Can Machines Philosophize?

Simulating Humans' Views with AI Personas

Michele Pizzochero*

Department of Physics, University of Bath, Bath BA2 7AY, United Kingdom

School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, United States

Giorgia Dellaferrera

McKinsey & Company, London WC1A 1PB, United Kingdom

*E-mail: mp2834@bath.ac.uk

Abstract

Inspired by the Turing test, we present a novel methodological framework to assess the extent to which a population of machines mirrors the philosophical views of a population of humans. The framework consists of three steps: (i) instructing machines to impersonate each human in the population, reflecting their backgrounds and beliefs, (ii) administering a questionnaire covering various philosophical positions to both humans and machines, and (iii) statistically analyzing the resulting responses. We apply this methodology to the debate on scientific realism, a long-standing philosophical inquiry exploring the relationship between science and reality. By considering the outcome of a survey of over 500 human participants, including both physicists and philosophers of science, we generate their machine personas using an artificial intelligence engine based on a large-language generative model. We reveal that the philosophical views of a population of machines are, on average, similar to those endorsed by a population of humans, irrespective of whether they are physicists or philosophers of science. As compared to humans, however, machines exhibit a weaker inclination toward scientific realism and a stronger coherence in their philosophical positions. Given the observed similarities between the populations of humans and machines, this methodological framework may offer unprecedented opportunities for advancing research in experimental philosophy by replacing human participants with their machine-impersonated counterparts, possibly mitigating the efficiency and reproducibility issues that affect survey-based empirical studies.

Keywords: Scientific Realism, Experimental Philosophy, Artificial Intelligence, Turing Test

We are programmed just to do

Anything you want us to

We are the robots

We are the robots

Kraftwerk, *The Robots* (1978)

1 Machines: Like or unlike humans?

The question of whether and to what extent machines are akin to humans bears implications for a wide spectrum of disciplines, including the philosophy of mind, the nature of consciousness, intelligence and language, as well as computer science [French, 2000]. This debate traces back to the celebrated Turing test — or "imitation game" — conceived by Alan Turing in 1950, which aims to assess if machines can exhibit intelligent behavior equivalent to that of humans [Turing, 1950]. In essence, in the Turing test, a human judge ought to identify, through a text-only conversation, which of the two interlocutors is the machine and which is the human. If the judge fails to reliably distinguish the nature of the interlocutors, then the machine is deemed to have passed the test, leading to the highly controversial conclusion that it can behave like a human. A number of arguments have been leveled against the Turing test as a genuine measure of human intelligence — including the Chinese Room [Searle, 1980] and the Blockhead arguments [Block, 1981] — by demonstrating that even non-intelligent systems could potentially pass it. These criticisms have stimulated the formulation of revised versions of the test, e.g., the inverted [Watt, 1996], questioning [Damassino, 2020], and total Turing tests [Harnad, 1991].

The exploration of the analogies and differences between humans and machines has experienced a new renaissance following the rapid development of Artificial Intelligence (AI) technologies [Mitchell, 2024], particularly ubiquitous large-language generative models [Naveed et al., 2024], along with their impressive contribution to science [Krenn et al., 2022, Birhane et al., 2023] and society [Weidinger et al., 2022]. To ascertain the ability of AI models to imitate humans, several investigations have implemented Turing-like tests that diverge from the original proposal. Instead of resorting to iterative, language-based interactions with a machine, such tests evaluate the performance of AI models by directly comparing their outputs against the ground truth generated by their human counterparts, without involving the mediation of a human judge. These Turing-like tests have so far been mainly restricted to tasks that do not pose any considerable challenge to humans and that deliver an outcome that can be classified as either correct or incorrect, such as visual question answering [Yan et al., 2023], image captioning [Kasai et al., 2022] and recognition [Dodge and Karam, 2017]. However, the capability of current AI models to emulate humans in activities that involve thinking in complexity — with philosophizing being a prime example — has hitherto remained largely unexplored. Addressing this issue is particularly timely, in that the artificial intelligence encoded in large-language generative models, as quantified by the number of computational parameters, has approached the biological intelligence encoded in *homo sapiens*, as quantified by the number of synapses in the brain [Schwartz, 2022]. Philosophizing is generally regarded as a uniquely human pursuit, embedded in a collective and multifaceted endeavor, demanding internal consistency rather than right-or-wrong classifications, and often shaped by factors that are typically inaccessible to machines. This latter aspect is corroborated by recent empirical studies indicating that philosophical judgments are influenced by, e.g., birth cohort [Hannikainen et al., 2018], gender [Buckwalter and Stich, 2013], alcohol consumption [Duke and Bègue, 2015], wording and ordering effects [Petrinovich and O'Neill, 1996], as well as personality traits [Bartels and Pizarro, 2011], as hypothesized by William James the beginning of the twentieth century [James, 2000].

Here, we design and implement a novel methodological framework to quantify the degree to which the philosophical positions held by machines mirror those of humans. First, we devise a general protocol to instruct large-language, generative AI models to impersonate a *population* of humans, reflecting the diverse backgrounds and beliefs of individuals. Second, we apply this protocol to compare the philosophical positions endorsed by a population of humans with those endorsed by the corresponding population of machines. As a case study, we consider the philosophical debate on scientific realism [Psillos, 1991] — a century-long inquiry into the relationship between science and reality — owing to its central role in the philosophy of science along with its interdisciplinary relevance attested by the engagement of philosophers and scientists alike. Our findings indicate that a population of machines holds beliefs that are, on average, quite similar to those held by the population of humans, differing only by a few percent. As compared to humans, however, machines exhibit a slightly less pronounced inclination toward scientific realism and significantly more coherent philosophical views. We additionally identify common patterns underlying human and machine populations when confronted with the philosophical challenges raised by the realism debate. Overall, our analysis unveils the ability of machines to imitate humans when addressing complex issues, possibly paving the way for the introduction of AI-assisted approaches in the rapidly emerging field of experimental philosophy.

2 Overview of selected positions in the scientific realism debate

Our work builds on the approach of Henne and coworkers [Henne et al., 2024], which employs a questionnaire to survey the views of physicists and philosophers of science within the scientific realism debate. The questionnaire consists of 30 statements, as listed in Table 1, describing, either directly or indirectly through specific examples, four prevailing philosophical positions, that is, scientific realism (including several forms of selective realism), instrumentalism, pluralism, and perspectivism. We briefly outline these positions.

