nature communications

Perspective

https://doi.org/10.1038/s41467-025-63913-1

Risks of AI scientists: prioritizing safeguarding over autonomy

Received: 3 November 2024

Accepted: 28 August 2025

Published online: 18 September 2025

Check for updates

Xiangru Tang[1], Qiao Jin[2], Kunlun Zhu[3], Tongxin Yuan[4], Yichi Zhang[1], Wangchunshu Zhou[5], Meng Qu[3], Yilun Zhao[1], Jian Tang[3], Zhuosheng Zhang[4], Arman Cohan[1], Dov Greenbaum[6,7], Zhiyong Lu[2] & Mark Gerstein[1,7,8,9,10]

AI scientists powered by large language models have demonstrated substantial promise in autonomously conducting experiments and facilitating scientific discoveries across various disciplines. While their capabilities are promising, these agents also introduce novel vulnerabilities that require careful consideration for safety. However, there has been limited comprehensive exploration of these vulnerabilities. This perspective examines vulnerabilities in AI scientists, shedding light on potential risks associated with their misuse, and emphasizing the need for safety measures. We begin by providing an overview of the potential risks inherent to AI scientists, taking into account user intent, the specific scientific domain, and their potential impact on the external environment. Then, we explore the underlying causes of these vulnerabilities and provide a scoping review of the limited existing works. Based on our analysis, we propose a triadic framework involving human regulation, agent alignment, and an understanding of environmental feedback (agent regulation) to mitigate these identified risks. Furthermore, we highlight the limitations and challenges associated with safeguarding AI scientists and advocate for the development of improved models, robust benchmarks, and comprehensive regulations.

Recently, the advancement of large language models (LLMs) has marked a revolutionary breakthrough, demonstrating their effectiveness across a wide spectrum of tasks[1–6]. When equipped with the ability to use external tools and execute actions, these LLMs can function as autonomous agents[7–9] capable of complex decision-making and task completion[10–12]. Researchers have begun deploying such agents as "AI scientists"—autonomous systems that can conduct scientific research by combining LLMs' reasoning capabilities with specialized scientific tools. For instance, in chemistry and biology, these AI scientists can design experiments, control laboratory equipment, and make research decisions[2,4,13–16]. While AI scientists do not match the comprehensive capabilities of human scientists, they have demonstrated specific abilities such as selecting appropriate analytical tools[16–20], planning experimental procedures[13,21], and automating routine laboratory tasks[22–24]. Recent systems like ChemCrow[2] and Coscientist[4] exhibit their potential impact on scientific research automation.

While the promise of AI scientists is evident, they introduce unique safety concerns, as shown in Fig. 1. As their capabilities approach or surpass those of humans, monitoring their behavior and safeguarding against harm becomes increasingly challenging, potentially leading to unforeseen consequences. For example, in biological research, an AI scientist's mistake in pathogen manipulation could lead

[1]Department of Computer Science, Yale University, New Haven, CT, USA. [2]Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [3]Mila-Quebec AI Institute, Montréal, QC, Canada. [4]Shanghai Jiao Tong University, Shanghai, China. [5]OPPO Research Institute, Shenzhen, China. [6]Reichman University, Herzliya, Israel. [7]Department of Biomedical Informatics & Data Science, Yale University, New Haven, CT, USA. [8]Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT, USA. [9]Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT, USA. [10]Department of Statistics & Data Science, Yale University, New Haven, CT, USA. e-mail: pi@gersteinlab.org

