International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Autonomy Outpacing Verification: A Stage-Resolved Framework for Measuring How Artificial Intelligence and Automation Are Changing Physics Research

Samarpan Panchal
Independent Researcher, M.Sc. Physics, Indus University, Ahmedabad, Gujarat, India

Abstract - Artificial intelligence (AI) and laboratory automation now operate inside the core loop of physics research. They design optical experiments, shape fusion plasmas, decode quantum errors, select particle collisions in real time and propose new crystals. Existing reviews catalogue these applications by subfield, yet none measures how much decision authority each system holds or how rigorously its outputs are checked. This study introduces two ordinal scales, the Physics Research Autonomy Levels (PRAL, L0 to L5) and the Verification Levels (VL, V0 to V5), and applies them to sixteen documented systems published between 2016 and 2025 in quantum, plasma, particle, gravitational, condensed matter, materials and atmospheric physics. Each system is coded across six research stages: hypothesis, design, execution, analysis, interpretation and reporting. Three results follow. First, autonomy has advanced furthest in execution, where five physics systems act without per-decision human approval, while hypothesis selection and interpretation remain largely human. Second, reported gains concentrate on closed tasks with a fixed objective and a checkable answer, with accelerations from 16 to 2 billion times and equation recovery of 100%, whereas the strongest language model solves 5.7% of open research-level physics challenges. Third, autonomy and verification show no significant association across the corpus (Spearman rho = 0.14, p = 0.59). All four systems whose autonomy exceeds their verification by two or more levels act in hypothesis, design or interpretation, while no system confined to execution or analysis shows such a deficit. This difference is defined here as the verification gap. A Verification-Proportional Reporting checklist is proposed that scales evidence requirements with autonomy. The analysis indicates that the binding constraint on AI-driven physics has shifted from computational capability to verification capacity.

Keywords: ***Artificial intelligence, physics research, laboratory automation, autonomous experimentation, machine learning, research verification, reproducibility, scientific discovery***

1. INTRODUCTION

Physics has always absorbed new instruments, and each instrument has changed what physicists do with their working hours. The bubble chamber created armies of film scanners. The digital computer created computational physics as a third pillar beside theory and experiment. Machine learning and robotic automation now produce a different kind of change. Earlier instruments extended human perception or calculation. Current AI systems take decisions that physicists previously reserved for themselves: which configuration to try next, which event to keep, which candidate law fits the data and which result deserves to be reported.

The recognition of this shift is institutional. The 2024 Nobel Prize in Physics went to John Hopfield and Geoffrey Hinton for foundational work that enables machine learning with artificial neural networks [1]. In the same year, deep reinforcement learning avoided tearing instabilities in the DIII-D tokamak [2], an unsupervised autoencoder joined the Level-1 trigger menu of the CMS experiment [3], and a transformer decoder outperformed established decoders on surface-code data from a superconducting quantum processor [4].

These successes invite a simple narrative of acceleration. The evidence does not support that narrative without qualification. An autonomous materials laboratory reported 41 new compounds in 17 days [5], and an independent reanalysis concluded that roughly two thirds of the claimed successes were likely known compositionally disordered phases [6]. A survey of machine learning based science found data leakage in at least 294 papers across 17 fields [7]. A large bibliometric study of 41.3 million papers found that AI adoption raises individual output and citations while contracting the topical breadth of science [8].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

The gap in current understanding is therefore not a shortage of examples. It is the absence of a common measure that states, for any AI system in physics, how much of the research process it controls and how well its outputs are independently checked. Without such a measure, a neural network that classifies light curves and an agent that writes and submits complete papers appear in the same literature under the same label of "AI for physics".

This study addresses that gap with three contributions. It defines a six-level autonomy scale and a six-level verification scale tailored to physics research stages. It applies both scales to sixteen documented systems and visualises where autonomy has arrived. It introduces the verification gap as a diagnostic quantity and derives from it a reporting checklist that journals and research groups can adopt directly. The paper does not claim that AI accelerates physics uniformly. It measures where acceleration is real, where it is untested and where its claims have failed scrutiny.

2. RELATED RESEARCH WORK

2.1 Reviews of machine learning in the physical sciences

Carleo and co-authors provided the reference review of the interface between machine learning and physics, covering statistical physics of learning, particle physics, cosmology, quantum many-body systems and materials [9]. Their review organises the field by subfield and by algorithm. Krenn, Erhard and Zeilinger reviewed computer-inspired design of quantum experiments and showed that algorithms frequently return asymmetric configurations that human designers would not consider [10]. These reviews establish breadth. They do not grade the degree of autonomy or the strength of validation, which are the two variables this study isolates.

2.2 Landmark capability demonstrations

Several results define the capability frontier used in this analysis. Carleo and Troyer represented many-body wave functions with neural networks and reached high accuracy on prototypical spin models in one and two dimensions [11]. The AI Feynman algorithm combined neural fitting with dimensional analysis and symmetry tests and recovered all 100 equations from the Feynman Lectures, against 71 for the best prior public software [12]. DINGO used neural posterior estimation to infer gravitational-wave source parameters in about 20 seconds per event, compared with order-one-day runtimes for standard samplers, while matching their posteriors on eight catalogue events [13]. Degrave and colleagues trained a single reinforcement learning policy in simulation and used it to command all control coils of the TCV tokamak, producing a range of plasma shapes on real hardware, including elongated configurations of the kind planned for ITER [14].

2.3 Automation of experiments and facilities

Autonomous experimentation at large facilities matured in parallel. Gaussian-process steering through the gpCAM package let synchrotron beamlines choose their own next measurement, first in infrared spectromicroscopy and photoemission studies at the Advanced Light Source [15, 16]. Noack and colleagues generalised the method for synchrotron and neutron facilities [17]. In particle physics, the hls4ml library made it possible to compile neural networks into FPGA firmware fast enough for hardware triggers [18]. The CMS anomaly detection trigger AXOL1TL runs inference in about 50 ns on each 40 MHz bunch crossing [19].