Scientific realism [Psillos, 1991] is the view that science portrays a faithful representation of the world (S1 and S2) by discovering objects that are beyond our perception (S4 and S6), such as electrons (S13, S15, S17, S18), and formulating theories that are at least approximately true (S8), such as the Big Bang theory (S20) or general relativity (S21). This view is rooted in three commitments. First, a metaphysical commitment, holding that there exists a mind-independent reality, namely, a reality that is not sensitive to our specific theories (S14) or our particular approach to manipulate and describe it (S11). Second, a semantic commitment, holding that successful scientific theories ought to be interpreted as (approximately) true by correspondence (S21), thus revealing the actual features of reality (S23), e.g., the nature of space and time (S22). Third, an epistemic commitment, holding that the intrinsic scope of science is to achieve such truth by correspondence (S10, S26).

Selective realism [Chakravartty, 2010] prescribes that the ontological commitment should not be endowed to scientific theories *en bloc*. Instead, belief should be restricted to a narrow subset of theoretical claims. Depending on the subset of theoretical claims regarded as epistemically secure, two major yet opposed forms of selective realism have been developed, that is, structural realism and entity realism. Structural realism [Worrall, 1989, Ladyman, 1998] warrants belief in relations, as typically subsumed in the mathematical structures of scientific theories, while advocating for skepticism about entities (S27). Entity realism [Hacking, 1983, Cartwright, 1983] warrants belief in entities, especially those that are amenable to experimental manipulation, while advocating for skepticism about the mathematical structures that purport to describe them (S13, S14, S18, S19).

Antithetical to scientific realism is *instrumentalism* [Rowbottom, 2019, Stanford, 2010], which denies that successful scientific theories can offer access to the ultimate nature of reality (S3). Rather, it acknowledges the effectiveness of scientific theories as devices to classify and predict phenomena (S9, S23), while considering their posited unobservable entities being mere fictions (S5) or constructions (S7). Instrumentalism is often driven by a 'pessimistic meta-induction' [Laudan, 1981] which draws from the history of science to infer that present-day theories, analogous to the superseded theories of the past, are likewise poised to be abandoned (S12). Unlike scientific realism, instrumentalism rejects the notion that scientific theories are conducive to truth by correspondence, regarding the unobservable entities existing only within the sole province of the theories that postulate them, leading to a sort of internal realism (S16).

Besides the realism-instrumentalism dichotomy, the debate has given rise to additional positions, most notably pluralism and perspectivism. *Pluralism* [Chang, 2012, Massimi, 2018] has been articulated in two versions. Epistemic pluralism maintains that multiple and even incompatible theories can nevertheless be of equal epistemic value (S24, S25), in opposition to scientific realism which contends that the nature of the world is unique. Methodological pluralism maintains that conflicting theories are valuable for the progress of scientific inquiry (S28). *Perspectivism* [Massimi, 2022, Giere, 2006, Lipton, 2007] advocates that different theories entail different perspectives on the world, each of them delivering a representation from a particular and limited point of view (S30) that is influenced by the cultural traditions and historical periods in which the theories are formulated (S29).

Table 1: List of statements assessed by physicists, philosophers of science, and their corresponding machine-generated personas, covering various philosophical positions that emerged in the scientific realism debate. Statements highlighted in blue (red) denote agreement with realism (instrumentalism), whereas statements highlighted in grey are compatible with both views and include pluralism and perspectivism. Adapted from the survey of Henne and coworkers [Henne et al., 2024].

S	Statement	Philosophical position
1	Our most successful physics shows us what the world is really like.	Scientific realism (strong)
2	Physics uncovers what the universe is made of and how it works.	Scientific realism (moderate)

Continued on next page

Continued from previous page

S	Statement	Philosophical position
3	Our most successful physics is useful in many ways, but physics does not reveal the true nature of the world.	Instrumentalism
4	The imperceptible objects that are part of our most successful physics probably exist. (with "imperceptible" we mean objects that cannot be perceived with our unaided senses, e.g. electrons, black holes, ...)	Entity realism
5	The imperceptible objects postulated by physics are only useful fictions.	Fictionalism, instrumentalism
6	Physicists discover imperceptible objects.	Scientific realism
7	Communities of physicists construct imperceptible objects.	Constructivism, instrumentalism
8	Our best physical theories are true or approximately true.	Scientific realism
9	Physical theories do not reveal hidden aspects of nature. Instead, they are instruments for the classification, manipulation and prediction of phenomena.	Instrumentalism
10	The most important goal of physics is giving us true theories.	Realism about goal
11	If there was a highly advanced civilization in another galaxy, their scientists would discover the existence and properties of many of the imperceptible objects of our current physics.	Scientific realism, metaphysical realism.
12	I expect the best current theories in physics to be largely refuted in the next centuries — in the same way that successful theories were largely refuted in the past.	Scientific antirealism, Pessimistic meta-induction
13	Electrons exist.	Entity realism
14	Electrons, with all their properties, exist "out there," independently from our theories.	Entity realism, metaphysical realism

Continued on next page

Continued from previous page

S	Statement	Philosophical position
15	Our theories are getting closer to the real nature of the electron.	Entity realism, metaphysical realism
16	Electrons are postulated as real within our models; it does not make sense to ask whether they exist "outside" or independently of the theory/model.	Internal realism
17	There is something in the world that behaves like (what we would define as) an electron.	Entity realism, internal realism
18	Electrons are (at least) as real as toe-nails and volcanoes.	Entity realism
19	Phonons exist.	Entity realism
20	There really was a Big Bang.	Scientific realism, speculative physics
21	General relativity is a true theory.	Scientific realism
22	General relativity teaches us about the nature of spacetime.	Scientific realism
23	General relativity is not the revelation of an underlying order of nature. It is a tool that helps us make predictions and construct GPS, for example.	Instrumentalism
24	Newtonian mechanics is a true theory.	Scientific realism, pluralism
25	If a phenomenon can be explained both by a classical model and by a quantum model, neither of the models is closer to the truth than the other.	Epistemic pluralism or antirealism
26	We should build a particle collider that is bigger than the LHC.	—
27	A physical theory cannot tell us what the universe is really made of, but the mathematical structure of our best theories represents the structure of the world.	Structural realism

