How to Critically Analyze Literature on Autonomous Experimentation

A. Gilad Kusne [1,2] (ORCID: 0000-0001-8904-2087)
Austin McDannald [2] (ORCID: 0000-0002-3767-926X)
Ichiro Takeuchi [3] (ORCID: 0000-0003-2625-0553)

1 Electrical and Computer Engineering Department, University of Maryland, College Park MD 20742, USA

2 Material Data Division, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg MD 20899, USA

3 Materials Science and Engineering Department, University of Maryland, College Park MD 20742, USA

Abstract: Autonomous experimentation systems, also known as self-driving labs, help researchers fail smarter, learn faster, and spend less. These efficiencies arise from coupling artificial intelligence with automated lab tools. Of critical importance is how to define success. Here, we discuss some questions to better understand a work's novelty, utility, and implementation.

Introduction:

Research and development in autonomous experimentation systems (AES, a.k.a. self-driving labs, also sometimes referred to as SDLs) are rapidly growing in popularity. These systems combine novel artificial intelligence (AI) with automated research tools (both physical and *in silico*) allowing the AI to perform experiment design, execution, and analysis in a closed loop[1]. While AES may apply to computational or physical studies, we confine our discussion to experimentation involving physical hardware, though the concepts carry over to computational experimentation with minimal modification. On every iteration, the AI identifies the experiments most likely to accelerate the user toward their goals, e.g., exploring an underlying mechanism of a physical phenomenon or discovering an optimal material. It then performs these experiments through tool automation and then collects and analyzes resulting data to inform the next iteration. This is in stark difference with systems with pre-programmed tasks, which are automated and not autonomous.

Across domains ranging from materials exploration and process optimization to fusion reactor control and drone flight path control, the term autonomous is applied to systems that rapidly navigate a multi-dimensional parameter space and make decisions on-the-fly, a task that is particularly difficult and time-consuming for humans. Our focus here is on AES for materials exploration and optimization, where the input parameter space is typically high-dimensional. Terminology in this area is not settled: autonomous research, autonomous experimentation, autonomous discovery, and self-driving laboratory all appear in the literature, with usages differing in how much weight they place on physical automation versus algorithmic selection of what to do next. For this paper we use the following definition: an AES couples AI-driven decision making that selects the next experiment (conditions, tools, tool parameters, etc.) with automated execution of those decisions. Figure 1 makes the two underlying axes

explicit, so a given work can be located regardless of the label it adopts.

For a typical AES, AI completes three tasks on each iteration before executing subsequent experiments: 1) *Analysis of past data*: Analyze all previously collected data, whether from physical experiments, simulations, the human user, external databases, literature, etc. This includes any necessary processing required to convert raw data into actionable data. 2) *Predictive modeling*: Use analysis from (1) to build a model or hypothesis (or set of models) capable of predicting the results of future experiments. Predictive models and hypotheses are paired with quantification of model uncertainty. Optimally, prediction uncertainty also includes propagated uncertainty from input data, 3) *Experiment design*: determine the next set of experiments to perform that will accelerate the research campaign toward the user's goals. On each iteration, an AES uses feedback signals to enable learning and on-the-fly decision making. This feedback includes capturing and analyzing experiment metadata, which AES can leverage to boost experiment reproducibility.

For the AI to execute an experiment and collect the resulting data requires complex interactions with automated tools. This includes communication between AI and the tools, setting each tool's control parameters, moving physical samples into/out of the tool, moving data over the network, managing tool-dependent data artifacts, interacting with computational tools located on high performance computing resources, etc.

As a result of these complex interactions, AES exists at the interface of many fields and requires expertise in artificial intelligence as well as computer science, physical sciences, data science, instrument sciences, and control systems among others. This complicates reviewing AES manuscripts and creates opportunities for readers and reviewer to misunderstand the work's novelty, implementation, and successes.