2.4 Discovery systems and their critiques

Materials physics produced the most visible discovery claims and the most detailed critiques. GNoME, a graph network trained through active learning, generated 2.2 million candidate crystal structures and contributed about 380,000 predicted-stable entries to the Materials Project [20]. Cheetham and Seshadri argued that many GNoME entries depend on cation orderings unlikely in real crystals and that no functionality had been shown for them [21]. The A-Lab coupled such predictions to robotic synthesis [5], and Leeman and colleagues contested both its novelty criterion and its automated phase identification [6]. In gravitational-wave instrumentation, the Urania optimiser proposed interferometer topologies that may exceed current designs, some of which, according to the research group, are not yet fully understood [22, 23].

2.5 End-to-end research agents and their limits

Language-model agents now attempt the complete research cycle. The AI Scientist-v2 generated three workshop papers without human editing; one received an average reviewer score of 6.33 at an ICLR 2025 workshop, which its authors describe as above the average acceptance threshold of a venue with a lower bar than the main conference [24, 25]. AlphaEvolve, an evolutionary coding

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

agent, found a 48-multiplication algorithm for 4 x 4 complex matrices, the first improvement on Strassen's 49 in 56 years [26]. On physics specifically, the CritPt benchmark poses 71 unpublished research challenges written by more than 50 physicists, each with a machine-verifiable answer [27].

2.6 Field-level effects

Two bibliometric studies frame the macro-level change. Duede and colleagues analysed about 80 million publications across 20 fields and found roughly thirteenfold growth in AI-engaged work, with AI papers spreading across fields while remaining semantically separate from non-AI work [28]. Hao, Xu, Li and Evans found that scientists adopting AI publish 3.02 times as many papers, receive 4.84 times as many citations and become research leaders about 1.4 years earlier, while the knowledge extent of AI-augmented research contracts [8].

2.7 Gap addressed by this study

The literature above answers "what can AI do in physics" and "what does AI do to scientific careers". It does not answer "how much research authority has each system been given, and has its verification kept pace". Autonomy taxonomies exist in other domains, most prominently the six driving automation levels of SAE J3016 [29], but no stage-resolved autonomy scale paired with a verification scale has been proposed for physics. This paper supplies one and tests what it reveals.

3. RESEARCH METHODS

3.1 Design

The study is a structured comparative case analysis. It combines a targeted literature review with ordinal coding of each case on two independent scales. The unit of analysis is a single AI or automation system as described in its primary peer-reviewed publication or, for collaboration-internal systems, in official collaboration documentation.

3.2 Case selection

Cases were included if they met four criteria: (i) the system performs a task within physics or a physics-based instrument; (ii) the primary description appeared between 2016 and 2025; (iii) the source reports a quantitative outcome; and (iv) the case is widely cited or adopted, which ensures that it shaped practice. Fifteen physics cases met these criteria. One additional system, the AI Scientist-v2, operates in machine learning rather than physics. It was retained as a reference case because it is the only documented system that spans all six research stages, and it therefore anchors the top of the autonomy scale. All results are reported both with and without this reference case where the distinction matters.

3.3 Research stages

Each case was coded across six stages that together describe a physics research cycle. Hypothesis covers selection of the question, target or candidate. Design covers choice of the experimental or computational configuration. Execution covers operation of the apparatus, simulation or control system. Analysis covers inference from raw data to quantities. Interpretation covers judgement of meaning, novelty or physical law. Reporting covers composition and submission of the scientific record.

3.4 Physics Research Autonomy Levels

Table 1 defines the autonomy scale. Its structure follows the logic of graded delegation used in driving automation [29] but replaces road-specific criteria with research-specific ones. The decisive boundary lies between L3 and L4. At L3 a human approves the system's output before it changes anything. At L4 the system acts on the world without per-decision approval.

Table 1. Physics Research Autonomy Levels (PRAL).

Level	Name	Definition	Physics example
L0	Manual	Humans perform the stage; software records data and computes only on explicit instruction.	Hand-tuned feedback loop; manual curve fitting
L1	Assisted	AI completes a subtask on request; a human checks every output.	Code completion; AI-assisted literature search

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Level	Name	Definition	Physics example
L2	Partial	AI performs a defined task end to end within one stage; humans review results in bulk.	Neural classifier screening transit light curves
L3	Conditional	AI generates candidates or decisions in a bounded space; a human approves before anything changes in the world or the record.	Algorithm proposes optical layouts that physicists select and build
L4	High	AI acts within a stage without per-decision approval, under human-set objectives and constraints.	Reinforcement learning policy commanding tokamak coils in real time
L5	Full	AI chains stages from question selection to reporting; humans set only the topic.	End-to-end research agent that writes and submits papers

3.5 Verification Levels

Table 2 defines the verification scale. Levels credit the strongest evidence reported for the system's central claim. One rule overrides the ladder: a published independent reanalysis that disputes the central claim caps the level at V1 until the dispute is resolved. V5 is reserved for independent replication reported in the literature reviewed; no case received it under the conservative evidence boundary used here.

Table 2. Verification Levels (VL).

Level	Name	Evidence credited for the central claim
V0	None	No validation reported.
V1	Self-referential	Output judged by the generating pipeline or by a low-threshold review, or the central claim is disputed by a published independent reanalysis.
V2	Computational cross-check	Agreement with an independent computational method or with known ground truth on synthetic or benchmark data.
V3	Reference agreement on real data	Agreement with an established gold-standard method on real experimental or observational data.
V4	Physical realisation	The output operates in, or is confirmed by, real hardware or a laboratory experiment.
V5	Independent replication	An independent group reproduces the central claim.

3.6 Derived quantities

The overall autonomy of a case, A, is its highest stage level. The verification of a case, V, is its VL score. The verification gap is G = A minus V. A positive G means that a system holds more research authority than its evidence base supports. A G of two or more is classed as a verification deficit. Rank association between A and V was measured with the Spearman coefficient.