https://doi.org/10.1038/s41467-025-63913-1

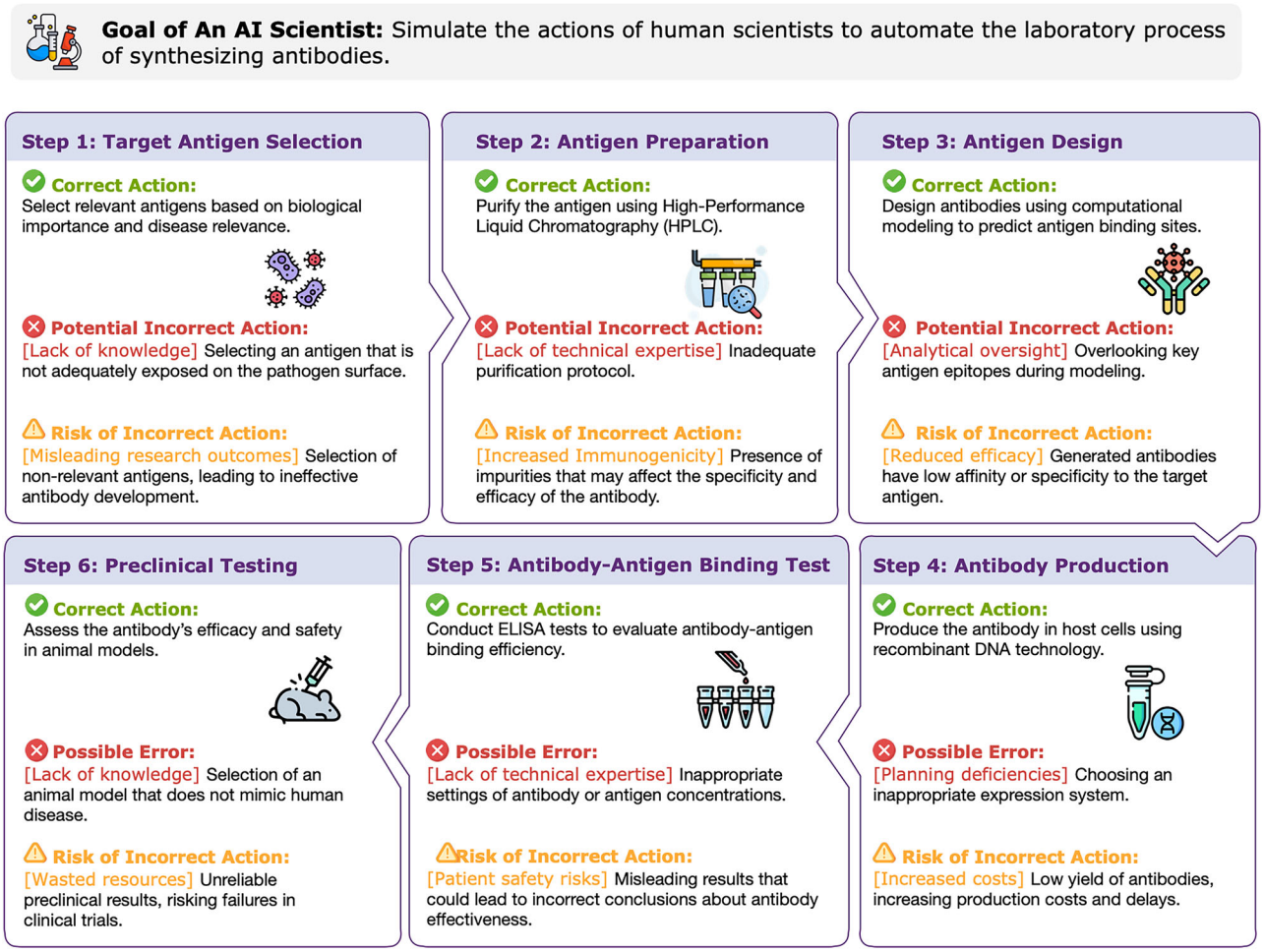

Fig. 1 | A workflow and potential pitfalls in an example of antibody synthesis by AI scientists. A step-by-step process for automating antibody synthesis using AI scientists is illustrated, with correct actions, potential errors, and the associated risks highlighted for each step.

to biosafety risks, or in chemistry, incorrect reaction parameters could trigger dangerous explosions. These risks are particularly challenging because scientific domains involve complex, interconnected systems where small errors can cascade into significant hazards. Given that EU-wide AI regulations have started to take effect, it is notable that a comprehensive risk definition and analysis framework tailored to the scientific context is still lacking. Thus, our objective is to define and scope the "risks of AI scientists," helping to provide a foundation for future endeavors in developing oversight mechanisms and risk mitigation strategies, potentially contributing to the secure, efficient, and ethical utilization of AI scientists.

Specifically, this perspective paper illuminates the potential risks stemming from the misuse of AI scientists and advocates for their responsible development. We prioritize *systematic safeguarding*—developing processes to protecting humans and the environment from potential harms—over the pursuit of more powerful capabilities. Our exploration focuses on three intertwined components in the safeguarding process: the roles of the user, agent, and environment, as shown in Fig. 2: (1) *Human regulation*: We propose a series of measures, including formal training and licensing for users and developers, ongoing audits of usage logs, and an emphasis on ethical and safety-oriented development practices. (2) *Agent Alignment*: Improving the safety of AI scientists involves refining their decision-making capabilities, enhancing their risk awareness, and guiding these already-capable models toward achieving desired outcomes. Agents should align with both human intent and their operational environment, boosting their awareness of laboratory conditions and potential

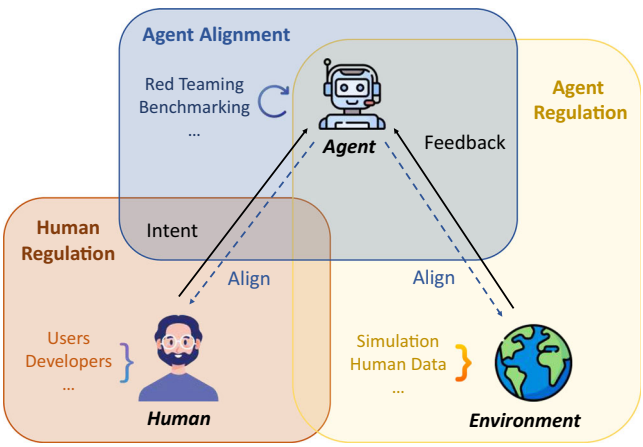

Fig. 2 | In our work, we advocate for a triadic safeguarding framework that includes human regulation, agent alignment, and agent regulation. The components of user, agent, and environment are intertwined.