One solution would be an AES-focused venue, journal or conference, that is not merely interdisciplinary in scope but jointly run by a computer science and a physical science community, with editorial boards and reviewer pools drawn from both, so that each manuscript is vetted by reviewers from each side. Current popular venues are products of only one community, either of physical scientists or computer scientists, and hope to draw authors and reviewers of the other background. More immediately, the challenge of misunderstanding AES manuscripts can be mitigated by ensuring that key pieces of information are included in the main manuscript. We aim to provide critical questions that a reader or reviewer can ask to ensure the inclusion of these key pieces of information and to boost understanding of the AES manuscript and its associated work.

It is important to note that excluding such information, whether intentionally or not, can mistakenly increase an AES work's perceived novelty and its perceived importance to a journal or conference's audience. We hope that the provided questions serve to reinforce the foundations of AES research and foster greater transparency. We recognize that applying these questions places additional demands on reviewers, and that some papers not meeting every criterion here will still be published. Our goal is not to gatekeep but to equip both reviewers and readers with a shared critical lens, so that the field matures with its claims and its standards of evidence evolving together.

The Research Spectrum

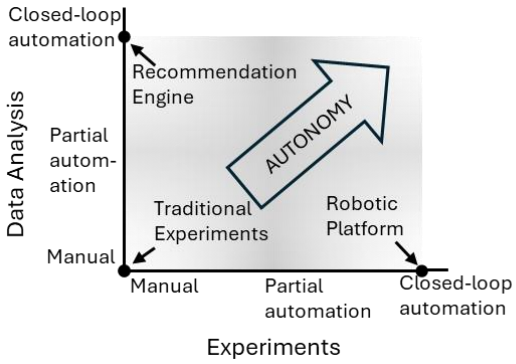


Figure 1. The Research Methodology Chart

Landscape of autonomous experimentation as a function of the levels of automation in physical experiments and data analysis. Traditional experimentation is shown in the bottom left, with humans driving experiment design, execution, and analysis. Increasing automation of the research's physical activities develops a robotic platform with reduced human physical involvement and the potential for greater reproducibility and execution provenance. As we move right in the plot, humans retain physical tasks that are less and less impactful on the research, e.g., moving a sample from one tool to another. As an alternative to automating physical tasks, increasing automation of the data analysis (including decision making) drives up in the plot, toward a recommendation engine that may offer improved analysis reproducibility and decision provenance. Combining automation of data analysis and physical tasks increases the ability for the system to gather actionable information more quickly and increases reproducibility and experiment provenance. Automation of data analysis where all the necessary data is previously collected (not collected by automated tools) falls under simulation of AES or post-hoc AES.

Figure 1 maps the relationship between AES and other related research approaches. Experiment operation can vary from being fully manual to full closed loop automation. The same is true of data analysis, which can vary from pen and paper data analysis to AI and other algorithmic solutions driving closed-loop data analysis automation. The map indicates the relationship between traditional experiments where humans perform all the research tasks and experiments with different types of automation. Increasing automation of physical tasks leads to robotic experiment platforms with the potential for greater reproducibility and execution provenance. Increasing automation of data analysis (including decision making) leads to algorithmic recommendation engines with the potential for greater analysis reproducibility and decision provenance. Here, humans perform the experiments and feed the results to the AI, which analyzes the results and determines the next set of experiments to perform. Increasing automation of both data analysis and physical tasks increases autonomy with the potential of accelerating knowledge capture, improving reproducibility, and improving provenance.

A 'gray area' exists between manual and fully closed-loop automation along each axis, where human assistance is required for the success of the research campaign. For experiment execution, an example is a human moving samples from one tool to another due to a lack of appropriate sample handling robotics. Along the data analysis axis, a human may be needed to perform complex data analysis tasks that have yet to be automated. Along the direction from fully traditional experiments to fully autonomous, the gray area includes a human monitoring and modifying the AI driving a research campaign, with modifications serving to pull the platform out of suboptimal behavior.

Human involvement is divided into two types. In the first type, the human supplies physical task(s), executing a step that the AI has specified. This may persist because automating a simple transfer or loading step

can cost more than it returns. In the second, the human contributes analysis or decision making, such as interpreting complex data where software analysis automation is not yet possible, or deciding on which experiment to run next. Both may be essential to completing the campaign, but the second changes who is steering it, and therefore what the work can or should claim.