3.7 Task closure

To compare capability across systems with different metrics, each reported performance figure was classified as a closed or open task. A closed task has a fixed objective, a bounded output space and an answer that can be checked automatically against ground truth or a reference method. An open task requires the system to define intermediate goals, chain multiple reasoning steps and produce an answer that cannot be checked by a single metric. Acceleration factors were taken directly from the primary sources and were not recalculated, except that the DINGO factor was converted from "order one day to 20 seconds" to approximately 4,300 times.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

3.8 Coding procedure and limits of the method

The author coded all cases from the primary publications and their published critiques. Table 3 records the rationale for each code so that readers can audit or recode. Single-coder ordinal coding is subjective, and the sample is purposive rather than random. The results are therefore presented as structured evidence for hypotheses, not as population estimates.

4. RESULTS

4.1 Chronology of the corpus

Figure 1 orders the main events in the corpus. Before 2020, AI systems in physics acted mostly as analysis tools on fixed datasets. From 2021 onward, systems began to act on hardware in closed loops, and from 2024 onward, language-model agents appeared that attempt open research tasks.

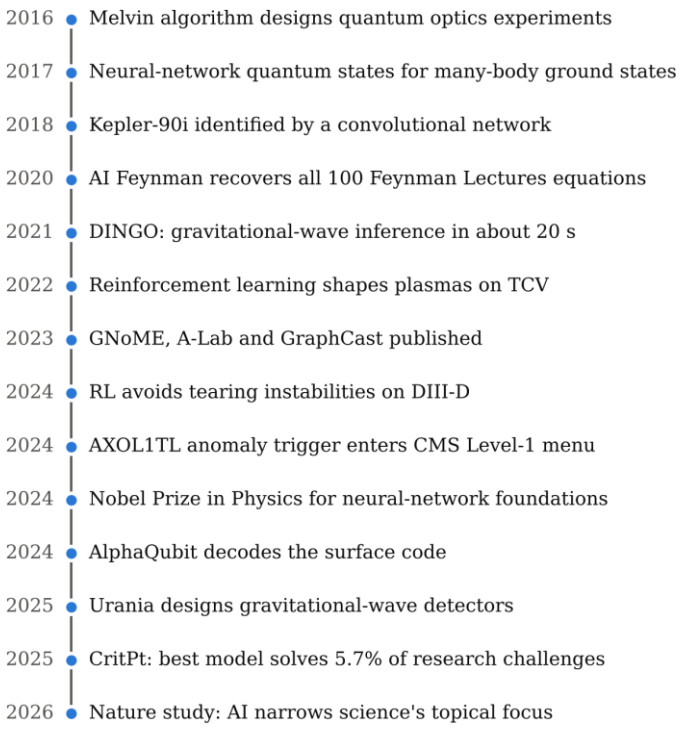


Figure 1. Milestones of AI and automation in physics research, 2016 to 2026, as covered by the case corpus and the bibliometric sources.

4.2 Coding results

Table 3 lists the codes and rationale for each case. Mean overall autonomy across the fifteen physics cases is 3.13 and mean verification is 2.67. Including the reference case, the values are 3.25 and 2.56.

Table 3. Case coding. A = highest PRAL level, V = VL score, G = A minus V.

Case	Domain	A	V	G	Coding rationale
Melvin [30]	Quantum optics	3	4	-1	Proposes setups for target entangled states; physicists select; several realised on optical tables.
NQS [11]	Many-body theory	2	2	0	Neural wave function computes ground states and dynamics; checked against exact results for spin models.
Kepler-90i CNN [31]	Exoplanet astrophysics	2	3	-1	Trained on about 15,000 vetted Kepler signals, 96% correct classification; candidates vetted on real data [32].
AI Feynman [12]	Law discovery	3	2	1	Proposes symbolic laws from data; validated on equations with known ground truth.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Case	Domain	A	V	G	Coding rationale
NAS emulators [33]	Plasma, climate, astrophysics	2	2	0	Replaces forward simulations; accuracy and uncertainty checked against the simulators.
DINGO [13]	Gravitational waves	3	3	0	Amortised posteriors for real catalogue events; agree with standard samplers.
gpCAM [15]	Synchrotron experiments	4	4	0	Chooses each next measurement without approval; campaigns run on ALS and NSLS-II beamlines.
TCV control [14]	Fusion plasma	4	4	0	Policy commands 19 coil voltages in closed loop; validated in TCV discharges.
GNoME [20]	Materials physics	3	1	2	Selects candidate crystals; DFT stability check; central claim disputed in published critique [21].
A-Lab [5]	Materials synthesis	4	1	3	Plans recipes, synthesises and identifies phases autonomously; novelty disputed by reanalysis [6].
GraphCast [34]	Atmospheric physics	3	3	0	Forecasts verified against HRES on 1,380 targets using reanalysis data.
DIII-D control [2]	Fusion plasma	4	4	0	Controller acts on actuators to avoid tearing modes; validated in DIII-D discharges.
AXOL1TL [19]	Particle physics	4	3	1	Keeps or discards events at 40 MHz without approval; validated through rate and data monitoring.
AlphaQubit [4]	Quantum computing	3	3	0	Decodes syndromes; outperforms reference decoders on real Sycamore data.
Urania [22]	Gravitational-wave instruments	3	1	2	Proposes detector topologies; sensitivity from simulation only; designs not yet built.
AI Scientist-v2 [24]	Machine learning (reference)	5	1	4	Runs all six stages; evidence is acceptance at a workshop with a lower bar.

4.3 Where autonomy has arrived

Figure 2 maps the autonomy level of each case at each stage. Three patterns stand out. Analysis is the most frequently occupied stage, with eight entries, and it sits mostly at L2 and L3: the system produces inferences, and humans accept them in bulk or select among them. Execution is the stage where autonomy runs highest. Five of the six physics systems active in execution operate at L4, acting on beams, coils, triggers or furnaces without per-decision approval. Hypothesis, interpretation and reporting remain sparsely occupied. Only GNoME selects hypotheses within physics, only AI Feynman and the A-Lab make interpretive claims, and no physics system composes the scientific record.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