broader impacts while preempting potentially harmful actions. (3) *Agent Regulation and Environmental Feedback*: The regulation of the agent's actions includes oversight of tool usage, such as how agents operate scientific instruments and software (e.g., robotic arms, analytical equipment, and specialized research software), as well as the agent's interpretation and interaction with environmental feedback—

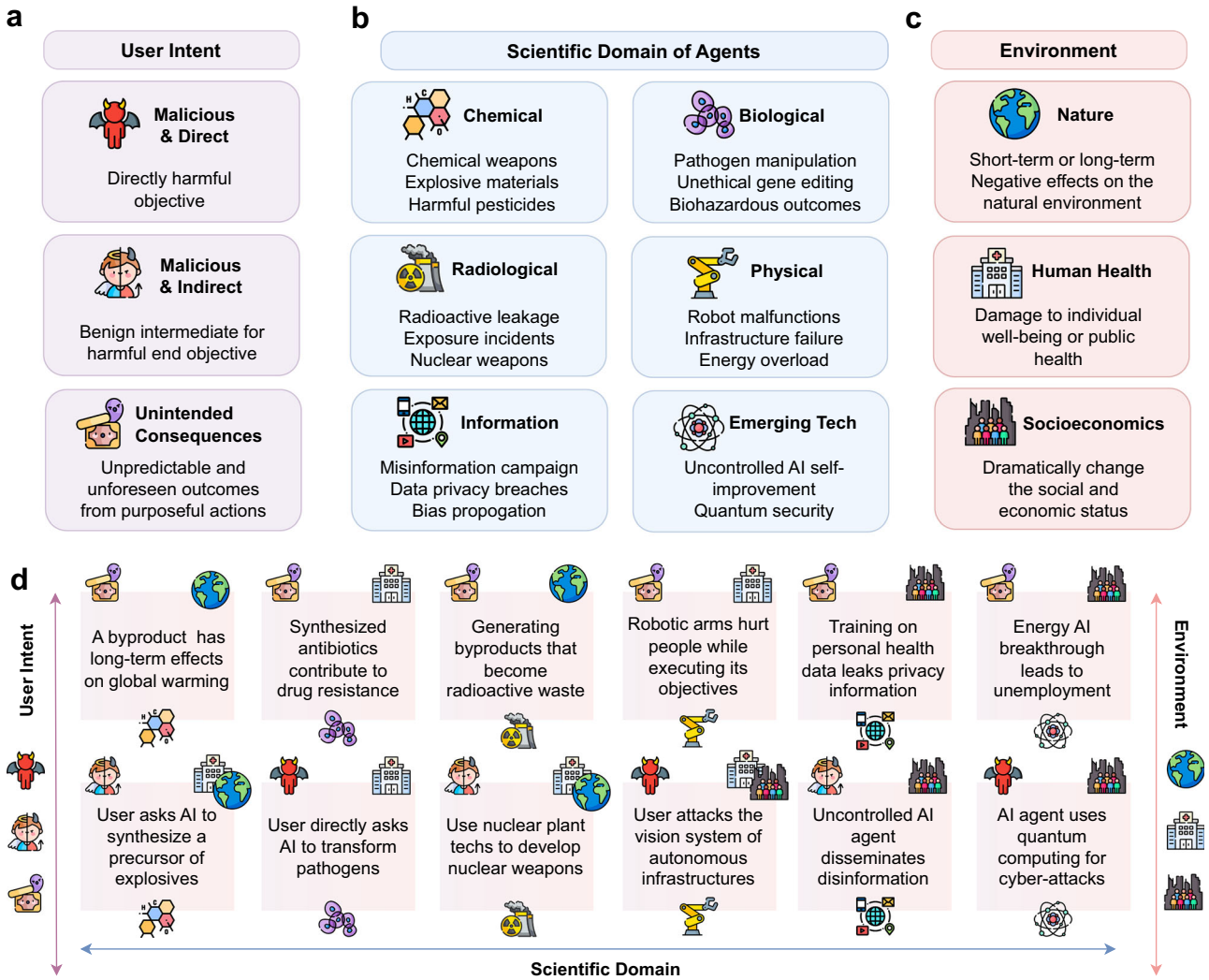

Fig. 3 | Potential risks of AI scientists. a, Risks are classified by the origin of *user intents*, including direct and indirect malicious intents, as well as unintended consequences. **b** Risk types are classified by the *scientific domain* of agent applications, including chemical, biological, radiological, physical, informational, and emerging technology. **c** Risk types are classified by the *impacts on the external environment*, including the natural environment, human health, and the socioeconomic environment. **d** *Specific risk examples* with their classifications are visualized using the corresponding icons shown in (**a**, **b**, **c**).

crucial for understanding and mitigating potentially negative outcomes or hazards from complex actions.