One can build AES that push beyond the bounds of Figure 1 to achieve even greater research efficacy. For such systems, humans perform tasks that serve to accelerate the performance of the AES. For instance, a human expert could impart their prior expertise to the AI to boost AI prediction accuracy in the presence of few experiment observations[2]. Or AES can collaborate to tackle challenges of greater complexity than any one AES can address on their own[3]. Such interactions fall under the category of AI-human teaming or collaboration[4].

Questions

The questions below are summarized in **Table 1** and discussed in turn in the remainder of this section.

Utility: Should this be autonomous? I.e., what is the utility?

Here we focus on the question of whether adding autonomous control to a particular research system provides practical benefits. The question of utility is relevant to a reviewer in gauging audience interest for a journal or conference.

AES development is motivated by a need to accelerate research campaigns and limit resource use. For example, a user obtains a limited window of instrument access at a user facility. The user then wishes to exact the maximum information they can during their limited access time. AES allows a user to achieve their goals with fewer experiments, in less time, and with fewer resources, allowing the user to pursue larger research goals during the same access window. Alternatively, the user can complete their task more quickly, freeing up the instrument for other users.

However, there are cases where the benefits of autonomous control may be more difficult to demonstrate. For instance, consider the case of a scanning measurement tool with optimal hardware and firmware capable of rapid raster scanning a sample. One can integrate AI control to convert the system into an AES, with the AES optimally navigating across the sample, seeking to estimate full-sample measurement response with the minimal number of measured sample points. However, for a scanning measurement system with fast enough measurements and scanning rates, simple rastering may provide similar information in similar or less time, and the added complexity of AI control must be justified. In such cases, authors should clearly articulate the specific regime in which AI control provides benefit. Examples include beam- or dose-sensitive samples where exhaustive rastering causes damage, sparse region-of-interest search across large sample areas, drift or contamination correction during long scans, or measurement campaigns where each point carries a high time or resource cost. The question is not whether AI control is warranted in general for scanning tools, but whether the authors have identified and demonstrated the regime in which it is warranted for their specific problem.

Another example is investigating the impact of only one control variable on target material properties. Unless the experiments or analysis are overly onerous, costly, resource intensive, or time intensive, AES may not be necessary, one can simply run a grid of experiments for the one control variable.

Cost, resource use, and the time for AI computations, measurements, scanning, etc. must be clearly stated for readers to gauge the added value of autonomous operation. Furthermore, AES-based tool operation should be compared to common operation, i.e., how would a human expert typically use the same tool? Does AES operation improve on such manual operation?

A key deciding factor in the preferability of AES is the compute time required by the AI driving the AES. If compute time is significantly longer than experiment time, (e.g., compute time is 10 minutes while serial experiments take 1 minute) AES may not out-perform simpler experimentation. Here, a fair comparison would also include the post-measurement analysis in a conventional workflow. A common solution in such cases is to utilize AI-based forecasting to estimate the results of future experiments and make predictions and decisions based on those estimates. However, these decisions tend to be suboptimal due to the lack of real data. Generally speaking, the larger the compute time (relative to the experiment time), the further out forecasting must be performed, and the less optimal the decisions made. Furthermore, a key quantity to report is the total time to complete an iteration loop of decision, execution, and analysis. The iteration time aids in understanding the AES acceleration of knowledge capture.

Additionally, achieving autonomous operation may require tool modification and increasing the experiment's complexities, again reducing the value of AES operation. For instance, a typical raster scanning tool may need significant modifications and significant increase in complexity to be able to hop from one specified point to another. Any AES-derived benefits should be measured against any increase in system complexity.

Furthermore, experiment design has a long history prior to AES, including classical methods for design of experiments[5]. Before performing any experiments, these methodologies plan out the full set of experiments to perform that will achieve maximal information. On the other hand, AES learn from each new piece of information and optimize the selection of each subsequent experiment on-the-fly[6]. Common methods fall under active learning[7] and Bayesian optimization[8]. One must ask if this on-the-fly optimization provides benefits over traditional experiment design methods.

Novelty: Does the background exclude relevant work essential in judging novelty?

