Philosophy of Autonomous Science: Ten Questions for the Coming Age of Artificial Scientists

Mario Krenn & Heather Champion

Artificial intelligence is beginning to make remarkable contributions in science, from protein design to materials discovery and experiment design. Yet we believe that moving from powerful tools to autonomous scientists is not just a technical challenge but also a deep philosophical and conceptual one. For that reason, we argue for a Philosophy of Autonomous Science (PAS): a program that translates core epistemic aims, including understanding, curiosity, surprise, interest, creativity, and novelty, into computable, nonanthropocentric objectives, while ensuring the safe and successful deployment of artificial scientists. Drawing on our own research experience, we show how insights from the philosophy of science can directly inform the design of AI systems for scientific discovery; outline ten inaugural questions that define the agenda of PAS; and invite philosophers, scientists, and AI researchers to collaborate in shaping the principles of the coming era of artificial scientists.

Artificial intelligence has started to contribute significantly to science. Let's consider biology, where Google DeepMind's AlphaFold program has solved one of the most significant challenges in structural biology – the protein prediction problem – leading to the 2024 Nobel Prize in Chemistry being awarded to DeepMind's John Jumper and Demis Hassabis (shared with biochemist David Baker for his work on computational protein design).[1] Similarly, in materials science, intelligent programs exploring the vast space of functional materials are beginning to discover molecules with exceptional properties.[2] In physics, AI algorithms have enabled, for the first time, the prediction of the sky location of neutron star mergers from gravitational wave signals before the mergers occur, opening the door to unprecedented multimessenger astronomy in the near future.[3]

Unquestionably, these are exciting achievements. However, if we aim to progress from AI being a useful scientific tool toward it being capable of autonomously conducting science, we must first recognize that this is not merely a technical challenge and it cannot be solved just by technical advances. Rather, it requires understanding how the best scientists perform science. For example, why and how are humans

https://doi.org/10.1162/DAED.a.991

creative? What constitutes scientific creativity? Why are humans curious, and what fundamental inner drives motivate scientists to deepen their understanding of the world? How do humans achieve understanding, and what precisely does scientific understanding entail? We believe that answering these questions and artificially recreating these traits is essential to the success of *autonomous science* (that is, science conducted by autonomous machine systems), even if not all of these human traits are replicable.

We strongly believe that the philosophy of science community can help answer these questions and thereby make significant contributions to the future of fully autonomous science. We propose a focused research program that studies how to translate traits of successful human researchers to autonomous machine scientists. The answers should be formulated in a nonanthropocentric way in the sense that they should be applicable to both biological and artificial scientists. We also encourage the study of the potential consequences of autonomous science, for example, on scientific understanding, one of the main aims of science.

To explain why we are so convinced that insights from philosophy could advance the automation of science, we offer our first personal anecdote. One main direction of our research (namely, that of Mario Krenn) is the automated discovery of new physics experiments.[4] Our AI programs have discovered experimental setups that no human physicists could find (such as the generation of complex quantum entanglement or new quantum transformations, or the design of extremely sensitive gravitational wave detectors), and numerous experiments have been implemented in laboratories. Whenever a machine finds a solution to a question that humans cannot answer, that solution must contain new ideas and tricks that humans can learn from. The problem is that learning from and understanding computer-designed solutions can be extremely challenging. While it has recently been possible to do so – for example, with our 2017 discovery of a new way to generate photon entanglement – it usually takes weeks to understand the underlying physical processes.[5] Around the year 2019, we started to think more deeply about how we can better gain new understanding with artificial intelligence – both for our particular questions and in general. We soon realized that because we could not consistently spell out what scientific understanding is, we couldn't even begin to answer the question. At this point, we discovered the work of Dutch philosopher Henk W. de Regt, who, in his 2017 book *Understanding Scientific Understanding* (which earned him the 2019 Lakatos Award for outstanding contribution to the philosophy of science), elucidated exactly what we were searching for: clear-cut and nonanthropocentric criteria to identify what scientific understanding is.[6] De Regt's insight that a core principle of scientific understanding is the ability to apply new insights in different scenarios – that is, to generalize without full computation – helped us to develop an entirely new algorithm. We built the algorithm, Theseus, to enable humans to use the computer-designed solutions to quantum physics problems in different

contexts.[7] We translated this goal into the application of a very abstract representation of quantum optics: based on graphs, together with highly efficient pruning and simplification techniques that produce solutions that conventionally can be understood immediately. Theseus allowed us to gain new insights into the principles of quantum entanglement generation and quantum interference. For example, by solving a question about the generation of heralded quantum entanglement for qubits, we immediately recognized the underlying structure that made the solution work and were able to generalize it to high-dimensional heralded entanglement generation that could be important in future quantum networks. We introduced Theseus in 2021 and still use the algorithm's generalizations for most quantum-optics design tasks today.[8]

This anecdote demonstrates that insights from the philosophy of science can directly contribute to the advancement and automation of physics. We are convinced that this was not a singular exception but that more insights from the philosophy of science could in the near future enable progress in AI for science and for autonomous science.