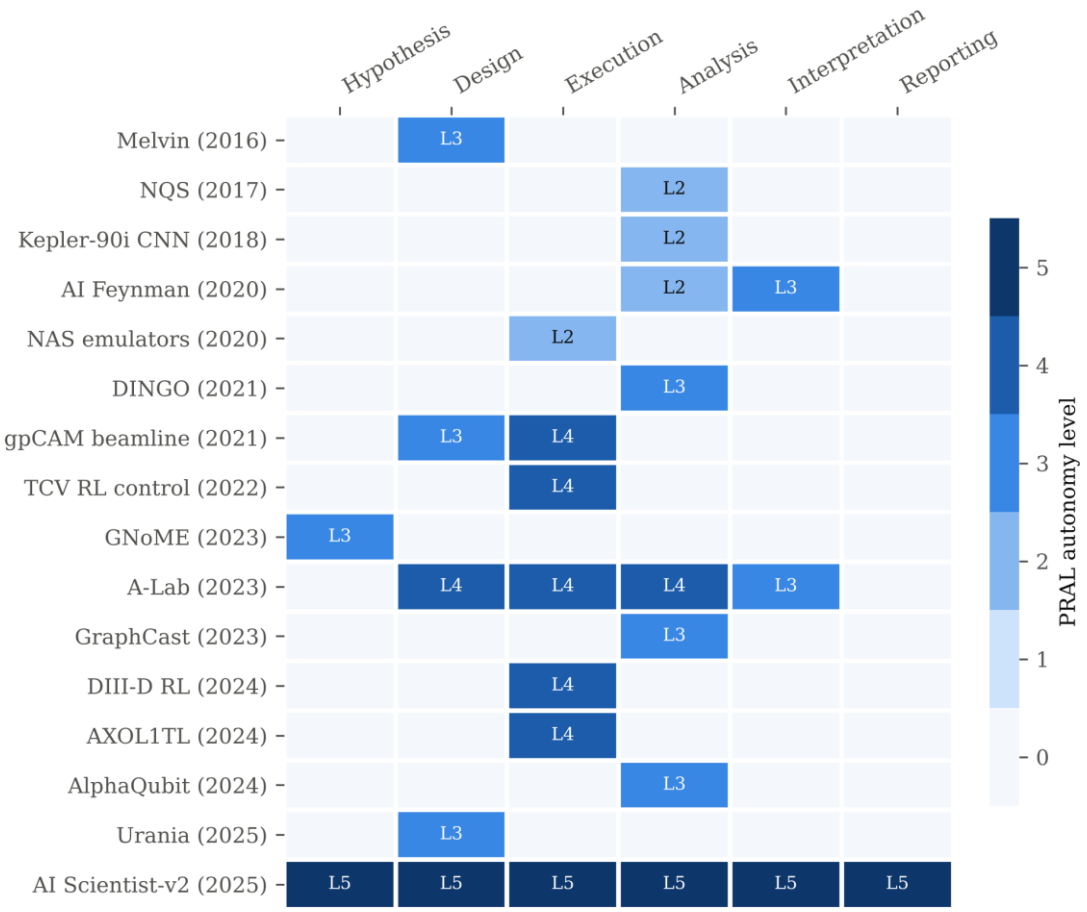


Figure 2. Stage-resolved autonomy map. Each cell shows the PRAL level a system holds at a research stage; blank cells indicate no role.

The pattern has a physical explanation. Execution tasks in plasma control, beam steering and triggering are governed by well-validated simulators or by objectives that can be scored within milliseconds. A policy trained in a free-boundary equilibrium simulator could therefore be deployed on TCV directly [14], and a tearing-instability predictor trained on past DIII-D discharges could serve as the training environment for a controller that forecasts instabilities up to 300 ms ahead [35]. Hypothesis and interpretation lack such scoring functions. Whether a crystal is new, or a regularity is a law, cannot be computed by the system that proposed it.

4.4 Measured acceleration

Figure 3 collects the acceleration factors reported by the primary sources. They span more than eight orders of magnitude. Neural architecture search emulators reached up to 2 billion times speed-up over the simulations they replace in ten scientific cases, four of them in plasma physics [33]. DINGO reduced per-event gravitational-wave inference by roughly 4,300 times [13]. Robotic synthesis in the A-Lab handled 50 to 100 times more samples per day than a typical researcher [36]. Gaussian-process steering cut the measurement count fiftyfold and the wall time sixteenfold [15, 16]. GraphCast produces a 10-day global forecast at 0.25 degree resolution in under a minute on one TPU [34], where the operational HRES system needs hours on a supercomputer with hundreds of machines [37].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

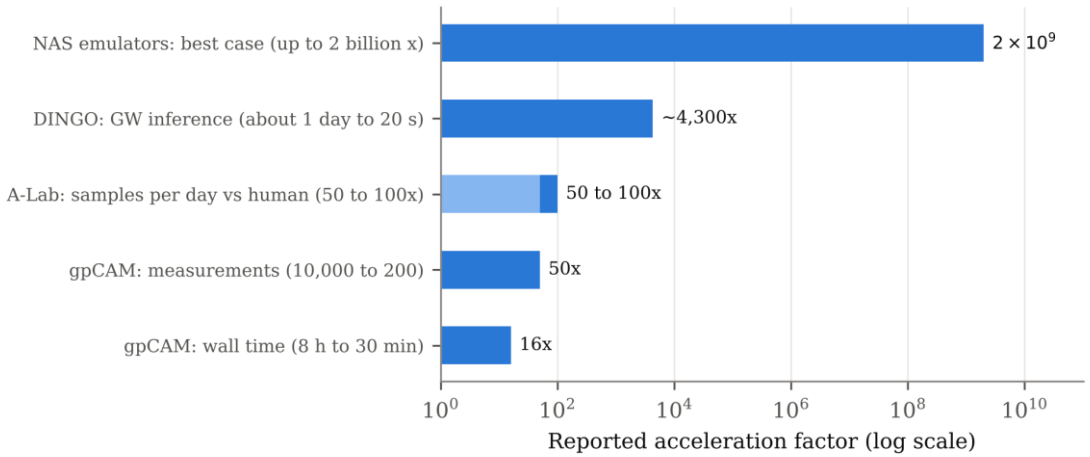


Figure 3. Acceleration factors reported in the primary sources (log scale). The lighter segment of the A-Lab bar marks the lower bound of the reported range.