Due to the interdisciplinary nature of the AES field, journals find it hard to properly review papers – e.g., a materials journal must find a reviewer knowledgeable of pertinent AI, and an AI conference must find a reviewer knowledgeable of the pertinent materials science. Without appropriate vetting, papers with incomplete background sections are more likely to be published, boosting the perceived novelty of the work.

For example, an AES paper may provide background for two aspects of the work – the physical science performed, and the AI utilized. The paper may however exclude discussion of previous work applying similar AI to similar physical science research challenges. This lack of essential background information may boost the perceived novelty of the work, whether intentionally or not. This oversight may be due to the lack of awareness of relevant literature. The interdisciplinary field of AES must deal with many conflicting jargons of contributing fields that can confound proper background investigations.

We have previously mentioned that collaboration between physical science and

AI communities in manuscript review can work to ensure that appropriate background information is included. Another solution is to ensure that reviewers and readers are better informed, which is challenging in a new and fast-moving field.

Implementation: Was the autonomous research campaign post-hoc?

An AES paper may claim success in accelerating or improving the efficiency of a research campaign, e.g., reducing the number of real-world physical experiments, amount of resources used, time taken, etc. However, in making these claims, the authors must clearly state in the main manuscript whether the research campaign was performed in simulation or if the research campaign was performed on an instrument. Without this information, a reader or reviewer may incorrectly understand the reported acceleration or efficiencies as describing real world AES operation.

Simulated AES include performing the closed-loop research campaign using simulations of physical equipment or through post-hoc AI simulations. Here post-hoc AI simulations describes the process of first performing all relevant real-world experiments, building a database, and then having the AI sample that database rather than performing real-world experiments. Post-hoc autonomous experiments have significant value. They provide initial estimates for real-world AES performance and a platform to work out potential AES workflow and algorithm issues. Simulation-only AES can also provide a digital twin for benchmarking the AI driving AES. However, papers should clearly indicate that they are post-hoc or that they use surrogate simulated data. Critically, a simulated AES is not a true AES, and a manuscript claiming successful AES operation must demonstrate real-time operation. The same holds for closed-loop computational AES, which must contend with its own real-time scheduling and resource-dependent variability.

Why is real world execution important? Real-world AES execution requires managing numerous issues associated with running live experiments, issues that dramatically complicate the research campaign. Issues include managing complex instrument communication protocols and instrument scheduling and navigating network firewalls. One must also mitigate instrument-dependent data artifacts, instrument idiosyncrasies (e.g., “ringing” in setting a temperature), environment dependences (temperature, humidity, vibration, source voltage and current fluctuations, etc.), and sample idiosyncrasies (e.g., materials grain size and shape distribution, defects, etc.). For an AI to truly manage the research campaign, it must be able to tackle these real-world challenges on-the-fly. Furthermore, *only in real world execution of experiments can new data be acquired*.

Additionally, post-hoc simulations of AES may point to potential concerns. When testing or developing an AES’s AI on previously acquired data, that data must either be exhaustive (with resolution at or finer than the live experiment will achieve) or it will need some sort of interpolation. If the previously acquired data is exhaustive, then utility of the AES must be clearly stated - if you can acquire exhaustive data, why do you need an AES? It could be that the AES could allow for expanded range, higher dimensions, or be much faster. Each goal will have its own success measure which should illuminate the utility of evaluating the performance of the AES on such an exhaustive data set.

If, however, the AES AI research campaign is evaluated by interpolating between previously acquired data points this makes strong

assumptions about unmeasured locations. There may be new and interesting behavior at these unmeasured locations, violating the assumptions of the interpolation, and preventing the opportunity for discoveries. As a result, quantifications of AES utility and performance may be incorrect.

If AES was simulated, the authors should include information about how the simulation was performed, if simulated physical tools were used, if the data was exhaustively collected prior to the simulation, and whether interpolated data was used.

Implementation: What is the level of autonomy in data analysis, decision making and hardware?

Does the AES require human intervention to succeed? How, where, and why? To answer these questions, one must often navigate the supplemental material. For example, a paper may report AI-driven convergence to an optimal result, while the supplemental shows that regular human intervention was needed to modify the experiments, tune the AI parameters, goals, or convergence criteria, etc. The role of the human may not be stated at all, making the degree of autonomous control uncertain. Again, to mitigate the potential for misunderstanding, the authors should clearly state the role of humans throughout the autonomous methodology, and this information should be placed in the main manuscript. If this information is lacking, reviewers should request it.