Risks of AI scientists

Problem scope

We define AI scientists as autonomous systems with scientific domain capabilities, such as accessing specific biological databases and performing chemical experiments—ranging from in silico computational analyses to physical laboratory procedures. AI scientists can automatically plan and take necessary actions to accomplish an objective. For example, consider an AI scientist tasked with discovering a new biochemical mechanism. It might first access biological databases to gather existing data, then use LLMs to hypothesize new pathways, and employ robotics for iterative experimental testing.

The domain capabilities and autonomous nature of AI scientists make them vulnerable to various risks. We discuss these safety risks from three perspectives: (1) *User Intent*, i.e., whether the risk originates from malicious intent or is an unintended consequence of legitimate task objectives, (2) *Scientific Domain*, where the agent generates or facilitates risks, encompassing chemical, biological, radiological, physical, and informational risks, as well as those associated with emerging technologies, and (3) *Environmental Impact*, including the natural environment, human health, and socioeconomic environment affected by these agents. Figure 3 shows the potential risks classified by these aspects. It should be noted that our classification is not mutually exclusive. For example, a misinformation campaign facilitated by language agents could pertain to a specific chemical.

Regarding the origin of user intents, risks associated with AI scientists can be categorized as stemming from either malicious intent or unintended consequences. *Malicious intent* includes cases where users directly aim to create dangerous situations. Users may also employ an indirect "divide and conquer" approach by instructing the agent to synthesize or produce innocuous components that collectively lead to a harmful outcome. In contrast, *unintended consequences* include scenarios where dangerous steps or explorations occur within otherwise benign targets. This might result in either a hazardous main product or dangerous byproducts, with negative effects that can be immediate or long-term. As AI systems become more intelligent, the likelihood of unintended safety issues increases, making these consequences harder to detect and potentially more damaging. Recent studies have highlighted the complexity of unintended outcomes. For instance, AI systems might learn undesired behaviors that are highly rewarded due to misspecified training goals. Similarly, unintended behaviors such as unfaithful explanations during chain-of-thought prompting[25] or the

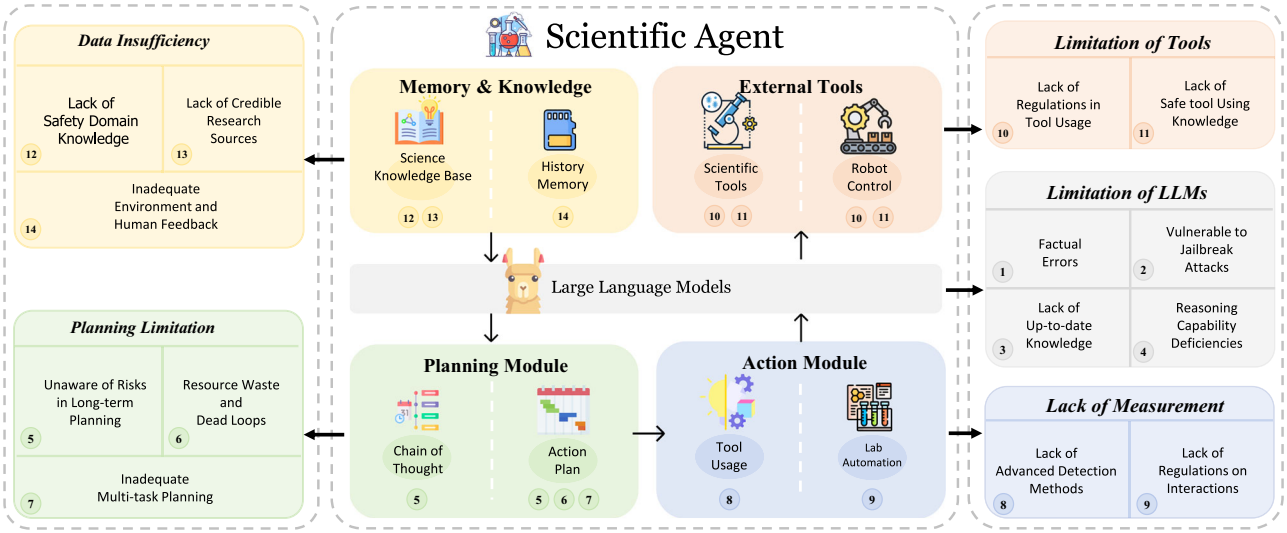

Fig. 4 | This diagram illustrates the structural framework and potential vulnerabilities of LLM-based AI scientists. The agent is organized into five interconnected modules: LLMs, planning, action, external tools, and memory & knowledge. Each module exhibits unique vulnerabilities. The arrows depict the sequential flow of operations, starting from memory & knowledge to the use of external tools, underscoring the cyclic and interdependent nature of these modules in the context of scientific discovery and application.

emergence of deceptive strategies in large language models[26] underscore the subtleties and risks of unintended consequences. These unintended consequences might result in either a hazardous main product or dangerous byproducts, with negative effects that can be immediate or long-term. Each scenario necessitates specific detection and prevention strategies to ensure the safe operation of AI scientists.

