew D White, Glen M Hocky, Heta A Gandhi, Mehrad Ansari, Sam Cox, Geemi P Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, et al. Do large language models know chemistry? *ChemRxiv*, 2022.

[120] Kan Hatakeyama-Sato, Naoki Yamane, Yasuhiko Igarashi, Yuta Nabae, and Teruaki Hayakawa. Prompt engineering of gpt-4 for chemical research: what can/cannot be done? *ChemRxiv*, 2023.

[121] Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023.

[122] Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[123] Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*, 2023.

[124] Chen Qian, Huayi Tang, Zhirui Yang, Hong Liang, and Yong Liu. Can large language models empower molecular property prediction? *arXiv preprint arXiv:2307.07443*, 2023.

[125] Magdalena Wysocka, Oskar Wysocki, Maxime Delmas, Vincent Mutel, and Andre Freitas. Large language models, scientific knowledge and factuality: A systematic analysis in antibiotic discovery. *arXiv preprint arXiv:2305.17819*, 2023.

[126] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[127] Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*, 2023.

[128] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[129] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.

[130] Yiran Wu, Feiran Jia, Shaokun Zhang, Qingyun Wu, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, and Chi Wang. An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337*, 2023.

[131] Shujian Zhang, Chengyue Gong, Lemeng Wu, Xingchao Liu, and Mingyuan Zhou. Automl-gpt: Automatic machine learning with gpt. *arXiv preprint arXiv:2305.02499*, 2023.

[132] Sebastian Bordt and Ulrike von Luxburg. Chatgpt participates in a computer science exam. *arXiv preprint arXiv:2303.09461*, 2023.

[133] Vinay Pursnani, Yusuf Sermet, Musa Kurt, and Ibrahim Demir. Performance of chatgpt on the us fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence*, page 100183, 2023.

[134] Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970*, 2023.

[135] Vijay Viswanathan, Chenyang Zhao, Amanda Bertsch, Tongshuang Wu, and Graham Neubig. Prompt2model: Generating deployable models from natural language instructions. *arXiv preprint arXiv:2308.12261*, 2023.

[136] Noah Hollmann, Samuel Müller, and Frank Hutter. Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[137] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023.

[138] Christopher A Bail. Can generative ai improve social science? *SocArXiv*, 2023.

[139] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.

[140] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[141] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

[142] Anton Korinek. Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4), 2023.

[143] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[144] Adhari AlZaabi, Amira ALamri, Halima Albalushi, Ruqaya Aljabri, and Abdulrahman Aal-Abdulsallam. Chatgpt applications in academic research: A review of benefits, concerns, and recommendations. *bioRxiv*, pages 2023–08, 2023.

[145] Elicit. https://elicit.org/. Accessed on 2023-04-06.

[146] SCISPACE. https://scispace.com/. Accessed on 2023-04-06.

[147] Gpt Generative Pretrained Transformer, Almira Osmanovic Thunström, and Steinn Steingrimsson. Can gpt-3 write an academic paper on itself, with minimal human input? *HAL Open Science*, 2022.

[148] Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, 6(1):75, 2023.

[149] Ömer Aydın and Enis Karaarslan. Openai chatgpt generated literature review: Digital twin in healthcare. *Available at SSRN 4308687*, 2022.

[150] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. In *arXiv preprint arXiv:2310.01783*, 2023.

[151] Ryan Liu and Nihar B Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.

[152] Zachary Robertson. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*, 2023.

[153] Mohammad Hosseini and Serge PJM Horbach. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8(1):4, 2023.

[154] Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, et al. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769, 2022.

[155] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[156] KillianLucas. Open interpreter, 2023. Accessed: 2023-09-24, License: MIT.

[157] OpenAI. Chatgpt plugins - code interpreter. https://openai.com/blog/chatgpt-plugins#code-interpreter, 2023. Accessed: 2023-12-03.

[158] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[159] Adept. Act-1: Transformer for actions, 2022. Accessed: 2023-09-25.

[160] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[161] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[162] Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix, et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1484, 2023.

[163] Marius Lindauer and Frank Hutter. Best practices for scientific research on neural architecture search. *The Journal of Machine Learning Research*, 21(1):9820–9837, 2020.

[164] Colin White, Mahmoud Safari, Rhea Sukthanker, Binxin Ru, Thomas Elsken, Arber Zela, Debadeepta Dey, and Frank Hutter. Neural architecture search: Insights from 1000 papers. *arXiv preprint arXiv:2301.08727*, 2023.

[165] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access*, 2023.

[166] Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.

[167] Weizhe Yuan and Pengfei Liu. Kid-review: Knowledge-guided scientific review generation with oracle pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11639–11647, 2022.

[168] Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. Reviewrobot: Explainable paper review generation based on knowledge synthesis. *arXiv preprint arXiv:2010.06119*, 2020.

[169] Robert Schulz, Adrian Barnett, René Bernard, Nicholas JL Brown, Jennifer A Byrne, Peter Eckmann, Małgorzata A Gazda, Halil Kilicoglu, Eric M Prager, Maia Salholz-Hillel, et al. Is the future of peer review automated? *BMC Research Notes*, 15(1):1–5, 2022.

[170] Xiquan Zhao and Yangsen Zhang. Reviewer assignment algorithms for peer review automation: A survey. *Information Processing & Management*, 59(5):103028, 2022.