Utility: Are the discoveries practically useful? Are there industrial applications?

AI-driven materials discovery has resulted in reported discoveries of many materials. However, many of these reports do not demonstrate that the novel materials out-perform the current relevant best-in-class materials. Instead, discovered materials are often simple stochiometric substitutions of known compounds and are not superior in cost, function, or mechanism, or present a novel phenomenon. This is similar to the large number of materials discovered without the use of AI. In such instances, the discovered materials are not likely to make it to industrial use.

For AES, reporting that a material was discovered can have a larger impact than the discovered material itself. Readers may look to the AES's AI, robotic platform, or methodology for a means of accelerating discovery of novel and industry-relevant materials as well as novel and interesting underlying mechanisms. This potential AES impact and utility grows with the discovery of a best-in-class material, a material exhibiting novel phenomenon, or a functional material with novel underlying physical mechanism.

As a result, any materials discovered should be compared to relevant best-in-class materials. If one suspects that they have discovered a novel, superior material, that material must be fully vetted. X-ray diffraction is typically not enough validation. Similarly, materials 'discovered' in simulation or computations must be validated with physical experiments, and appropriate experimental data must be provided for critical evaluation by reviewers.

Additionally, for AES seeking to discover novel physical mechanisms, many AI-driven 'discoveries' of physical mechanisms are demonstrations of AI discovering well-known mechanism or heuristic. As a result, practical utility remains to be proven. Any demonstration of AI-discovered physical mechanisms should be framed in terms of known physics.

Implementation: Did the authors use the appropriate AI decision making

methodology for their work? I.e., should this be active learning, Bayesian optimization, reinforcement learning, combinations thereof, or something else?

Different AI decision making strategies have different strengths and weaknesses and, as a result, differ in the classes of challenges to which they are typically applied. I.e., the choice of appropriate decision-making strategy for experiment design is determined by the research challenge. Common challenges are to explore an unknown materials behavior or searching for an optimal material. Toward the former goal, active learning seeks to learn and represent the unknown behavior with a high-fidelity surrogate function. Toward the latter goal, Bayesian optimization seeks to build a similar surrogate function with the goal of identifying a function extremum. Active learning and Bayesian optimization assume that one seeks to explore a new material space and learn a new behavior, where the action of performing an experiment does not impact the material behavior.

Reinforcement learning frames the challenge differently. Actions change the state of the system, action consequences may appear only many steps later, and the goal is to maximize reward accumulated over a whole trajectory. The task is to learn a policy mapping state to action. Active learning and Bayesian optimization can be considered limits of reinforcement learning, where each experiment is independent and no credit assignment over time is needed. Reinforcement learning is more suited to path-dependent systems, where the current action alters the state that later actions act upon, e.g., controlling a deposition or annealing process, or tuning an instrument whose configuration carries over, rather than to mapping a static response surface.

Authors should clearly state their decision making strategy to facilitate critical analysis.

Performance: Are performed AI selected experiments overly similar to past experiments?

AI algorithms predict most reliably when new experiments fall within the distribution of the training data. As a result, measuring the prediction accuracy of an AES algorithm over data highly similar to that used for training will likely suggest greater interpolation or extrapolation capability than is achievable in real operation.

This matters because an AES cannot guarantee that newly acquired data will resemble its training data. In fact, moving beyond that data is often the point of running the campaign, e.g., the search for novel phenomena. Performance measures (such as accuracy) should therefore consider the distance from the training (and pre-training) data. One caveat: input-space similarity alone is not a sufficient measure of a test's value. Measuring at inputs very close or identical to those already acquired is often necessary, since replicates quantify noise in the measured outputs. Such analysis of replicants should be reported as a noise estimate rather than as evidence of predictive capability.

Performance: Has the AES converged to the best result? Was the strategy properly benchmarked?