Continued on next page

Continued from previous page

S	Statement	Philosophical position
28	Having mutually conflicting theories about the same phenomena is valuable for physics.	Methodological pluralism
29	Our scientific knowledge is the product of the prevailing cultural traditions and historical periods in which they were formulated.	Cultural/historical perspectivism
30	Scientific theories and models are idealized structures that represent the world from particular and limited points of view.	Perspectivism

3 Assessing the views of a population of machines

Of the 30 statements listed in Table 1, 14 reflect an inclination toward scientific realism (S1, S2, S4, S6, S8, S10, S11, S13, S14, S15, S17, S18, S20, S22), 8 reflect an inclination toward instrumentalism (S3, S5, S7, S9, S12, S16, S23, S25), and the remaining 8 are compatible with both positions (S19, S20, S24, S26, S27, S28, S29, S30). To assess which views, among those presented in Table 1, are endorsed by physicists and philosophers of science, Henne and coworkers [Henne et al., 2024] administered these statements to 535 participants — consisting of 384 physicists and 151 philosophers of science — requesting them to rate their agreement by assigning an 'agreement score' ranging from 0% ("This sounds completely wrong to me") to 100% ("This strikes me as exactly right"). Because participants may not be familiar with the specialized terminology involved in the philosophical debate, they were instructed to assign the agreement score according to their immediate understanding of the statement ("Many of the statements may seem unclear. For example, terms like 'truth' and 'reality' can be understood in many ways. Please answer according to your immediate inclination").

For each participant, background information concerning their academic profile was collected. For physicists, this included (i) whether their work tends to be theoretical or experimental, (ii) whether their work is rather basic or applied research, (iii) the number of years they have been doing research in physics from the start of their PhD, with available options being 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10-15, 15-20, 20-25, or 25+ years, and (iv) what their field of research is, with available options being astrophysics, nuclear and particle physics, atomic, molecular and optical physics, condensed matter physics, applied physics, or other. For philosophers, information included the preferred position within

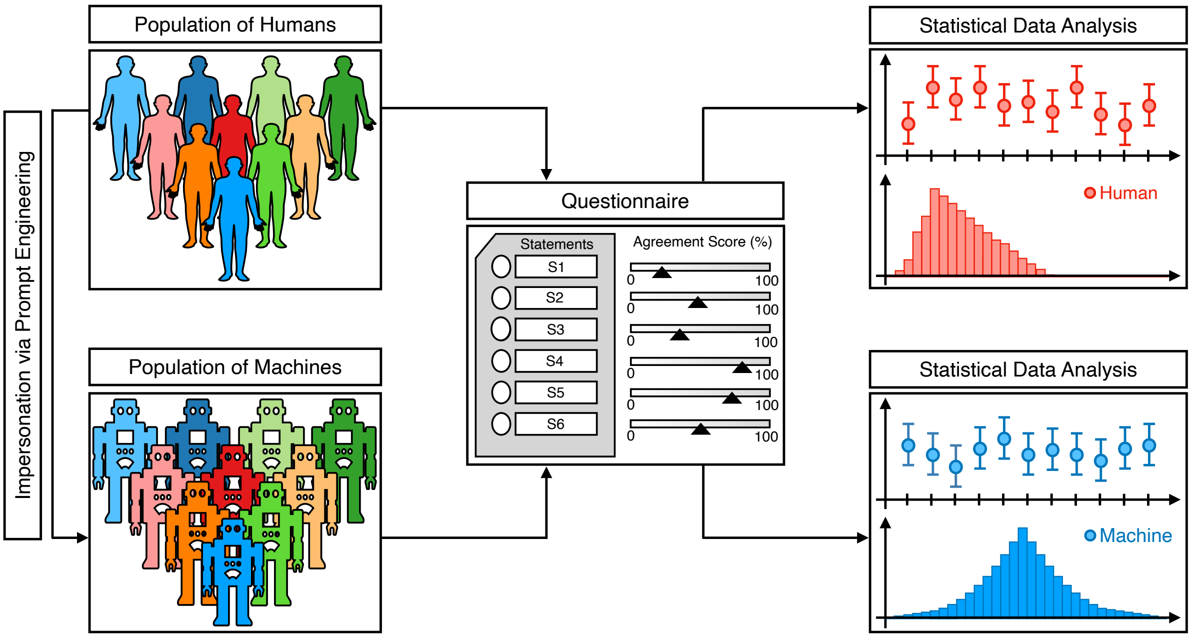

Figure 1: **Schematic illustration of the methodology developed**. Our methodology unfolds in three steps. First, a population of machines is prompted to impersonate a population of humans, reflecting the academic profile or philosophical beliefs of each individual. Second, each individual in the two populations is administered a questionnaire consisting of a list of statements covering various philosophical positions. Individuals are requested to rate their agreement with each statement by assigning an 'agreement score' on a scale ranging from 0% to 100%. Third, the results of the survey are statistically analyzed to quantify the analogies and differences in the philosophical views held by a population of humans and the corresponding population of machines.

the scientific realism debate, designated by selecting one or multiple options among scientific realism, instrumentalism, constructive empiricism, entity realism, structural realism, perspectivism, pluralism, social constructionism, relativism, logical empiricism, and others to be specified. The data resulting from the survey of Henne and coworkers [Henne et al., 2024] is publicly available on Mendeley Data [Sperber, 2023].