In this essay, we describe how the philosophy of science might change and expand its focus toward a new discipline, the *philosophy of autonomous science* (PAS), and how PAS could and should emerge, directly contributing to the design of artificial scientists to explore the universe, from its smallest to largest structures. While philosophers have discussed related issues, we believe their importance for autonomous science warrants far greater, sustained attention. For that reason, we propose ten crucial, initial research questions for this new field.

1 Scientific understanding. Imagine an oracle that correctly predicts the outcome of every particle physics experiment, the products of every chemical reaction, or the function of every protein. Such an oracle would revolutionize science and technology as we know them. However, as scientists, we would not be satisfied with the oracle alone. We desire more. We would want to comprehend how the oracle arrived at these predictions. This capability, known as *scientific understanding*, is frequently recognized as the essential goal of science.[9] But what exactly does scientific understanding mean? How could artificial intelligence contribute to it? How could we recognize that an AI system possesses scientific understanding? And, finally, as AI systems become increasingly advanced, how can we continue to comprehend their superhuman solutions, concepts, and ideas?

Interestingly, notable philosophers in the early twentieth century developed comprehensive theories about scientific explanation while arguing that philosophers should disregard "understanding," since these concepts are psychological and pragmatic rather than logical. Not all scientists agreed with this viewpoint. Physicists including Lord Kelvin, Erwin Schrödinger, and Richard Feynman developed their own models of scientific understanding.

In 2005, philosophers of science Henk W. de Regt and Dennis Dieks developed a new theory on what might constitute "scientific understanding," foreshadowing de Regt's work in *Understanding Scientific Understanding*.[10] The de Regt-Dieks theory is contextual and pragmatic; it refers to observable properties rather than psychological states of mind. They regard previous concepts like mechanical models or visualizations as "tools for understanding" within a broader general framework. Their model, particularly in de Regt's book, is strongly driven by the principle that a "satisfactory conception of scientific understanding should reflect the actual (contemporary and historical) practice of science."

Their view is that "a phenomenon can be understood if there exists an intelligible theory T, such that scientists can qualitatively recognize characteristic consequences of T without performing exact calculations." They explicitly define two interrelated criteria explaining what it means to understand a phenomenon and what constitutes an intelligible theory.

This theory is defined independently of humans; thus, it can – in principle – be applied to entities beyond humans, including machines. As such, it was a great inspiration for us to develop new algorithms and to classify scientific understanding through AI in general. However, one of the most powerful ideas in the definition – "without performing exact calculations" – makes a direct application to AI algorithms difficult. For that reason, we have transformed their definition into a criterion for teaching (see question eight), related to a Turing test. Unfortunately, this method cannot be scaled. Thus, large-scale exploration of AI-understanding cannot be tested. For that reason, one critical question is whether we can find one (or even more) nonanthropocentric and computable approximation of scientific understanding. And if so, what is its relation to de Regt's criteria? Is it possible to maximize the new computable notion of scientific understanding?

2 **Scientific motivations**. What drives human researchers, such that they get up every day and work tirelessly to explore the world? At the core of this drive lie *epistemic emotions*, which are directly connected to knowledge, learning, and understanding. For example, curiosity motivates us to explore, scientific interest guides the direction of exploration, and surprise and wonder sustain engagement and attention when we encounter the unexpected or unknown, while transforming uncertainty into fascination that keeps our mind open to new possibilities and deeper understanding. (We will discuss curiosity, surprise, and interest in subsequent questions.)

What are other motivations for scientists, and how can those be recreated artificially? For example, the epistemic emotion of awe and wonder is intertwined with cultural practices.[11] What could constitute an AI "culture" that leads to related emergent phenomena? Another motivation comes from direct competition between scientists – often at the verge of an (expected) breakthrough. How could one recre-

ate true competitive behavior between artificial systems? Another motivation for some scientists comes from the awareness of our own mortality and the reality that each discovery is made against the backdrop of a finite life. Could the introduction of artificial finiteness in machines lead to more motivation to explore the world?