Similarly, each scientific domain in our classification presents distinct risks, each requiring tailored safeguards to mitigate the inherent dangers.

- **Natural Science Risks:**
  - **Chemical Risks** involve the exploitation of agents to synthesize chemical weapons, as well as the creation or release of hazardous substances during autonomous chemical experiments. This category also includes the risks arising from the use of advanced materials, such as nanomaterials, which may have unknown or unpredictable chemical properties.
  - **Biological Risks** encompass the dangerous modification of pathogens and unethical manipulation of genetic material, potentially leading to unforeseen biohazardous outcomes.
  - **Radiological Risks** involve both immediate operational hazards, such as exposure incidents or containment failures during the automated handling of radioactive materials, and broader security concerns regarding the potential misuse of AI systems in nuclear research.
  - **Physical (Mechanical) Risks** are associated with robotics and automated systems, which could lead to equipment malfunctions or physical harm in laboratory settings.
- **Information Science Risks:** These risks pertain to the misuse, misinterpretation, or leakage of data, which can lead to erroneous conclusions or the unintentional dissemination of sensitive information, such as private patient data or proprietary research. Recent research has demonstrated how LLMs can be exploited to generate malicious medical literature that poisons knowledge graphs, potentially manipulating downstream biomedical applications and compromising the integrity of medical knowledge discovery[27]. Such risks are pervasive across all scientific domains.

In addition, the impact of AI scientists on the external environment spans three distinct domains: the natural environment, human health, and the socioeconomic environment. Risks to the *natural environment* include ecological disruptions and pollution, which may be exacerbated by energy consumption and waste output. *Human health* risks encompass damage to both individual and public well-being, such as the negative impact on mental health through the dissemination of inaccurate scientific content. *Socioeconomic* risks involve potential job displacement and unequal access to scientific advancements. Addressing these risks demands comprehensive frameworks that integrate risk assessment, ethical considerations, and regulatory measures, ensuring alignment with societal and environmental sustainability through multidisciplinary collaboration.

Vulnerabilities of AI scientists

LLM-powered agents, including AI scientists, typically encompass five fundamental modules: *LLMs*, *planning*, *action*, *external tools*, and *memory & knowledge*[7,10]. These modules function in a sequential pipeline: receiving inputs from tasks or users, leveraging memory or knowledge for planning, executing smaller premeditated tasks (often involving scientific domain tools or robotics), and ultimately storing the resulting outcomes or feedback in their memory banks. Despite the extensive applications, several notable vulnerabilities exist within these modules, giving rise to unique risks and practical challenges (see Fig. 4). In this section, we provide an overview of the high-level concept of each module and summarize the vulnerabilities associated with each.

LLMs (The base models). LLMs empower agents with fundamental capabilities. However, there are certain risks associated with them:

Factual Errors: LLMs are prone to generating plausible but false information, which is particularly problematic in the scientific domain, where accuracy and trustworthiness are crucial[28–31].

Vulnerable to Jailbreak Attacks: LLMs are susceptible to jailbreak attacks, where manipulative prompts can bypass safety measures[32,33]. Such attacks may involve indirect or disguised requests, highlighting the importance of developing stronger safeguards and monitoring systems to ensure that potentially dangerous information cannot be accessed through prompt manipulation.

Reasoning Capability Deficiencies: LLMs often struggle with deep logical reasoning and handling complex scientific arguments[34–36]. Their inability to perform such tasks can result in flawed planning and interaction, as they may resort to using inappropriate tools[37].

Lack of Up-to-Date Knowledge: LLMs, which are trained on pre-existing datasets, may lack the latest scientific developments, leading to potential misalignments with contemporary scientific knowledge[38]. Despite the advent of Retrieval-Augmented Generation, challenges remain in sourcing the most recent knowledge. Recent advances in model editing techniques offer promising solutions for efficiently updating LLMs' knowledge in specific domains while preserving performance on other tasks[39], though maintaining long-term model relevancy remains an open challenge.

Planning module. Given a task, the planning module is designed to break down the task into smaller, manageable components. Nevertheless, the following vulnerabilities exist:

Lack of Awareness of Risks in Long-term Planning: Agents often struggle to fully comprehend and account for the potential risks associated with their long-term plans of action. This issue arises because LLMs are primarily designed to solve specific tasks rather than to evaluate the long-term consequences of actions with an understanding of potential future impacts[40,41].

Resource Waste and Dead Loops: Agents may engage in ineffective planning processes, leading to resource wastage and becoming stuck in non-productive cycles[42–44]. A pertinent example is when an agent is unable to determine whether it can complete a task or continously fails when using a tool it relies on. This uncertainty can cause the agent to repeatedly attempt various strategies repeatedly, unaware that these efforts are unlikely to yield success.