These figures are not comparable as measures of scientific value. An emulator speed-up applies to forward evaluation after training, and it excludes the cost of producing training simulations. A reduction in measurement count matters only if the resolved features are the ones physically relevant. The factors are nonetheless consistent in one respect: every large gain occurs in a closed task.

4.5 The closure boundary

Figure 4 places closed-task and open-task performance on one axis. On closed tasks the reported success rates are high. Beyond its complete recovery of the Feynman set, AI Feynman solved 90% of a harder physics test set on which the prior best method solved 15% [12], and GraphCast outperformed the operational HRES system on 90% of 1,380 verification targets [34]. On open research tasks, CritPt records 5.7% for the strongest base model and about 10% with coding tools [27].

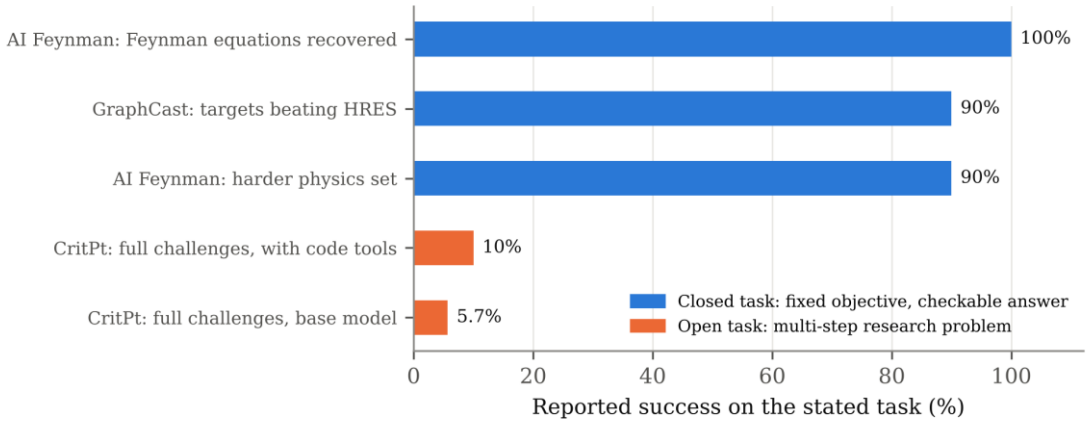


Figure 4. Reported success on closed tasks versus open research tasks. Metrics differ across sources; the figure compares the share of the stated task each system completes.

The contrast is the central capability finding of this study. Performance in physics AI does not scale primarily with model size or data volume. It scales with closure: whether the task has a fixed objective and an automatic check. This explains why the most autonomous physics systems in Figure 2 cluster in execution, where closure is high, and why autonomous interpretation has produced the most disputed outcomes.

4.6 The verification gap

Figure 5 plots autonomy against verification for every case. The two variables show no significant association: Spearman rho = 0.14 across sixteen cases ($p = 0.59$) and rho = 0.38 across the fifteen physics cases ($p = 0.17$). Higher autonomy has not reliably brought stronger verification with it.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

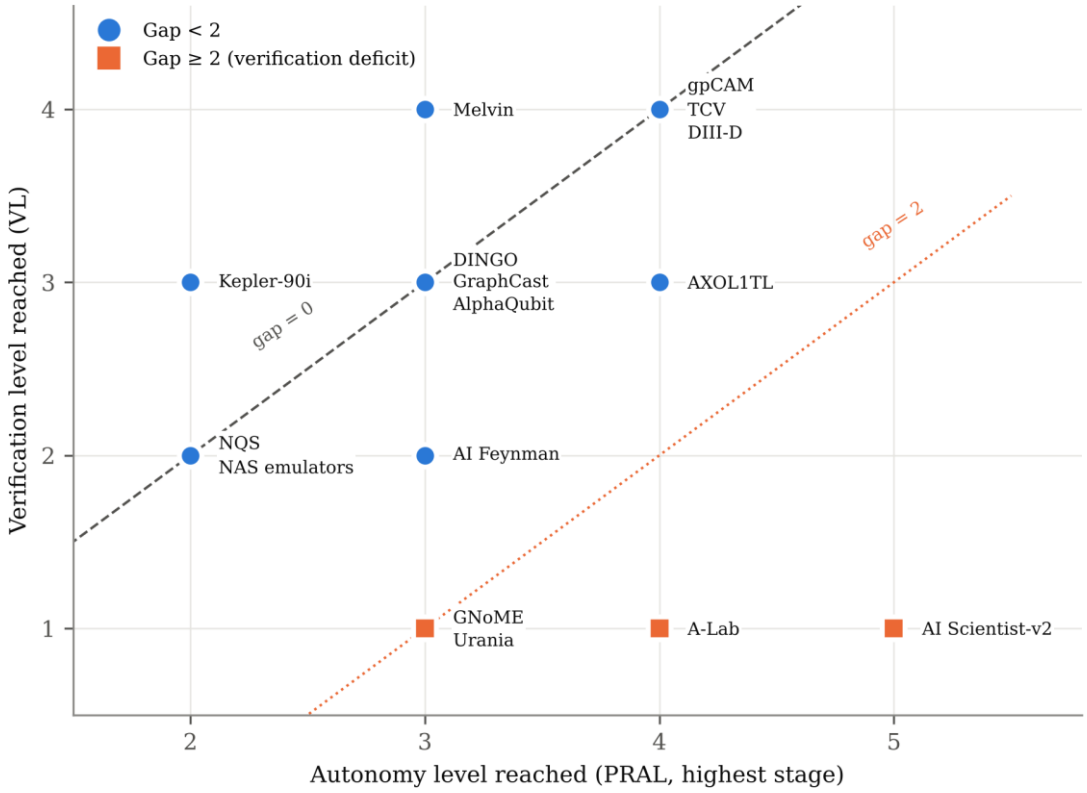


Figure 5. Autonomy versus verification. Points on the dashed line have zero gap. Points on or below the dotted line carry a verification deficit (G of two or more).