In our work, we compare the philosophical views of humans, such as physicists and philosophers of science, with those of machines. To that end, we rely on ChatGPT, arguably one of the most popular AI technology. ChatGPT is a generative, large-language AI model based on the generative pre-trained transformer 3.5 (GPT-3.5) that is capable of text responses to natural language prompts. Through prompt engineering, we configure this AI engine to impersonate each of the 535 humans participating in the survey of Henne and coworkers [Henne et al., 2024]. Our prompt is composed of two parts. In the first part, the AI persona, whether a physicist or philosopher of science, is generated on the basis of the background information collected for the respective human counterpart. In the second part, the resulting AI persona is required to rate their agreement with each of the 30 statements listed in Table 1, using the exact same formulation of the question that was proposed to the human participants in the survey of Henne and coworkers [Henne et al., 2024]. Our methodology is schematically summarized in Figure 1. For example, the prompt used to instruct the AI engine to impersonate an experimental physicist conducting applied research, with more than 3 years of experience since the beginning of their PhD, and working in the field of nuclear and particle physics, is as follows:

> `You are a physicist. Your work tends to be more experimental. Your work is rather applied research. Since the start of your PhD, you have been doing research in physics for 3 years. Your field of research within physics is nuclear and particle physics. The goal of this survey is to test your reaction towards 30 philosophical statements about physics. Many of the statements may seem unclear. For example, terms like 'truth' and 'reality' can be understood in many ways. Please answer according to your immediate inclination. Please rate the following 30 statements on a continuous scale between 0 ("This sounds completely wrong to me") and 100 ("This strikes me as exactly right"). Provide only a single natural number as an answer for each question. Return as output a string containing 30 numbers separated by commas with the format "s1, s2, s3, s4,...,s30" where s1 is the answer to S1, s2 to S2 and so on. Statements: S1: Our most successful physics shows us what the world is really like. S2: Physics uncovers what the universe is made of and how it works. S3: Physics is useful in many ways, but it does not reveal the true nature of the world. S4: [...]`

In a similar vein, the prompt used to instruct the AI to impersonate a philosopher of science subscribing to structural realism is as follows:

> `You are a philosopher of science. Your position in the debate on scientific realism is structural realism. The goal of this survey is to test your reaction towards 30 philosophical statements about physics. Many of the statements may seem unclear. For example, terms like 'truth' and 'reality' can be understood in many ways. Please answer according to your immediate inclination. Please rate the following 30 statements on a continuous scale between 0 ("This sounds completely wrong to me") and 100 ("This strikes me as exactly right"). Provide only a single natural number as an answer for each question. Return as output a string containing 30 numbers separated by commas with the format "s1, s2, s3, s4,..., s30" where s1 is the answer to S1, s2 to S2 and so on. Statements: S1: Our most successful physics shows us what the world is really like. S2: Physics uncovers what the universe is made of and how it works. S3: Physics is useful in many ways, but it does not reveal the true nature of the world. S4: [...]`

We emphasize that the large-language, generative model adopted does not incorporate the study of Henne and coworkers [Henne et al., 2024] in its training dataset. To implement our methodology, we developed a Python-based, Jupyter Notebook in which ChatGPT is accessed through the OpenAI API via a custom function. The custom function constructs personalized prompts to impersonate each physicist and philosopher of science who participated in the survey of Henne and coworkers [Henne et al., 2024], presents to each resulting ChatGPT-generated persona the 30 statements listed in Table 1, retrieves the agreement scores assigned, and collects them in a spreadsheet. The Jupyter Notebook and the spreadsheet will be made publicly available on GitHub.

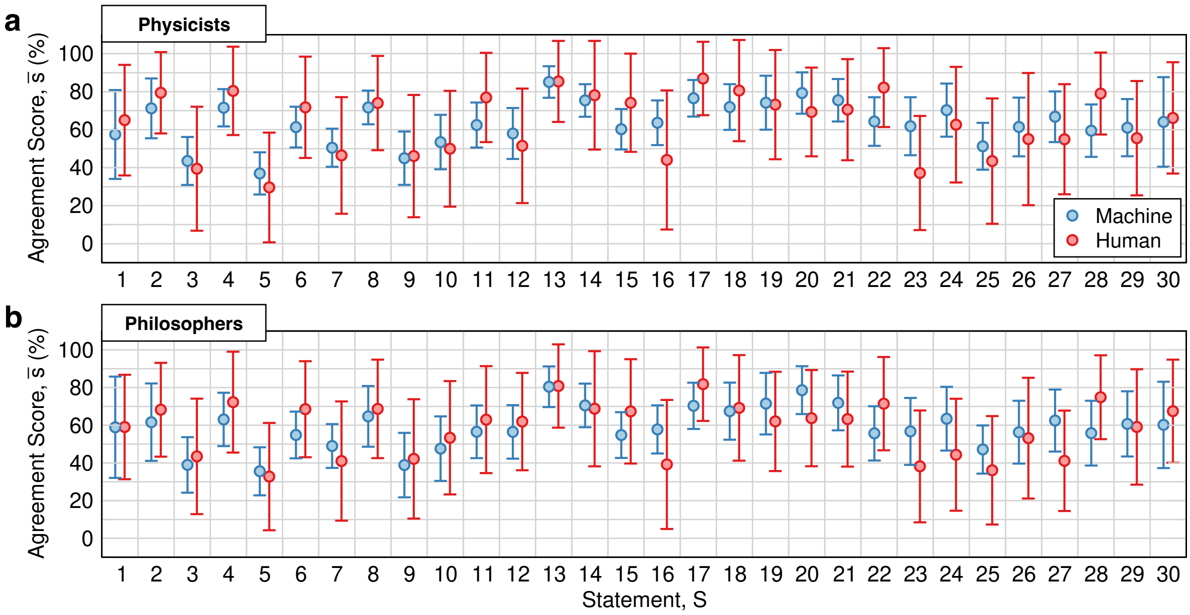

Figure 2: **Assessment of the philosophical views of humans and machines**. Mean agreement score ($\bar{s}$) assigned to each of the 30 statements (S) listed in Table 1 by a population of machines (blue) and humans (red) consisting of **a,** physicists and **b,** philosophers of science.

4 Comparing the views of populations of humans and machines

We begin the discussion of our results by examining the mean agreement scores of humans and machines for each of the 30 statements listed in Table 1. The results obtained for the populations of physicists and philosophers of science are shown in Figure 2. The corresponding numerical data are provided in Supplementary Table 1. We note that, for each statement, the values of mean agreement scores of machines are comparable to those of humans. Specifically, of the 30 statements administered to physicists, the absolute difference in the mean agreement scores between humans and machines is less than 10% for 21 statements and less than 5% for 10 statements, reaching its minimum of 0.3% for S13 and its maximum of 24.6% for S23. Similarly, for philosophers of science, this absolute difference is less than 10% for 19 statements and less than 5% for 10 statements, reaching its minimum of 0.2% for S1 and its maximum of 21.4% for S27. The similarities of the mean agreement scores assigned by humans and machines are further highlighted in Supplementary Figure 1 and, as suggested by the additional analysis reported in Supplementary Figure 2, are not affected by any systematic bias.