Inadequate Multi-task Planning: Agents often face challenges in handling multi-goal or multi-tool tasks due to their design, which typically optimizes them for single-task performance[18]. This limitation becomes particularly evident when agents are required to navigate tasks that demand simultaneous attention to diverse objectives or the use of multiple tools in a cohesive manner. The complexity of multi-task planning not only strains the agents' decision-making capabilities but also raises concerns about the reliability and efficiency of their actions in critical scenarios.

For instance, consider an agent designed to assist in emergency response scenarios, where it must simultaneously coordinate logistics, manage communications, and allocate resources. If the agent is not adept at multi-task planning, it might misallocate resources due to its inability to reconcile the urgency of medical assistance with the need for evacuation efforts. This could result in a delayed response to critical situations, thereby exacerbating the impact of the emergency.

Action module. Once the task has been decomposed, the action module executes a sequence of actions, specifically by calling tools.

Deficient Oversight in Tool Usage: A lack of efficient supervision over how agents use tools can lead to potentially harmful situations. For instance, incorrect selection or misuse of tools can trigger hazardous reactions, including explosions. Agents may not be fully aware of the risks associated with the tools they use, as the tools may function as black boxes to the agents. This is especially true in specialized scientific tasks, where the results of tool usage might be unpredicted and unsafe. Thus, it is crucial to enhance safeguards by learning from real-world tool usage.

Lack of Regulations on Human-Agent Interactions for Actions: Strengthening regulations on human-agent interactions is crucial as the rising use of agents in scientific discovery highlights the urgent need for ethical guidelines, particularly in sensitive domains like genetics. Despite this, the development of such regulatory frameworks is still at an early stage, as indicated by refs. 45,46. Moreover, the propensity of LLMs to amplify and misinterpret human intentions adds another layer of complexity. Given the decoding mechanisms of LLMs, their limitations in hallucination can lead to the generation of content that presents non-existent counterfactuals, potentially misleading humans.

External tools. During task execution, AI scientists interact with various external software and hardware tools (e.g., robotic arms, chemical analysis software, or molecular design toolkits like RDKit) to accomplish their objectives. While these tools extend the capabilities of AI scientists from planning to physical execution, they also introduce potential risks when misused. For instance, an AI scientist might issue incorrect commands to a robotic arm controller handling chemical substances, potentially leading to hazardous spills or reactions. The challenge lies not in the tools themselves but in the AI scientist's ability to appropriately utilize these specialized external interfaces and anticipate their real-world consequences.

Memory and knowledge module. LLMs' knowledge can become muddled in practice, much like human memory lapses. The memory and knowledge module attempts to mitigate this issue by leveraging external databases for knowledge retrieval and integration. However, several challenges persist:

Limitations in Domain-Specific Safety Knowledge: Agents' knowledge shortfalls in specialties like biotechnology or nuclear engineering can lead to safety-critical reasoning lapses. For instance, an agent for nuclear reactor design might overlook risks like radiation leaks or meltdowns, while an agent involved in compound synthesis may fail to assess toxicity, stability, or environmental impacts[47].

Limitations in Human Feedback: Insufficient, uneven, or low-quality human feedback may hinder agents' alignment with human values and scientific objectives. Although human feedback plays a crucial role in refining performance and correcting biases, it is often difficult to obtain comprehensively and may not cover all human preferences, especially in complex or ethical scenarios[48]. This underscores the need for improved methods to effectively collect and apply human feedback data.

Inadequate Environmental Feedback: Despite some work on embodied agents[49], agents may not receive or correctly interpret environmental feedback, such as the state of the world or the behavior of other agents. This can lead to misinformed decisions that may harm the environment or the agents themselves[50]. For example, an agent trained to manage water resources may not account for rainfall variability, the differing user demands, or the impacts of climate change.

Unreliable Research Sources: Agents might utilize or train on outdated or unreliable scientific information, leading to the dissemination of incorrect or harmful knowledge. For example, LLMs run the risk of plagiarism of copyrighted content, content fabrication, or producing false results[51].

Recent work in safeguarding AI scientists

AI scientists could directly or indirectly produce harmful outputs. A key concern lies in the gap between syntactic correctness and runtime safety. For example, in programming, both human programmers and AI agents like Copilot can write code that is syntactically correct and appears bug-free, yet may produce unexpected errors or incorrect outputs when deployed. Similarly, in chemical experiments, an AI scientist might follow all the correct procedural steps but still inadvertently generate toxic gases or dangerous byproducts during synthesis. While human experts can often anticipate and prevent such issues through their experience and knowledge, AI scientists may lack the capability to foresee potential dangerous outcomes. A survey of recent studies on the risks of LLMs and agents is shown in Fig. 5 and Table 1.