If an AES is driven by a poorly performing AI or non-informative signals, it may not have enough information to home-in on the user's goals. Instead, the AES may stumble through many experiments and happen upon an optimal experiment. This is a realistic possibility due to the surprisingly good performance of random sampling in exploring an experiment's parameter space[9]. It would

then be misleading to report that the AES found the optimal experiment on its own merits.

In validating AES convergence, reviewers and readers should consider the number and distribution of experiments performed before reaching the optimal experiment. By tracing the experiments performed over time, one can roughly determine if the AES is performing deterministically. This is commonly done by recording the order of experiments performed and then visualizing this path from experiment to experiment (see Figure 2).

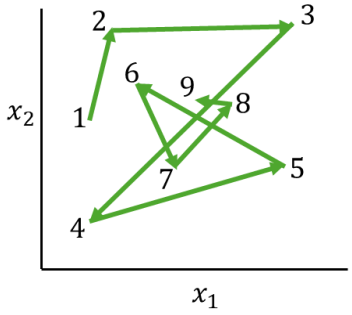


Figure 2. Demonstration of tracing the experiments performed by an AES. The green arrows show the path taken by the AES in the X domain.

Additionally, experiments are innately stochastic processes with randomness introduced at many levels from equipment control parameter value uncertainties to user decisions in initializing the system. As a result, one should report performance distribution of the AES strategy across relevant sources of randomness. This more clearly demonstrates the value of the AES strategy. For example, if one initiates the AES with a random sample, evaluate the distribution of AES performance as a function of that random sampling. If the feedback signal is highly noisy, evaluate the distribution of AES performance as a function of the real-world variations in the feedback signal.

Similarly, if one is introducing a novel AES AI, it should be benchmarked against relevant algorithms including those off-the-shelf. For example, if a paper claims to demonstrate a novel experiment design strategy (i.e., active learning or Bayesian optimization acquisition function), it should be benchmark against relevant acquisition functions including random selection. This will then seed the conversation about the value proposition the new AI provides, when is there a performance boost and how that boost compares to the benchmark.

More broadly, an experimentalist may have only a single measurement tool available and may not yet know whether that tool provides the critical information needed to drive autonomous optimization. This should not discourage them from setting up an AES. Rather, one explicit goal of such a study should be to determine whether the available measurement signal is informative enough to drive the AI, and comparison against random sampling is the cleanest way to reveal this. If the AES does not outperform random sampling, that is itself a useful finding: it indicates the signal may be insufficient, and the community learns which tools are and are not adequate drivers of autonomous optimization for that class of problem.

Implementation: Is the AES measuring the right thing for the stated goal?

As mentioned above, an AES measuring a non-informative signal may still happen upon an optimal solution due to the surprisingly good performance of random sampling. One must critically evaluate the path the AI took toward the solution to identify the role of the AI's logic.

Similarly, an AES may measure a surrogate signal instead of directly measuring the most indicative signals. As a result, any output does

not directly translate to a discovery. Instead, outputs point to further, more thorough materials investigations.

Implementation: What were the stopping criteria?

Due to the exploratory nature of research, an AES can never know if it truly found an optimal result. As a result, one must look at the AI's stopping criteria. This could depend on the number of experiments performed, materials resources used, or instrument time etc. It could also depend on measures of experiment optimality or convergence criteria. A reviewer should ask themselves if the stopping criteria is relevant to the challenge and indicative of a successful AES. For example, an AES seeking an optimal material would be expected to converge to that optimal. If the AES is stopped before convergence, it may indicate that the optimal result found may not have been deterministically found or that the indicated optimum may not be the algorithm's preferred optimum. Similarly, for an active learning driven AES that seeks to explore a new space, authors can plot the exploration signal (aka acquisition function) to demonstrate that utility of exploration had dropped beneath some threshold.

Further artificial intelligence-based questions

Due to AES's dependence on AI, many questions relevant to critical evaluation of AI methods also apply to evaluating AES value and performance. For example, one can report an inappropriate measure of performance accuracy, such as applying a binary classification accuracy measure to a multi-class challenge. Especially relevant pitfall for AES is to use accuracy as a performance measure when the goal of the AES is to discover the classes present. Consider a "needle-in-a-haystack" type problem: the goal is to discover the small number of samples or region of interest. A model could have high accuracy by calling all the samples "hay" missing the "needle" entirely. And so accuracy has no bearing on the performance of AES for that problem. As a result, each step in the AES AI pipeline should be evaluated and justified with respect to the clearly stated goal of the AES.