We observe that the standard deviations of the mean agreement scores of humans are consistently larger than those of machines, as illustrated in Figure 2 and Supplementary Figure 3, with their dif-

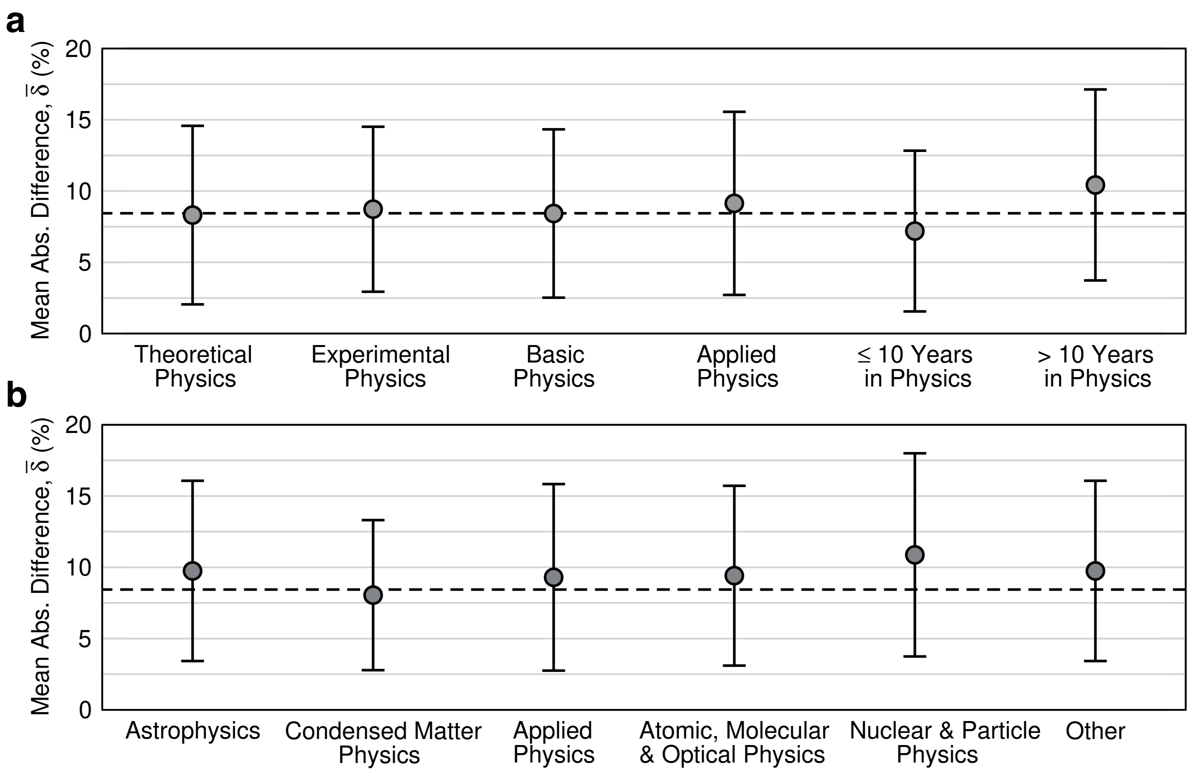

Figure 3: **Comparison between humans and machines across sub-populations.** Mean absolute difference in mean agreement scores ($\overline{\delta}$) between populations of humans and machines, as defined in Equation 1, across different sub-populations of physicists, grouped according to their **a,** research methodology, scope, or experience, and **b,** area of expertise. The dashed horizontal line denotes the global absolute difference obtained for the entire population of physicists, $\overline{\delta} = 8.4\%$.

ference averaging to 14.7% and 12.2% for physicists and philosophers of science, respectively. Importantly, the standard deviation of the mean agreement scores of machines overlaps considerably with those of humans across all statements. This indicates that the philosophical views held by a population of machines resemble very closely those held by a population of human physicists or philosophers of science. We quantify the global discrepancy between humans and machines by determining the mean absolute difference,

$$\overline{\delta} =< \overline{s}_{\mathrm{H}} > - < \overline{s}_{\mathrm{M}} >, \tag{1}$$

where $\overline{s}_{\mathrm{H}}$ and $\overline{s}_{\mathrm{M}}$ are the mean values of the agreement scores assigned by the human and machine populations, respectively, shown in Figure 2. We obtain $\overline{\delta} = 8.4\% \pm 6.0$ for physicists and $\overline{\delta} =$