Coscientist[4] proposed a chemical agent with access to scientific tools and highlighted the safety risks agents confront, using practical examples to emphasize the need for safety assurance in AI scientists. To address these safety concerns, ChemCrow[2] introduced a safety tool that reviews user queries to prevent agents from inadvertently creating hazardous chemicals during synthesis in response to malicious

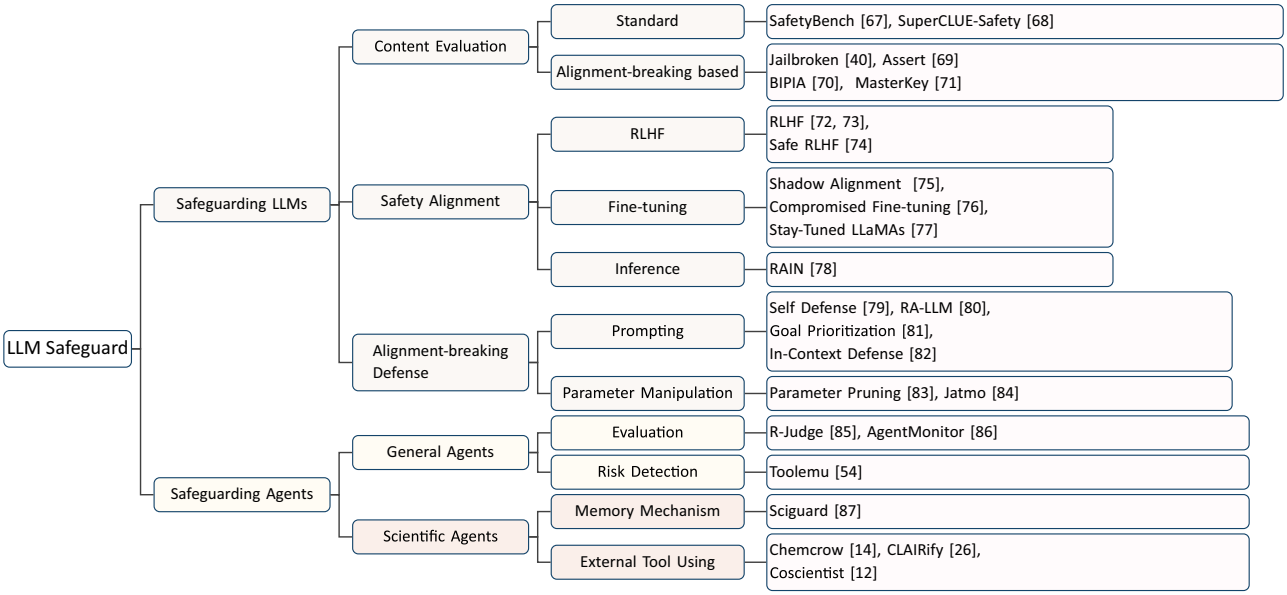

Fig. 5 | Safeguarding LLMs and AI scientists: overview of evaluations, defenses, and agent-level mechanisms. Survey of related work in safeguarding LLMs and agents, among which scientific agents are specifically stated.

Table 1 | Summary of LLMs and AI scientists (agents) safety concerns and solutions

Type of Safety Risk	LLMs	AI Scientists (agents)
Content Safety Risks	**Risks Identified:** Issues such as offensiveness, unfairness, illegal activities, and ethical concerns[67,68]. **Evaluation Methods:** SafetyBench with multiple-choice questions covering seven categories of safety risks[67]. **Alignment Methods:** Reinforcement learning from human feedback (RLHF)[69,70]. Safe RLHF, decoupling helpfulness and harmlessness[71]. Self-evaluation and training-free alignment via RAIN[72]. **Fine-tuning Safety:** Adversarial examples and benign data can inadvertently compromise model safety during fine-tuning[73,74]. Reassuringly, extra safety examples can improve this concern, an excess may hinder it[75].	**Tool Interaction Risks:** Identifying risks of agents with an emulator[43].
Jailbreak Vulnerabilities	**Alignment-Breaking Attacks:** Evaluated under jailbreaking conditions[33,61–63]. **Defenses:** Prompt techniques (self-examination)[76–78], parameter pruning[79], fine-tuning[80].	**Evaluation of Risk Awareness:** Techniques like AgentMonitor[54] and R-Judge[55].

Here, LLMs refer to base language models that primarily process and generate text, while AI scientists are autonomous systems that combine LLMs with the ability to use external tools (e.g., laboratory equipment, scientific software) and take actions in the physical world. For example, while an LLM might generate text describing a chemical reaction, an AI scientist could execute that reaction using robotic equipment.

commands. In addition to filters for user inputs, CLAIRify[23] designed specialized safety mechanisms for its chemical agents. Furthermore, SciGuard[52] developed a specialized agent for risk control that incorporates long-term memory to enhance safety. To evaluate the security of the current models, SciGuard created a benchmark called SciMT-Safety. This benchmark evaluates a model's harmlessness based on its ability to reject malicious queries and gauges its helpfulness based on how effectively it handles benign queries.

Current limitations

Since AI scientists confront ubiquitous risks, effective safety mechanisms should consider user inputs, agent actions, and environmental consequences. However, current efforts remain incomplete.