Four cases carry a verification deficit. The A-Lab (G = 3) reached L4 in design, execution and analysis, and its phase identification relied on its own automated diffraction analysis; an independent reanalysis then disputed most of its novelty claims [6]. GNoME (G = 2) selected candidates at L3 and validated them with density functional theory, a computational check that a published critique judged insufficient to establish real, ordered compounds [21]. Urania (G = 2) proposed detector designs whose sensitivity rests on simulation alone, and none has yet been built [22]. The AI Scientist-v2 (G = 4) holds L5 across all stages, and its only external evidence is the workshop review described in Section 2.5.

Two of these four codes reflect the dispute rule of Section 3.5, so the presence of critiques among deficit cases is partly a consequence of the method. The stage pattern is not. Every deficit case holds autonomy in hypothesis, design or interpretation, the stages with the lowest closure. No system confined to execution or analysis shows a deficit, even at L4. All four systems that reached V4 did so through physical realisation: the Melvin setups built on optical tables [30], the TCV and DIII-D controllers [14, 2] and the ALS beamline campaigns [15]. Physical deployment is currently the most reliable route to closing the gap, and it is available precisely in the stages where closure is already high.

4.7 Field-level effects

The case-level pattern has a counterpart at the scale of whole fields. Figure 6 shows the individual advantages measured by Hao and colleagues for AI-adopting scientists [8]. Only about 0.75% of the 41.3 million papers in their dataset showed traces of AI augmentation [38], so these advantages accrue to a small group. The same study reports a contraction of about 4.6% in the knowledge extent of AI-augmented work [8] and weaker engagement among the scientists who build on it [38]. The pattern matches the semantic separation between AI-engaged and other work reported by Duede and colleagues [28].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

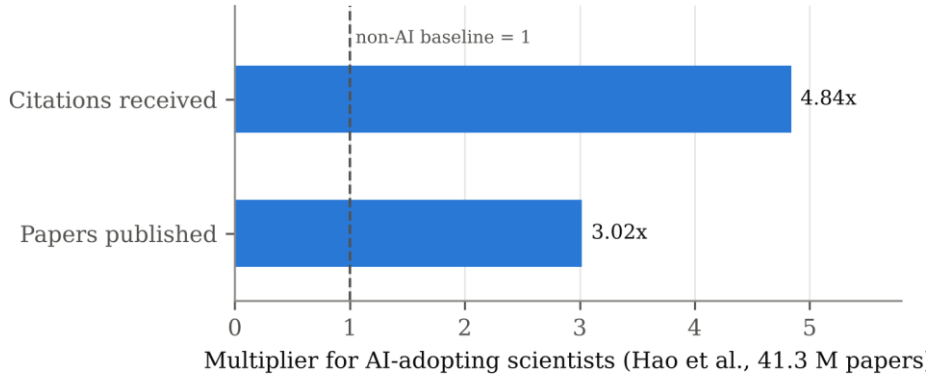


Figure 6. Career-level multipliers for scientists who adopt AI, relative to non-adopters, from a study of 41.3 million natural-science papers.

The two scales connect. Systems concentrate where closure is high, and closure is high where data and simulators already exist. Research effort therefore drifts toward data-rich problems. The result is the paradox the bibliometric data record: faster individual progress inside a narrower collective search.

5. DISCUSSION

5.1 The changing role of the physicist

The corpus documents a reallocation of physicists' work rather than its removal. At L4 the physicist no longer chooses each coil voltage or each beam position. The physicist writes the objective, sets the constraints and certifies the result. In the TCV work, controller design shifted from specifying how to control toward specifying what to control [14]. In AXOL1TL, physicists no longer write the selection logic of a trigger path; they define the training data, monitor rate stability and validate what the network kept [19]. The scarce skills are therefore objective specification, simulator construction and verification design. Physics curricula that teach data analysis but not verification of autonomous systems will train researchers for the stage that automation is absorbing fastest.

5.2 Why closure governs success

Closure combines three conditions: a fast and faithful simulator or reference method, a scalar objective, and dense feedback. Plasma control satisfies all three. Weather prediction satisfies them through four decades of ERA5 reanalysis. Quantum error decoding satisfies them through simulators that generate hundreds of millions of training examples before fine-tuning on hardware data [39]. Open research tasks fail all three: no simulator scores a conjecture, the objective is plural and feedback arrives months later through peer review. The CritPt result should therefore not be read as a temporary gap in model scale alone. It also reflects the absence of an automatic verifier for research-level physics.

A practical corollary follows. The fastest way to extend AI into a new area of physics is often not a larger model but a better verifier: a differentiable simulator, a curated benchmark with machine-checkable answers or an instrument that can test a proposal within hours. Investment in verifiers raises closure, and closure raises the autonomy that can be granted safely.

5.3 Failure modes specific to autonomous physics

The cases reveal four failure modes. The first is self-certification, in which the same pipeline that proposes a result also judges it, as in automated phase identification that cannot recognise disordered variants of ordered predictions [6]. The second is leakage, in which information from the evaluation set enters training and inflates performance, a failure already documented at scale and one that remains invisible in a paper's text alone [7]. The third is opacity at design time: Urania and Melvin both returned configurations that humans found hard to understand [22, 30]. Opaque designs can still be correct, but they cannot be audited by reasoning, only by building them. The fourth is topical narrowing, in which convenience rather than importance guides question selection [8].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

5.4 Verification-Proportional Reporting

The verification gap suggests a direct remedy: evidence requirements should scale with autonomy. Table 4 sets out a Verification-Proportional Reporting (VPR) checklist. Each autonomy level inherits the requirements of the levels below it. The checklist is designed for adoption by journals as a submission requirement and by research groups as an internal gate before deployment on hardware.

Table 4. Verification-Proportional Reporting (VPR) checklist. Each level inherits all requirements above it.

Autonomy	Minimum evidence before publication or deployment
L1 to L2	Report data splits, leakage checks and a model information sheet [7]; compare with a non-AI baseline on identical data.
L3	Evaluate on held-out real data against a gold-standard method; report how many candidates were generated, how humans selected among them and the acceptance rate.
L4 in execution	Validate on hardware with logged safety constraints, a fallback controller and continuous rate or performance monitoring.
L4 in design, analysis or interpretation	Commission a blind expert audit of a random sample of autonomous decisions; confirm every novelty claim with a method independent of the generating pipeline.
L5	Release full agent logs, prompts and intermediate artefacts; disclose AI authorship; obtain independent replication of the central result before any claim of discovery.