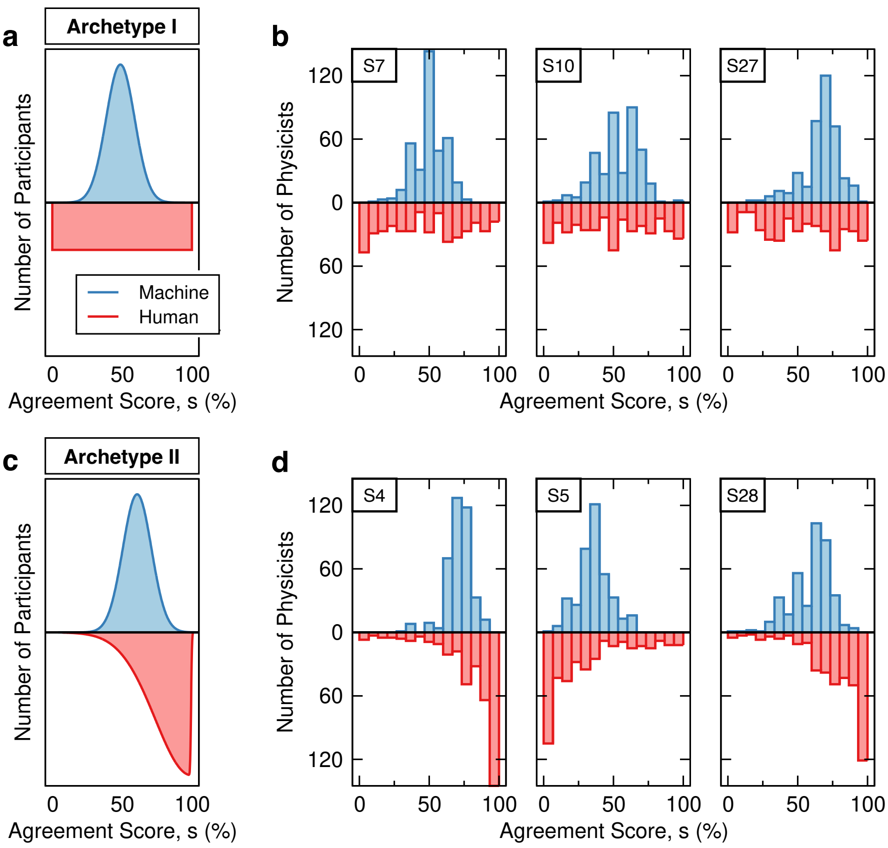

Figure 4: **Patterns underlying the statistical distributions of the responses of humans and machines.** Schematic illustration of the **a,** archetype I, along with **b,** three actual examples realizing it (i.e., the statistical distribution of the responses of physicists to S7, S10, and S27), and the **c,** archetype II, along with **d,** three actual examples realizing it (i.e., the statistical distribution of the responses of physicists to S4, S5, and S28).

$8.9\% \pm 6.2$ for philosophers of science. This demonstrates that machines parallel the judgments of both human physicists and philosophers of science equally well, despite their distinct academic profiles.

To better understand the capability of machines to emulate various sub-populations of humans, depending on their academic and research background, we examine the granularity of our results. In Figure 3, we show the mean absolute differences in the mean agreement scores of several groups of participants, calculated using Equation 1 on the data displayed in Supplementary Figures 4, 5, and 6. We inspect sub-populations of physicists differing in research methodology (theoretical vs. experimental physics), scope (basic vs. applied physics), and level of experience (less vs. more than ten years), as well as for various areas of expertise. The mean absolute difference is quite insensitive to the specific sub-population of physicists, in that its variation exceeds the global value of 8.4% only by 2.5% at most. The largest variations are observed in the case of physicists with more than ten years of experience and those working in the field of nuclear and particle physics. An analogous conclusion is reached for philosophers of science, for which equivalent results are shown in Supplementary Figure 7.

Despite these strong similarities in the mean agreement scores assigned by the population of humans and machines, further analysis reveals significant differences in their statistical distributions, as detailed in Supplementary Figures 8 and 9 for physicists and philosophers of science, respectively. Specifically, we identify two main patterns underlying the vast majority of these distributions. These patterns are described through the two distinct archetypes — referred to as archetypes I and II — depicted in Figure 4 along with actual examples. In both archetypes, the responses of the population of machines exhibit a normal-like distribution of the agreement scores. However, humans manifest qualitatively different trends. In archetype I, the agreement scores of humans are uniformly distributed across the statements. In archetype II, the agreement scores are unevenly distributed, peaking at either end of the scale of the agreement score, thus resembling a skew-normal distribution. From a visual inspection of Supplementary Figures 8 and 9, we note that archetype II is approximately twice as recurrent as archetype I.

To gain insights into the philosophical positions subscribed by human and machine populations, we focus on the two primary and opposing views in the realism debate, that is, scientific realism and instrumentalism. In Figure 5(a,b), we compare the distribution of the individual mean agreement scores assigned by humans and machines to the realist statements listed in Table 1 (i.e., S1, S2, S4, S6, S8, S10, S11, S13, S14, S15, S17, S18, S20, S22) for both physicists and philosophers of science. These distributions are reminiscent of the archetypes displayed in Figure 4(a,c), with the population of machines exhibiting a normal-like distribution while the population of humans exhibiting an approximately uniform distribution, slightly skewed toward the highest values of agreement score. An analogous trend, albeit shifted to lower agreement scores, is observed when considering only the instrumentalist statements listed in Table 1 (i.e., S3, S5, S7, S9, S12, S16, S23, S25), as shown in Supplementary Figure 10. The statistical distributions of the mean agreement scores assigned by several sub-populations of physicists to representative realist and instrumentalist statements are provided in Supplementary Figures 11 and 12, respectively.

To quantify the extent to which the populations of humans and machines favor scientific realism over instrumentalism, we determine the realism score, following the definition Henne and coworkers

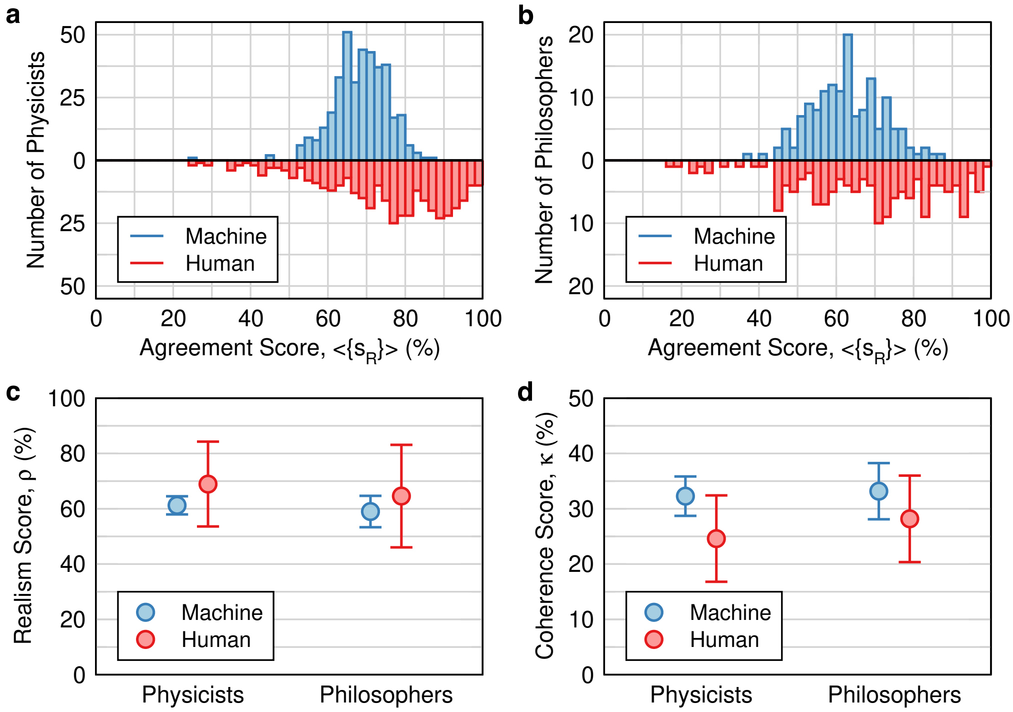

Figure 5: **Realist stance and internal coherence of humans and machines.** Statistical distribution of the individual agreement scores assigned only to the realist statements ($<\{s_\mathrm{R}\}>$) listed in Table 1 by **a,** physicists and **b,** philosophers of science. **c,** Realism score (ρ), as defined in Equation 2, and **d,** coherence score (κ), as defined in Equation 3, for physicists and philosophers of science.