These questions (and the questions in the subsequent three sections) are closely related to emotions, albeit epistemic ones. We essentially ask how such emotions could be recreated artificially. However, all current implementations of emotions are biological. Can their functional roles be captured *in silico*, or are certain forms of biological embodiment indispensable for (epistemic) emotions?

3 **Scientific curiosity**. Scientists are curious about the world, and this curiosity shapes what they research. But what does *curiosity* actually mean? In AI research, the term "curiosity" is used for exploring complex environments with sparse rewards. Think about a Super Mario game: to finish a level – the goal of the game – a large number of actions must be performed, and it is not immediately clear how or whether a specific action brings the player closer to the goal. This is troublesome not only for human players but also for AI agents playing similar games. To overcome the problem of sparse rewards, researchers introduced a different form of reward that is *intrinsic* to the agent.[12] Artificial agents are developed such that they want to be able to predict the consequences of their own actions in an environment. A good way for them to learn to do so is by taking actions for which they cannot predict the consequences well. The agents seem to mimic a simple form of curiosity; thus, researchers call this intrinsic reward a "curiosity-based" reward. Surprisingly, this intrinsic curiosity-based reward leads to agents that can play many computer games without relying on the scores provided by the game itself.[13]

This raises a number of important questions: How close is this form of artificial curiosity to human curiosity? Can human curiosity be described as motivation to act when the consequences are difficult to predict, or do we have entirely different forms of curiosity? An interest in knowing the unknown seems universally relevant. But perhaps curiosity should be distinguished from richer emotional experiences such as *wonder*, in which the unknowing agent also experiences a sense of admiration (for a variety of possible aesthetics).

Helen De Cruz, a philosopher of science, argued that awe and wonder have a defamiliarizing effect, allowing us to escape existing patterns of thought.[14] This makes them extremely useful for science; for example, in prompting revolutionary changes. Nonetheless, as some scientific discoveries have accidental triggers, both fleeting interests in novel features of an environment as well as deeper desires for elating new knowledge seem relevant. Can we find models of curiosity that range in degree and kind, closer to the complex forms experienced by humans but that we can measure in a variety of arbitrary entities, including humans, animals, and machines?

4 **Scientific surprise.** Surprising insights often open new ways of thinking about a system. They not only help individual scientists learn unexpected facts but, on a collective scale, can trigger unforeseen breakthroughs. It is empirically demonstrated that surprising connections lead to impactful scientific results.[15] But what does *surprise* mean, how can we identify and quantify it, and how can we maximize the surprise generated by AI agents in science?

Two computable proxies for the intuitive notion of surprise are frequently used in computer science: *Bayesian surprise* and *Shannon's surprisal*. Bayesian surprise measures how much a new observation changes your beliefs; it quantifies the information gained when updating from prior to posterior. Technically, it is the Kullback-Leibler divergence from the posterior to the prior, so it is always nonnegative and is zero only when the observation leaves your beliefs unchanged. Unlike Shannon's surprisal, which measures how unlikely the new data are under your current model, Bayesian surprise measures the size of the belief update. Bayesian surprise has previously been used in an influential experiment on computational creativity, as we will discuss in question six.

Quite clearly, these two information-theoretical interpretations of surprise do not capture the epistemic emotion described by philosophers and psychologists. But there are other ways to approximate surprise. For example, an empirical study of over two hundred scientists found that surprise often carries an aesthetic component, a *feeling of beauty or elegance* in the unexpected coherence of a new insight.[16] Could and should this aesthetic component be recreated to capture the notion of surprise? Could surprise naturally emerge as a consequence of the maximization of other qualities, or do we need *artificial* surprise to build truly powerful artificial scientists?

5 **Scientific interest.** Humans put much effort, money, and brainpower into things that interest them. In science, interest is one of the main drivers of research direction and is thereby responsible for scientific progress on both the individual and global scale. But what does *interest* mean? What interests humans, and in particular scientists? And how could we recreate scientific interest in an automated way and build machines that are intrinsically interested in exploring the universe?

In psychology, "interest" is often considered an epistemic emotion, along with surprise, confusion, and awe.[17] A standard psychological way of describing interest combines two concepts. First, the artifact needs to have high novelty-complexity. But novelty is not enough; it also needs to have high comprehensibility. That is, humans need to have the impression that they could understand the underlying idea.