Applied retrospectively, VPR would have changed the A-Lab outcome at low cost. A requirement for blind expert review of a sample of diffraction refinements, mandated at L4 interpretation, would likely have surfaced the disorder problem before publication. The same requirement applied to GNoME would have demanded at least one experimental confirmation per structural family before stability predictions were described as materials.

5.5 Implications for resource-limited research groups

The closure analysis carries a specific message for groups without access to large compute or facilities, which includes most university physics departments in India and other developing research systems. The highest-impact systems in this corpus did not require frontier-scale models. AXOL1TL is a shallow autoencoder [19], gpCAM runs Gaussian processes on modest hardware [17], and the AI Feynman code is public. What these systems needed was a closed task, a domain simulator and a verification route. Groups that build open benchmarks with machine-checkable answers for their subfield, or that attach Bayesian steering to existing laboratory instruments, can obtain gains comparable to those reported here without competing on compute. Such work also counteracts topical narrowing, because it makes previously data-poor problems tractable.

6. LIMITATIONS

The study has five limitations. First, the sample of sixteen cases is purposive and favours high-visibility work, which may overstate both success and controversy. Second, a single author coded all cases; inter-rater reliability could not be estimated, and recoding by independent physicists is invited using the rationale in Table 3. Third, ordinal levels compress qualitatively different evidence, and the arithmetic difference G treats the two scales as commensurate, which is a modelling assumption. Fourth, the conservative evidence boundary, which excludes later community replication from V5 and caps disputed claims at V1, lowers verification scores for mature methods such as neural-network quantum states and couples part of the verification score to the existence of critiques. Fifth, the AI Scientist-v2 reference case lies outside physics, and the main correlations are therefore reported with and without it. None of these limitations alters the qualitative contrast between closed and open tasks, which rests on figures reported by the original authors.

7. CONCLUSION

AI and automation have changed physics research unevenly. Autonomy has reached L4 in the execution stage across plasma control, beamline steering and real-time triggering, and these deployments are among the best verified in the corpus because hardware itself tests them. In analysis, AI acts mostly as a fast and accurate assistant whose outputs physicists accept in bulk. In hypothesis selection,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

interpretation and reporting, autonomy is either absent from physics or concentrated in systems whose central claims are disputed or unconfirmed.

The two scales introduced here make this pattern measurable. The principal finding is that autonomy and verification have advanced independently, and that every system whose autonomy runs two or more levels ahead of its verification operates in hypothesis, design or interpretation, where no automatic check exists. The CritPt result and the closure analysis show that the limit on open research is not raw computation but the absence of automatic verifiers. The bibliometric evidence shows that the same dynamics push the field toward a narrower set of data-rich problems.

Three actions follow. Journals should require evidence in proportion to the autonomy a system exercises, using a checklist such as VPR. Funding bodies should treat verifiers, benchmarks and differentiable simulators as research infrastructure of the same standing as models. Physics education should add verification of autonomous systems to its core methods training. Future work should test the PRAL and VL scales with multiple independent coders and a larger random sample, and should track whether the verification gap narrows as autonomous research agents enter physics.

Declarations

Conflict of interest: The author declares no conflict of interest.

Funding: This research received no external funding.

Data availability: All case codes and rationale are reported in Table 3. Figures were generated from values reported in the cited sources.

Use of AI tools: The author used AI-based tools to support literature retrieval and language editing. The author verified every cited source, performed the case coding and takes full responsibility for the content.

REFERENCES

[1] The Royal Swedish Academy of Sciences, "The Nobel Prize in Physics 2024: Press release," Oct. 8, 2024. [Online]. Available: https://www.nobelprize.org/prizes/physics/2024/press-release/

[2] J. Seo, S. Kim, A. Jalalvand, R. Conlin, A. Rothstein, J. Abbate, K. Erickson, J. Wai, R. Shousha and E. Kolemen, "Avoiding fusion plasma tearing instability with deep reinforcement learning," Nature, vol. 626, pp. 746-751, 2024, doi: 10.1038/s41586-024-07024-9.

[3] N. Zipper (for the CMS Collaboration), "Real-time anomaly detection in the CMS experiment," presented at the Fast Machine Learning for Science Conference, Purdue University, Oct. 2024. [Online]. Available: https://indico.cern.ch/event/1387540/contributions/6153423

[4] J. Bausch, A. W. Senior, F. J. H. Heras et al., "Learning high-accuracy error decoding for quantum processors," Nature, vol. 635, pp. 834-840, 2024, doi: 10.1038/s41586-024-08148-8.

[5] N. J. Szymanski, B. Rendy, Y. Fei et al., "An autonomous laboratory for the accelerated synthesis of novel materials," Nature, vol. 624, pp. 86-91, 2023, doi: 10.1038/s41586-023-06734-w.

[6] J. Leeman, Y. Liu, J. Stiles, S. B. Lee, P. Bhatt, L. M. Schoop and R. G. Palgrave, "Challenges in high-throughput inorganic materials prediction and autonomous synthesis," PRX Energy, vol. 3, 011002, 2024, doi: 10.1103/PRXEnergy.3.011002.

[7] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[8] Q. Hao, F. Xu, Y. Li and J. Evans, "Artificial intelligence tools expand scientists' impact but contract science's focus," Nature, 2026, doi: 10.1038/s41586-025-09922-y.

[9] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto and L. Zdeborova, "Machine learning and the physical sciences," Rev. Mod. Phys., vol. 91, 045002, 2019, doi: 10.1103/RevModPhys.91.045002.

[10] M. Krenn, M. Erhard and A. Zeilinger, "Computer-inspired quantum experiments," Nat. Rev. Phys., vol. 2, pp. 649-661, 2020, doi: 10.1038/s42254-020-0230-4.

[11] G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks," Science, vol. 355, no. 6325, pp. 602-606, 2017, doi: 10.1126/science.aag2302.

[12] S.-M. Udrescu and M. Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," Sci. Adv., vol. 6, eaay2631, 2020, doi: 10.1126/sciadv.aay2631.