Conclusion

We hope that these critical evaluation questions and discussions aid AES researchers to better evaluate and report their work and aid readers and reviewers in critical analysis of reported work.

Acknowledgements

This work was partly supported by the center for 3D Ferroelectric Microelectronics Manufacturing (3DFeM2), an Energy Frontier Research Center funded by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences under Award Number DE-SC0021118

Declaration of Interests

The authors declare no competing interests.

REFERENCES

1. Stach, E. *et al*. Autonomous experimentation systems for materials development: A community perspective. *Matter* **4**, 2702–2726 (2021).
2. Adams, F., McDannald, A., Takeuchi, I. & Kusne, A. G. Human-in-the-loop for Bayesian autonomous materials phase mapping. *Matter* **7**, 697–709 (2024).

3. Kusne, A. G. & McDannald, A. Scalable multi-agent lab framework for lab optimization. *Matter* (2023).
4. Kusne, A. & Mcdannald, A. Agent, Agentic, and Distributed Artificial Intelligence: From Managing Next-Generation Labs to the Philosophy of Science. https://www.authorea.com/doi/full/10.22541/au.175915787.70926747?commit=c273503c6653c5235c1deb4d31d85eb40b45d596.
5. NIST/SEMATECH e-Handbook of Statistical Methods. https://www.itl.nist.gov/div898/handbook/.
6. Braham, E. J., Davidson, R. D., Al-Hashimi, M., Arróyave, R. & Banerjee, S. Navigating the design space of inorganic materials synthesis using statistical methods and machine learning. *Dalton Trans.* **49**, 11480–11488 (2020).
7. Settles, B. Active learning literature survey. *Univ. Wis. Madison* **52**, 11 (2010).
8. Garnett, R. *Bayesian Optimization*. (Cambridge University Press, 2023).
9. Rohr, B. *et al.* Benchmarking the acceleration of materials discovery by sequential learning. *Chem. Sci.* **11**, 2696–2706 (2020).

Type	Question	Intent / Summary
Novelty	Does the background exclude relevant work essential to judging novelty?	Check that prior work applying similar AI to similar science is cited; interdisciplinary gaps can inflate perceived novelty.
Utility	Should this be autonomous?	Assess whether autonomous control yields practical benefit over manual operation and classical design of experiments, accounting for a fair comparison of complexity and compute cost.
	Are the discoveries practically useful? Are there industrial applications?	Compare discovered materials or mechanisms to best-in-class; X-ray diffraction alone is insufficient validation of a discovery.
Implementation	Was the autonomous research campaign post-hoc?	Distinguish real-time AES from simulated or post-hoc database sampling; claims of acceleration require real-world execution evidence.
	What is the level of autonomy in data analysis, decision making and hardware?	Require the role of human intervention to be stated in the main text, not buried in supplements.
	Was the appropriate AI decision-making method used (active learning, Bayesian optimization, reinforcement learning, etc.)?	The method should match the research goal (exploration vs. optimization vs. policy learning) and be stated explicitly.
	Is the AES measuring the right thing for the stated goal?	A surrogate or non-informative signal can still stumble onto an optimum; outputs only point to the need for further investigation.
	What were the stopping criteria?	Stopping criteria should be relevant to the goal; premature stopping can mask non-deterministic convergence.
	Do general AI evaluation pitfalls apply?	Standard AI evaluation concerns (e.g., inappropriate accuracy measures for needle-in-a-haystack goals) apply to each step of the AES pipeline.
Performance	Are AI-selected experiments overly similar to past experiments?	Guard against data leakage; pair prediction accuracy with train/test similarity analysis.
	Has the AES converged to the best result? Was the strategy properly benchmarked?	Benchmark against random sampling and off-the-shelf methods; report performance distributions over sources of randomness.

Table 1. Summary of critical questions for evaluating autonomous experimentation system (AES) reports.