Many AI researchers are attempting to build systems capable of drawing from millions of scientific papers to help scientists create new and exciting research ideas. But are those ideas actually interesting? In a number of these projects, researchers

have evaluated how experienced scientists rank the interest of AI-generated ideas. In one of our own projects, called SciMuse, we asked more than one hundred research group leaders in the Max Planck Society to help us evaluate how interesting the AI-generated ideas are.[18] While we now have more than four thousand human-evaluated ideas, we cannot yet explain what makes an idea or discovery interesting.

This leaves us with several critical questions again: How can comprehensibility be approximated computationally? Can we achieve true scientific interest by maximizing complexity and comprehensibility? What are other computable criteria that generate interest? Does interest emerge from other intrinsic motivations such as surprise or curiosity? Is scientific interest domain-dependent or are there universal, field-independent criteria? And how can a definition of scientific interest be informed by large-scale sociological experiments that involve many successful scientists?

6 **Scientific creativity**. Creativity is one of humanity's highest achievements; it is the driving force behind our most remarkable innovations. Although it is often celebrated in the arts – in paintings, poetry, or music – creativity is equally crucial in science and technology, where imaginative thinking paves the way for groundbreaking discoveries. But what is *creativity*? Does it require a special process like intentional reasoning, and, if so, can this be approximated with computational methods? How can we evaluate and potentially automate it in the scientific context?

Philosophers and scientists have long identified creativity using two crucial properties: novelty and usefulness. (Sometimes the closely related properties of originality and effectiveness are used, together with surprise, which has been advocated as a third criterion, to emphasize nontrivial forms of novelty.) This categorization is sometimes called the "standard definition of creativity."[19] In 2013, this abstract idea inspired a computational creativity experiment at IBM called Chef Watson. The goal was to develop an algorithm that could produce creative artifacts – in Chef Watson's case, creative culinary recipes and menus.[20] Through this experiment, the IBM team, in collaboration with chefs at the Institute of Culinary Education, worked to demonstrate how the two criteria of novelty and usefulness could be automatically evaluated and thus optimized. Toward these aims, the team collected a large dataset of ingredients, large datasets of international recipes of various dishes, and large datasets of experimental psychophysics data (human pleasantness ratings) and physicochemical features of flavor compounds. Novelty was estimated by a Bayesian surprise function – a Kullback-Leibler divergence that estimated how much a previously unseen dish changes the general recipe distribution. Usefulness was approximated by a score that estimated the pleasantness of the new dish computed from the large database of experimental pleasantness and flavors. The team then maximized combinations of the two

approximated concepts and indeed found numerous novel and pleasant recipes, as ranked by human experts (chefs).

This groundbreaking demonstration shows that complex ideas such as creativity can be approximated and maximized. Thus, the question of how we can approximate creativity in the natural science domain becomes crucial. Are there different forms of computable creativity, and what distinguishes artifacts that maximize one or the other approximation of creativity? Can different modes and qualities of creativity, such as those described by philosopher and cognitive scientist Margaret Boden – exploration of, association within, or transformation of a conceptual space – be computationally approximated?[21] Additionally, could creativity emerge as a side effect of a more general optimization strategy, such as surprise maximization or uncertainty minimization?

7 **Scientific novelty**. As scientists, we strive for novel results. This sounds simple – but what is *novel*, and what is *novelty* in science? Sometimes, a tiny modification of a system has severe consequences. Sometimes, in hindsight, a discovery looks obvious, even if it was overlooked for a long time. This question of novelty is not only useful by itself, but as a key component of other crucial traits such as creativity or interest; we need to understand it well.

There are many metrics that capture some notions of scientific novelty. For example, in information theory, the number of bits that are required to encode a new artifact under the current model approximates novelty. In knowledge graphs or latent spaces of large neural networks, the distance of the new artifact to other known artifacts is used to approximate novelty. In AI-driven idea generation, novelty is sometimes captured by the nonexistence of links between concepts in knowledge graphs. These notions clearly only capture a small aspect of novelty in science. Recently, large language models (LLMs) have been used to automatically evaluate whether findings, ideas, or hypotheses are new – a process that clearly introduces new biases by prioritizing a particular concept of novelty.