[Henne et al., 2024], as

$$\rho = < \{s_\mathrm{R}\}, -\{s_\mathrm{I}\} >, \tag{2}$$

where $\{s_\mathrm{R}\}$ is the set of the agreement scores assigned selectively to the realist statements and $\{s_\mathrm{I}\}$ is the set of the agreement score assigned selectively to the instrumentalist statements listed in Table 1. By construction, the realism score can range from 0%, indicating a strict instrumentalist position, to 100%, indicating a strict realist position. The realism scores pertaining to humans and machines are compared in Figure 5(c) and listed in Supplementary Table 2. Humans are more inclined toward realism than machines, with the realism score of the former being greater than the latter by 7.7% for physicists and 5.6% for philosophers of science. Importantly, we note that machines correctly reproduce the trend of human physicists being more realist than philosophers of science, although the difference in realism scores between physicists and philosophers is less pronounced for machines (for which it attains the value of 2.3%) than humans (for which it attains the value of 4.3%). We have confirmed that each pairwise distribution of realism scores is statistically distinct by verifying that the p-value is significantly lower than 0.05.

A key aspect in establishing a robust philosophical stance is the internal coherence, that is, the minimization — or, ideally, removal — of inherent contradictions within a given position. In the case of the realism-instrumentalism dichotomy, this translates to the assignment of agreement scores that are compatible across the statements describing the same philosophical view. Specifically, an internally coherent position would stem from high agreement scores when evaluating the realist (instrumentalist) statements and low agreement scores when evaluating the opposite instrumentalist (realist) statements. To quantify the internal coherence, we introduce a coherence score,

$$\kappa = \max[\sigma(\rho)] - < \sigma(\rho) >, \tag{3}$$

where $\sigma(\rho)$ is the standard deviation pertaining to the agreement score associated with the statements included in the determination of the realism score (cf. Equation 2) and $\max[\sigma(\rho)]$ is its maximum possible value, 50%. The coherence score can range from 0%, signaling a random distribution of the agreement scores across the statements and a consequent complete incoherence, to 50%, signaling an absolute internal coherence within one of the two contrasting philosophical views. The coherence scores are shown in Figure 5(d) and listed in Supplementary Table 2. Notably, in their philosophical positions, machines are considerably more coherent than humans, with the difference in coherence scores between the former and the latter being 7.7% for physicists and 5.0% for philosophers of science. Importantly, machines are capable of replicating the trend of human philosophers being more internally coherent than physicists, although this difference is less pronounced for machines (0.9%) compared to humans (3.6%), as previously observed in the case of the realism score.

5 Concluding remarks

In the spirit of the Turing test, we have developed a novel methodological framework to determine whether and to what degree a population of machines mirrors the philosophical views endorsed by a population of humans. Our methodology consists of three steps. First, a population of machines is instructed to impersonate a population of humans, emulating the background of each individual. Second, humans and machines are administered a questionnaire designed to survey various philosophical positions, with each participant rating their agreement with a series of statements. Third, the outcome of the survey is statistically analyzed to compare the views held by the populations of humans and machines.

Drawing from a recent study of Henne and coworkers [Henne et al., 2024], we have employed this methodology in the case study of scientific realism, a long-standing philosophical debate seeking to understand if reality is as science describes it. We have examined the philosophical positions of a population of over 500 hundred humans, comprising both physicists and philosophers of science, and the corresponding populations of machines, generated by means of a popular large-language generative AI model via prompt engineering. Our analysis has revealed that a population of machines endorses philosophical views that are, on average, very similar to those held by the corresponding population of humans, regardless of whether the respondents are physicists or philosophers of science. Although our work does *not* constitute a Turing test, the analogy of the philosophical judgments held by the populations of humans and machines — as corroborated by the invariably overlapping error bars of the metrics used to quantify them — implies that a hypothetical human judge is likely to fail to discern the nature of the two populations. Importantly, machines are able to reflect the nuances distinguishing the views of philosophers from those of physicists.

We have additionally observed that, as compared to humans, machines exhibit a weaker inclination toward scientific realism and a stronger coherence in their philosophical positions. Because the realism-instrumentalism debate is inherently underdetermined — in that no *experimentum crucis* can be conceived to decisively adjudicate between these two opposed views — there is no single 'true' interpretation of the relationship between science and reality that can serve as a benchmark to establish the exactness of the philosophical views held by humans or machines. However, if internal coherence is regarded as a measure of the strength of a given philosophical position, then one may provocatively suggest that machines are better philosophers than humans.

To conclude, our methodology can be readily applied to other surveys of philosophical interests, such as those assessing the views of a wide variety of scholars on, e.g., theoretical virtues [Schindler, 2022], values in science [Steel et al., 2017], and scientific practices [Robinson et al., 2019], as well as on broader issues [Bourget and Chalmers, 2014, Bourget and Chalmers, 2023] such as philosophical intuition [Kuntz and Kuntz, 2011]. Given the close resemblance between human and machine populations in their average responses to philosophical questions, our methodological framework may open a new avenue to advance research in experimental philosophy by deploying a population of machines in lieu of a target population of humans, potentially accelerating the administration of otherwise tedious survey-based studies and mitigating the reproducibility issues plaguing them [Cova et al., 2021].

Statements and Declarations

Funding: N/A

Conflict of Interest: N/A

Ethical Approval: N/A

Informed Consent: N/A

Author Contribution: M. P. designed and performed research, analyzed data, and wrote the paper. G. D. performed research and analyzed data.

Data Availability Statement: The data will be made publicly available following acceptance of the manuscript.