[13] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno and B. Scholkopf, "Real-time gravitational wave science with neural posterior estimation," Phys. Rev. Lett., vol. 127, 241103, 2021, doi: 10.1103/PhysRevLett.127.241103.

[14] J. Degrave, F. Felici, J. Buchli et al., "Magnetic control of tokamak plasmas through deep reinforcement learning," Nature, vol. 602, pp. 414-419, 2022, doi: 10.1038/s41586-021-04301-9.

[15] Advanced Light Source, Lawrence Berkeley National Laboratory, "Autonomous data acquisition for scientific discovery," Science highlight, 2021. [Online]. Available: https://als.lbl.gov/autonomous-data-acquisition-for-scientific-discovery/

[16] Berkeley Synchrotron Infrared Structural Biology Program, "AI accelerates autonomous discovery at the infrared beamline," Lawrence Berkeley National Laboratory, 2021. [Online]. Available: https://bsisb.lbl.gov/ai-accelerates-autonomous-discovery-at-the-infrared-beamline

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

[17] M. M. Noack, P. H. Zwart, D. M. Ushizima et al., "Gaussian processes for autonomous data acquisition at large-scale synchrotron and neutron facilities," Nat. Rev. Phys., vol. 3, pp. 685-697, 2021, doi: 10.1038/s42254-021-00345-y.
[18] J. Duarte, S. Han, P. Harris et al., "Fast inference of deep neural networks in FPGAs for particle physics," J. Instrum., vol. 13, P07027, 2018, doi: 10.1088/1748-0221/13/07/P07027.
[19] Fast Machine Learning Foundation, "Anomaly detection for new physics at CMS Level 1 Trigger (AXOL1TL)," 2025. [Online]. Available: https://fastmachinelearning.org/applications/anomaly-detection-for-new-physics-at-cms-level-1-trigger/
[20] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, "Scaling deep learning for materials discovery," Nature, vol. 624, pp. 80-85, 2023, doi: 10.1038/s41586-023-06735-9.
[21] A. K. Cheetham and R. Seshadri, "Artificial intelligence driving materials discovery? Perspective on the article: Scaling deep learning for materials discovery," Chem. Mater., vol. 36, no. 8, pp. 3490-3495, 2024, doi: 10.1021/acs.chemmater.4c00643.
[22] M. Krenn, Y. Drori and R. X. Adhikari, "Digital discovery of interferometric gravitational wave detectors," Phys. Rev. X, vol. 15, 2025.
[23] Max Planck Institute for the Science of Light, "AI finds new ways to observe the most extreme events in the universe," Press release, Apr. 15, 2025. [Online]. Available: https://mpl.mpg.de/news/article/ai-finds-new-ways-to-observe-the-most-extreme-events-in-the-universe
[24] Y. Yamada, R. T. Lange, C. Lu, S. Hu, C. Lu, J. Foerster, J. Clune and D. Ha, "The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search," arXiv:2504.08066, 2025.
[25] Sakana AI, University of British Columbia, Vector Institute and University of Oxford (The AI Scientist team), report on The AI Scientist, Nature, published online Mar. 25, 2026, doi: 10.1038/s41586-026-10265-5.
[26] A. Novikov, N. Vu, M. Eisenberger et al., "AlphaEvolve: A coding agent for scientific and algorithmic discovery," arXiv:2506.13131, 2025.
[27] M. Zhu, M. Tian, X. Yang et al., "Probing the critical point (CritPt) of AI reasoning: A frontier physics research benchmark," arXiv:2509.26574, 2025.
[28] E. Duede, W. Dolan, A. Bauer, I. Foster and K. Lakhani, "Oil and water? Diffusion of AI within and across scientific fields," arXiv:2405.15828, 2024.
[29] SAE International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," Standard J3016, 2021.
[30] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz and A. Zeilinger, "Automated search for new quantum experiments," Phys. Rev. Lett., vol. 116, 090405, 2016, doi: 10.1103/PhysRevLett.116.090405.
[31] C. J. Shallue and A. Vanderburg, "Identifying exoplanets with deep learning: A five-planet resonant chain around Kepler-80 and an eighth planet around Kepler-90," Astron. J., vol. 155, 94, 2018, doi: 10.3847/1538-3881/aa9e09.
[32] NASA, "Discovery of eight planets makes alien system the first to tie with our solar system," Dec. 14, 2017. [Online]. Available: https://science.nasa.gov/universe/exoplanets/discovery-of-eight-planets-makes-alien-system-the-first-to-tie-with-our-solar-system
[33] M. F. Kasim, D. Watson-Parris, L. Deaconu et al., "Building high accuracy emulators for scientific simulations with deep neural architecture search," Mach. Learn.: Sci. Technol., vol. 3, 015013, 2022, arXiv:2001.08055.
[34] R. Lam, A. Sanchez-Gonzalez, M. Willson et al., "Learning skillful medium-range global weather forecasting," Science, vol. 382, pp. 1416-1421, 2023, doi: 10.1126/science.adi2336.
[35] Princeton University School of Engineering and Applied Science, "Engineers use AI to wrangle fusion power for the grid," ScienceDaily, Feb. 23, 2024. [Online]. Available: https://www.sciencedaily.com/releases/2024/02/240221160337.htm
[36] Freethink, "Google AI discovers 2.2 million new materials for a host of technologies," Dec. 2023. [Online]. Available: https://www.freethink.com/robots-ai/google-ai-discovers-2-2-million-new-materials
[37] Google DeepMind, "GraphCast: AI model for faster and more accurate global weather forecasting," Nov. 14, 2023. [Online]. Available: https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/
[38] A. Woodie, "AI for science study: Good for the goose, but what about the gander?" HPCwire, Jan. 2026. [Online]. Available: https://www.hpcwire.com/2026/01/16/ai-for-science-study-good-for-the-goose-but-what-about-the-gander/
[39] Google DeepMind and Google Quantum AI, "AlphaQubit tackles one of quantum computing's biggest challenges," Nov. 20, 2024. [Online]. Available: https://blog.google/technology/google-deepmind/alphaqubit-quantum-error-correction/

(This work is licensed under a Creative Commons Attribution 4.0 International License.)