In scientific exploration, researchers seek to discover new phenomena that are not specifically predicted by theory, emergent phenomena that are not directly visible from their discipline's governing equations (for instance, superconductivity and the whole idea of quantum computing from Schrödinger's equation), or new equations and models that describe experimental data. So what types of novelty exist and how do they indicate success? One of us (Heather Champion) recently proposed several types of "strong novelty" for science that demonstrate key learning outcomes, and showed how current AI algorithms are already achieving many of them by generating surprise, reducing utility blindness, and eliminating deep ignorance, while potentially changing scientific concepts.[22]

In general terms, we ask: How can one identify novelty independent of the domain? And can general notions of scientific novelty be computable?

8 **Machine-to-human pedagogy and AI teachers.** In 2000, science fiction writer Ted Chiang wrote a prophetic one-page fiction story in *Nature* titled "Catching Crumbs from the Table." In this story, Chiang describes the emergence of "metahumans": genetically modified humans with enormous brain capacities. He portrays a world in which the last original human-authored research was submitted for publication twenty-five years prior, with humans unlikely ever to make any original scientific contributions again. In Chiang's scenario, some human scientists "left the field altogether, but those who stayed shifted their attentions away from original research and toward hermeneutics: interpreting the scientific work of metahumans."[23]

We believe we are on the verge of entering such a world. We observe it in our own research into using intelligent algorithms to design new physics experiments. Traditionally, creative and experienced human researchers devise such experimental designs. However, given the enormous number of possible experimental setups, it is questionable whether humans have already discovered all designs with exceptional properties (our own research indicates they have not). As our human-machine team competes against the best human-created designs – crafted through careful, rational design decisions – we know that if machine solutions outperform human creations, those solutions must contain new ideas and concepts from which humans can learn. (The machine solutions can be directly verified by external tools that are not dependent on the algorithm that discovered them.) And indeed, in many cases, we have discovered new generalizable concepts and ideas in quantum physics by analyzing what the machines have discovered (we describe one such instance in the introduction to this essay).[24] However, this is not always the case. With physics researcher Yehonathan Drori and experimental physicist Rana Adhikari, we recently used AI algorithms to design new gravitational wave detectors, some of the most sensitive measurement devices ever conceived by humanity. These detectors have, so far, been designed by a collaboration of hundreds of researchers over the last twenty to thirty years. Our AI system has discovered more than a dozen new and at least theoretically more sensitive detector designs.[25] However, even after analyzing them for half a year, we were not able to understand the underlying big picture of the idea. One can compute the designs by hand and with software, but we failed to understand why they worked.

It seems inevitable that this dynamic will increasingly occur in physics and other sciences: Machines will invent solutions from which humans could learn. If we are fortunate (as in the first quantum optics example), we will grasp the underlying principles and thereby enhance our scientific understanding.[26] However, we anticipate that we will increasingly encounter scenarios similar to the gravitational wave detectors, where we can confirm but not fully comprehend the principles behind AI-discovered improvements. This raises the critical question: *What should we do?* Perhaps "we need not be intimidated by the accomplishments of metahuman

science," as Chiang concludes in his prophetic text. Rather "we should always remember that the technologies that made metahumans possible were originally invented by humans, and they were no smarter than we."

What else could be done to solve this gap in understanding AI-discovered solutions? One powerful option would be developing AI systems capable of explaining the underlying concepts and ideas to us. Such systems would go far beyond the current scope of interpretable AI or explainable AI (XAI). In XAI, the goal is for humans to understand the decision-making processes of advanced AI, which is crucial for AI safety and ethics.

However, it is unclear whether looking inside the "machine's brain" is the most effective way for us to gain new understanding. Even if we can comprehend its internal workings, it remains uncertain whether humans will ever fully understand solutions in the same way as a system with superhuman intelligence. Our limited mental capacities in terms of memory and computational speed compel us to seek simplified models to understand phenomena. Conversely, an advanced system with access to different computational resources could develop scientific models beyond our capabilities.

Instead of examining the internal workings of powerful AI, we might look at how we learn from each other. When we want to learn what another human knows, we do not look into their neural circuits; rather, we communicate and learn from the teachings of other people. We might need to do the same for artificial systems: develop AI algorithms that can explain new physics models, concepts, and ideas at a level comprehensible to us. This capability closely aligns with the test for scientific understanding described in question one. In short, we should focus on building a Teacher AI (TAI) for science.

We can envision a TAI that either develops solutions or uses solutions from other AIs and explains them to humans. It could resemble a student-teacher dynamic, in which students ask questions for clarification and details, and the teacher provides insights. Likely, the TAI would need access to advanced scientific tools similar to those that humans use for visualization, simplification, or generalization. The TAI could generate exercises to test our knowledge and track our understanding and its progression during teaching.