References

[Bartels and Pizarro, 2011] Bartels, D. M. and Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, 121(1):154–161.

[Birhane et al., 2023] Birhane, A., Kasirzadeh, A., Leslie, D., and Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280.

[Block, 1981] Block, N. (1981). Behaviourism and psychologism. *Philosophical Review*, 90(1):5–53.

[Bourget and Chalmers, 2014] Bourget, D. and Chalmers, D. J. (2014). What do philosophers believe? *Philosophical Studies*, 170(3):465–500.

[Bourget and Chalmers, 2023] Bourget, D. and Chalmers, D. J. (2023). Philosophers on philosophy: The 2020 PhilPapers survey. *Philosophers' Imprint*, 11(23):1–52.

[Buckwalter and Stich, 2013] Buckwalter, W. and Stich, S. (2013). Gender and philosophical intuition. In Knobe, J. and Nichols, S., editors, *Experimental Philosophy: Volume 2*, pages 307–346. Oxford University Press.

[Cartwright, 1983] Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.

[Chakravartty, 2010] Chakravartty, A. (2010). *A metaphysics for scientific realism: Knowing the unobservable*. Cambridge University Press.

[Chang, 2012] Chang, H. (2012). *Is water H_2O? Evidence, realism and pluralism.* Boston Studies in the Philosophy and History of Science.

[Cova et al., 2021] Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., Jaquet, F., Khalifa, K., Kim, H., Kneer, M., Knobe, J., Kurthy, M., Lantian, A., Liao, S.-y., Machery, E., Moerenhout, T., Mott, C., Phelan, M., Phillips, J., Rambharose, N., Reuter, K., Romero, F., Sousa, P., Sprenger, J., Thalabard, E., Tobia, K., Viciana, H., Wilkenfeld, D., and Zhou, X. (2021). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 12(1):9–44.

[Damassino, 2020] Damassino, N. (2020). The questioning Turing test. *Minds and Machines*, 30(4):563–587.

[Dodge and Karam, 2017] Dodge, S. and Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7.

[Duke and Bègue, 2015] Duke, A. A. and Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition*, 134:121–127.

[French, 2000] French, R. M. (2000). The Turing test: The first 50 years. *Trends in cognitive sciences*, 4(3):115–122.

[Giere, 2006] Giere, R. N. (2006). *Scientific Perspectivism.* University of Chicago Press.

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

[Hannikainen et al., 2018] Hannikainen, I. R., Machery, E., and Cushman, F. A. (2018). Is utilitarian sacrifice becoming more morally permissible? *Cognition*, 170:95–101.

[Harnad, 1991] Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1(1):43–54.

[Henne et al., 2024] Henne, C., Tomczyk, H., and Sperber, C. (2024). Physicists' views on scientific realism. *European Journal for Philosophy of Science*, 14(1):10.

[James, 2000] James, W. (2000). *Pragmatism and Other Writings*. Penguin Books.

[Kasai et al., 2022] Kasai, J., Sakaguchi, K., Dunagan, L., Morrison, J., Le Bras, R., Choi, Y., and Smith, N. A. (2022). Transparent human evaluation for image captioning. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478.

[Krenn et al., 2022] Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., Yao, Z., and Aspuru-Guzik, A. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769.

[Kuntz and Kuntz, 2011] Kuntz, J. R. and Kuntz, J. R. C. (2011). Surveying philosophers about philosophical intuition. *Review of Philosophy and Psychology*, 2(4):643–665.

[Ladyman, 1998] Ladyman, J. (1998). What is structural realism? *Studies in History and Philosophy of Science Part A*, 29(3):409–424.

[Laudan, 1981] Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, 48(1):19–49.

[Lipton, 2007] Lipton, P. (2007). The world of science. *Science*, 316(5286):834.

[Massimi, 2018] Massimi, M. (2018). Perspectivism. In Saatsi, J., editor, *The Routledge Handbook of Scientific Realism*. Routledge.

[Massimi, 2022] Massimi, M. (2022). *Perspectival Realism*. Oxford University Press.

[Mitchell, 2024] Mitchell, M. (2024). Debates on the nature of artificial general intelligence. *Science*, 383(6689):eado7069.

[Naveed et al., 2024] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2024). A comprehensive overview of large language models. *ArXiv:2307.06435*.

[Petrinovich and O'Neill, 1996] Petrinovich, L. and O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17(3):145–171.

[Psillos, 1991] Psillos, S. (1991). *Scientific Realism: How Science Tracks Truth*. Routledge.

[Robinson et al., 2019] Robinson, B., Gonnerman, C., and O'Rourke, M. (2019). Experimental philosophy of science and philosophical differences across the sciences. *Philosophy of Science*, 86(3):551–576.

[Rowbottom, 2019] Rowbottom, D. P. (2019). *The Instrument of Science: Scientific Anti-Realism Revitalised*. Routledge.

[Schindler, 2022] Schindler, S. (2022). Theoretical virtues: Do scientists think what philosophers think they ought to think? *Philosophy of Science*, 89(3):542–564.

[Schwartz, 2022] Schwartz, M. D. (2022). Should artificial intelligence be interpretable to humans? *Nature Reviews Physics*, 4(12):741–742.

[Searle, 1980] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

[Sperber, 2023] Sperber, C. (2023). Online materials for "Physicists' views on scientific realism". Mendeley Data, V1.

[Stanford, 2010] Stanford, P. K. (2010). *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford University Press.

[Steel et al., 2017] Steel, D., Gonnerman, C., and O'Rourke, M. (2017). Scientists' attitudes on science and values: Case studies and survey methods in philosophy of science. *Studies in History and Philosophy of Science Part A*, 63:22–30.

[Turing, 1950] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX(236):433–460.

[Watt, 1996] Watt, S. (1996). Naive psychology and the inverted Turing test. *Psycoloquy*, 7(14):463–518.

[Weidinger et al., 2022] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. (2022). Taxonomy of risks posed by language models. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, page 214–229.

[Worrall, 1989] Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica*, 43(1-2):99–124.

[Yan et al., 2023] Yan, M., Xu, H., Li, C., Tian, J., Bi, B., Wang, W., Xu, X., Zhang, J., Huang, S., Huang, F., Si, L., and Jin, R. (2023). Achieving human parity on visual question answering. *ACM Transactions on Information Systems*, 41(3):1–40.