Developing such a system also requires the TAI to have access to the same mathematical and physical tools used by human scientists. For example, in quantum mechanics, it would benefit from autonomous access to open-source programs like Qiskit (IBM's quantum circuit simulator), NetKet (physicist Giuseppe Carleo's toolbox for simulating many-body quantum systems), QuTiP (a quantum optics simulator developed by scientist Franco Nori's research group at the RIKEN institute), and PyTheus (a quantum physics experiment discovery framework developed by our Krenn Research Group then-hosted at the Max Planck Institute for the Science of Light).[27] The TAI should not only provide verbal explanations but also fully

leverage the numerous "tools for understanding," as described by de Regt.[28] For instance, regarding visualization, the TAI should determine how concepts and ideas can be visually represented through graphs, diagrams, or potentially videos and 3D animations. It should use unification by generating new examples that employ the same concepts and create practical exercises to help us grasp the application of these concepts. We have not seen practical effort in this direction, but in order to avoid Chiang's dystopian future, such systems will be essential in the forthcoming years.

This potential raises numerous questions: How could machine-to-human pedagogy advance to ease human understanding of alien concepts? And what motivates human scientists to understand the universe when their role is degraded from active explorer to passive student?

9 **Automating the philosophy of science.** The philosophy of science has long shaped the rules of research, from Karl Popper's theory of falsification and Imre Lakatos's research programs on falsificationism to Paul Feyerabend's pluralism and methodological anarchism.[29] If we reason about autonomous science, we must think about the underlying foundations and rules of science, and should consider the possibility of automating the philosophy of science itself. But what does that mean?

One potential way to automate some aspects of the philosophy of science would be to approximate "the rules of doing science" as policies that describe ways of generating hypotheses, developing overthrowing theories, or designing experiments. Through simulated worlds, these policies could be evaluated on the quality of the systems' discoveries about the (simulated) worlds. The evaluation metrics could be formalized as objective functions to select high-performing policies – or, in the pluralistic spirit of Feyerabend, a diverse set of them – and thereby drive the discovery of new rules for exploring the world. Yet, as always, many questions arise here: What are good metrics for evaluating different policies, such as different rules of science? How can we guard against Goodhart's law (policies that game the metric) and ensure that this autophilosophy preserves desirable traits such as methodological pluralism?[30] How can these policies be modeled and parametrized? And can large surveys of the beliefs of working philosophers and scientists help to model and emulate their decision processes?[31]

10 **The role of humans in a future of autonomous science.** Let us assume a world with successful artificial scientists that can explore the universe independently, without human input. What, then, is our role in such a world? Humans will remain the strategic leaders, setting the grand directions and determining how resources are allocated. These decisions will not be purely scientific; they will also be guided by societal goals, ethical considerations, and collective needs. Developing effective and transparent mechanisms for such

high-level decision-making will be one of the central challenges for the future governance of science.

At the same time, it will be essential to establish rigorous methods that allow humans to verify, interpret, and safely validate the discoveries proposed by artificial scientists. Is this possible if we cannot genuinely understand the AI-generated ideas? How can we maximize the benefits of autonomous scientific discovery while mitigating the risks of potentially catastrophic outcomes (for instance, the accidental discovery of an easily reproducible bioweapon or an algorithm capable of breaking modern encryption schemes)? Equally important will be ensuring that artificial scientists do not develop systematic blind spots: regions of the scientific search space left unexplored not because they are uninteresting but because of hidden architectural biases or misaligned objectives.

As the era of artificial scientists approaches, we must now begin to identify which human roles, values, and forms of oversight will remain indispensable in guiding, interpreting, and safeguarding the future of knowledge itself.

We have proposed ten questions that could – and should – form the foundation of a new field: the philosophy of autonomous science. The goal of this field should be to help develop and guide the emerging era of artificial scientists by translating fundamental epistemic traits and values (such as curiosity, creativity, motivation, surprise, interest, understanding, and novelty) into nonanthropocentric, computable objectives, and by establishing norms, governance frameworks, and control mechanisms for autonomous scientific systems.

If we build them with care, artificial scientists will extend our capabilities for curing diseases, solving mathematical mysteries, and exploring the cosmos at a speed beyond imagination.

ABOUT THE AUTHORS

Mario Krenn is Professor of Machine Learning in Science at the University of Tübingen, where he leads the Artificial Scientist Lab (previously based at the Max Planck Institute for the Science of Light). He has published in such journals as *Physical Review X*, *Machine Learning: Science and Technology*, *Nature Machine Intelligence*, and *Nature Photonics*.

Heather Champion is a Guest Researcher in the Ethics & Philosophy Lab, part of the Cluster of Excellence–Machine Learning for Science at the University of Tübingen. She is also a PhD candidate in Philosophy at the Rotman Institute of Philosophy and at Western University. She has published in such journals as *Synthese* and *Philosophy of Science*.

ENDNOTES

[1] John Jumper, Richard Evans, Alexander Pritzel, et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596 (7873) (2021): 583–589.

[2] Robert Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, et al., "Data-Driven Strategies for Accelerated Materials Design," *Accounts of Chemical Research* 54 (4) (2021): 849–860.

[3] Maximilian Dax, Stephen R. Green, Jonathan Gair, et al., "Real-Time Inference for Binary Neutron Star Mergers Using Machine Learning," *Nature* 639 (8053) (2025): 49–53.

[4] Mario Krenn, Mehul Malik, Robert Fickler, et al., "Automated Search for New Quantum Experiments," *Physical Review Letters* 116 (9) (2016): 090405; and Mario Krenn, Manuel Erhard, and Anton Zeilinger, "Computer-Inspired Quantum Experiments," *Nature Reviews Physics* 2 (11) (2020): 649–661.

[5] Mario Krenn, Armin Hochrainer, Mayukh Lahiri, and Anton Zeilinger, "Entanglement by Path Identity," *Physical Review Letters* 118 (2017): 080401.

[6] Henk W. de Regt, *Understanding Scientific Understanding* (Oxford University Press, 2017); and The London School of Economics and Political Science, "Lakatos Award Lecture," https://www.lse.ac.uk/philosophy/events/lakatos-award-lecture (accessed January 30, 2026).

[7] Mario Krenn, Jakob S. Kottmann, Nora Tischler, and Alán Aspuru-Guzik, "Conceptual Understanding through Efficient Automated Design of Quantum Optical Experiments," *Physical Review X* 11 (2021): 031044.

[8] Carlos Ruiz-Gonzalez, Sören Arlt, Jan Petermann, et al., "Digital Discovery of 100 Diverse Quantum Experiments with PyTheus," *Quantum* 7 (2023): 1204.

[9] Henk W. de Regt and Dennis Dieks, "A Contextual Approach to Scientific Understanding," *Synthese* 144 (1) (2005): 137–170; and Angela Potochnik, *Idealization and the Aims of Science* (University of Chicago Press, 2017).

[10] De Regt and Dieks, "A Contextual Approach to Scientific Understanding"; and De Regt, *Understanding Scientific Understanding*.

[11] Helen De Cruz, *Wonderstruck: How Wonder and Awe Shape the Way We Think* (Princeton University Press, 2024).

[12] Jürgen Schmidhuber, "Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)," *IEEE Transactions on Autonomous Mental Development* 2 (3) (2010): 230–247; and Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell, "Curiosity-Driven Exploration by Self-Supervised Prediction," in *ICML '17: Proceedings of the 34th International Conference on Machine Learning – Volume 70*, ed. Doina Precup and Yee Whye Teh (Association for Computing Machinery, 2017), 2778–2787.

[13] Yuri Burda, Harri Edwards, Deepak Pathak, et al., "Large-Scale Study of Curiosity-Driven Learning," paper presented at the Seventh International Conference on Learning Representations (ICLR 2019), New Orleans, Louisiana, May 6–9, 2019.

[14] Taylor Carman, "Wonderstruck: How Wonder and Awe Shape the Way We Think," *Notre Dame Philosophical Reviews*, June 6, 2025; and De Cruz, *Wonderstruck*.

[15] Feng Shi and James Evans, "Surprising Combinations of Research Contents and Contexts Are Related to Impact and Emerge with Scientific Outsiders from Distant Disciplines," *Nature Communications* 14 (2023): 1641.

[16] Milena Ivanova and Brandon Vaidyanathan, "Surprise in Science: A Qualitative Study," *Erkenntnis: An International Journal of Scientific Philosophy* 91 (1) (2026): 57–80.

[17] Paul J. Silvia, "Interest–the Curious Emotion," *Current Directions in Psychological Science* 17 (1) (2008): 57–60.

[18] Xuemei Gu and Mario Krenn, "Interesting Scientific Idea Generation Using Knowledge Graphs and LLMs: Evaluations with 100 Research Group Leaders," arXiv (2024), https://doi.org/10.48550/arXiv.2405.17044.

[19] Mark A. Runco and Garrett J. Jaeger, "The Standard Definition of Creativity," *Creativity Research Journal* 24 (1) (2012): 92–96.

[20] Lav R. Varshney, Florian Pinel, Kush R. Varshney, et al., "A Big Data Approach to Computational Creativity," arXiv (2013), https://doi.org/10.48550/arXiv.1311.1213; and Lav R. Varshney, Florian Pinel, Kush R. Varshney, et al., "A Big Data Approach to Computational Creativity: The Curious Case of Chef Watson," *IBM Journal of Research and Development* 63 (1) (2019): 7.1–7.18, https://doi.org/10.1147/JRD.2019.2893905.

[21] Margaret A. Boden, "Computer Models of Creativity," *AI Magazine* 30 (3) (2009): 3; and Margaret A. Boden, *The Creative Mind: Myths and Mechanisms*, 2nd ed. (Routledge, 2004). As Boden explains in her introduction to *The Creative Mind*, reprinted by *Interalia Magazine* in 2016: "'What is going on' isn't magic–and it's different in each type of case. For creativity can happen in three main ways, which correspond to the three sorts of surprise. The first involves making unfamiliar combinations of familiar ideas. Examples include poetic imagery, collage in painting or textile art, and analogies. These new combinations can be generated either deliberately or, often, unconsciously. . . . The other two types of creativity are interestingly different from the first. They involve the *exploration*, and in the most surprising cases the *transformation*, of conceptual spaces in people's minds. Conceptual spaces are structured styles of thought. They're normally picked up from one's own culture or peer-group, but are occasionally borrowed from other cultures." See Margaret A. Boden, "Creativity in a Nutshell," *Interalia Magazine*, https://www.interaliamag.org/articles/margaret-boden-creativity-in-a-nutshell (accessed February 5, 2026).

[22] Heather Champion, "Strong Novelty Regained: High-Impact Outcomes of Machine Learning for Science," *Synthese* 206 (3) (2025): 1–23.

[23] Ted Chiang, "Catching Crumbs from the Table," *Nature* 405 (6786) (2000): 517, https://doi.org/10.1038/35014679.

[24] Krenn, Hochrainer, Lahiri, and Zeilinger, "Entanglement by Path Identity."

[25] Mario Krenn, Yehonathan Drori, and Rana X. Adhikari, "Digital Discovery of Interferometric Gravitational Wave Detectors," *Physical Review X* 15 (2) (2025): 021012.

[26] Krenn, Malik, Fickler, et al., "Automated Search for New Quantum Experiments."

[27] IBM Quantum Computing, "Qiskit," https://www.ibm.com/quantum/qiskit (accessed February 6, 2026); Giuseppe Carleo, Kenny Choo, Damian Hofmann, et al., "NetKet: A Machine Learning Toolkit for Many-Body Quantum Systems," *SoftwareX* 10 (2019): 100311, https://doi.org/10.1016/j.softx.2019.100311; J. Robert Johansson, Paul D. Nation, and Franco Nori, "QuTiP: An Open-Source Python Framework for the Dynamics of Open Quantum Systems," *Computer Physics Communications* 183 (8) (2012): 1760–1772, https://doi.org/10.1016/j.cpc.2012.02.021; and Ruiz-Gonzalez, Arlt, Petermann, et al., "Digital Discovery of 100 Diverse Quantum Experiments with PyTheus."

[28] De Regt, *Understanding Scientific Understanding*.

[29] Karl Popper, *The Logic of Scientific Discovery* (Hutchinson & Co, 1959); Imre Lakatos, "Falsification and the Methodology of Scientific Research Programmes," in *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*, ed. Imre Lakatos and Alan Musgrave (Cambridge University Press, 1970); and Paul Feyerabend, *Against Method: Outline of an Anarchistic Theory of Knowledge*, 4th ed. (Verso, 2010).

[30] Eric Oberheim, "Rediscovering Einstein's Legacy: How Einstein Anticipates Kuhn and Feyerabend on the Nature of Science," *Studies in History and Philosophy of Science Part A* 57 (2016): 17–26; and Charles Goodhart, "Problems of Monetary Management: The U.K. Experience," in *Papers in Monetary Economics, 1975* (Reserve Bank of Australia, 1975).

[31] David Bourget and David J. Chalmers, "What Do Philosophers Believe?" *Philosophical Studies* 170 (3) (2014): 465–500.