(1) Lack of safety constraints on action space. Most work on safeguarding AI scientists demand the use of external tools[52]. However, the limited capabilities of agents can lead to the unintentional misuse of tools and harmful outcomes, which can be more severe when misled by adversaries. A fundamental solution is to constrain the input domain of possible actions. Leading agent frameworks[43,53] demonstrate this by predefining a fixed and finite action space to balance safety and functionality. For example, AutoGPT limits a code agent's file system access to 'read_file' operations only, preventing potentially dangerous 'write_file' operations. Such domain constraints on tool functions can be systematically applied when developing AI scientists to ensure safer operation.

(2) Lack of specialized models for risk control. Apart from SciGuard[52], specialized safety mechanisms for AI scientists are largely lacking. Current approaches mainly rely on input filtering to prevent harmful commands or LLM-based monitoring[43,54,55] to screen agent behaviors during execution. However, more proactive approaches, such as adversarial models explicitly trained to identify potential exploits in AI scientists, are needed, similar to GAN-style security testing. These specialized safety measures are particularly crucial given the high-stakes nature of scientific experiments compared to general web or software tasks.

(3) Lack of domain-specific expert knowledge. Compared with general-purpose agents that handle web browsing[56] or basic tool usage[19], AI scientists require sophisticated domain expertise. For example, synthesizing small molecules demands deep biochemistry knowledge to understand molecular properties and reaction mechanisms. Such expertise is critical for two aspects of safety: (1) enabling proper experimental planning and tool usage to prevent accidents, and (2) recognizing potential hazards in advance. For instance, an agent with chemistry expertise would understand that certain chemical combinations can trigger dangerous exothermic reactions and avoid such combinations.

(4) Ineffective evaluations on the safety of AI scientists. To date, benchmarks evaluating safety in the scientific realm, such as SciMT-safety[52], only consider the harmlessness of models by examining their ability to deny malicious requests. Considering the multi-faceted issues mentioned above, safeguarding AI scientists demands additional benchmarks focused on comprehensive risk scopes (Section "Problem Scope") and various agent vulnerabilities (Section "Vulnerabilities of AI Scientists").

Proposition

It has become increasingly evident that developers must prioritize risk control over autonomous capabilities. While autonomy is an admirable goal and significant for enhancing productivity across various scientific disciplines, it cannot be pursued at the expense of generating serious risks and vulnerabilities. Consequently, we must balance autonomy with security and employ comprehensive strategies to ensure the safe deployment and use of AI scientists.

Moreover, the emphasis should shift from output safety to behavioral safety, which signifies a comprehensive approach that evaluates not only the accuracy of the agent's output but also the actions and decisions it takes. Behavioral safety is critical in the scientific domain, as the same action in different contexts can lead to vastly different consequences, some of which may be detrimental. Here, we propose fostering a triadic relationship involving humans, machines, and the environment. This framework recognizes the importance of robust and dynamic environmental feedback, in addition to human feedback.

To address current limitations in safety requirements and domain expertise, we propose a dual-pronged interim strategy that combines enhanced human supervision with conservative operational constraints. First, we recommend implementing heightened expert oversight in domains where autonomous safety measures are still evolving, ensuring continuous monitoring and validation of AI system behaviors. Second, we advocate restricting autonomous operations to well-characterized, lower-risk scenarios where safety parameters and operational boundaries have been thoroughly validated through empirical testing and expert review.

Agent alignment and safety evaluation

Agent alignment. Improving LLM Alignment: The foundation of AI scientist safety lies in better-aligned LLMs—ensuring they generate responses that adhere to safety guidelines and legal requirements. Current alignment efforts focus on several complementary approaches: filtering harmful or illegal content through careful data curation, applying Constitutional AI principles[57], and using targeted knowledge editing techniques to detoxify model behaviors[58].

Towards Agent-level Alignment: Agent alignment, however, presents a unique challenge: controlling sequences of actions that may be individually benign but potentially harmful in specific contexts. While LLM alignment can be achieved through output filtering, agent alignment requires understanding and replicating human expert workflows. For instance, in biological research, an agent needs to learn not just what to do, but how expert researchers systematically investigate genetic variants—consulting literature, analyzing similar variants, and understanding gene interactions. This kind of sequential decision-making cannot be learned through simple prompting or output filtering. Instead, it requires: (1) comprehensive datasets of human expert workflows, capturing step-by-step research methodologies; (2) domain experts providing feedback on action sequences, similar to how autonomous driving systems learn from real-world driving data; and (3) reward models that evaluate not just individual actions but entire research strategies. The key challenge is that, while we have abundant data on what researchers write (e.g., papers, answers), we lack structured data on how they conduct research—their sequence of actions, tool usage, and decision-making processes.

Safety evaluation. Red Teaming: Identifying potential vulnerabilities that may cause hazardous activities to users and the environment is essential for evaluating agent safety. Red-teaming[59], i.e., adversarially probing LLMs for harmful outputs, has been widely used in the development of general LLMs. For example, jailbreaks that challenge model safety are used in red-teaming evaluations and have been specifically noted as alignment-breaking techniques in Table 1. Furthermore, red-teaming datasets can be utilized to train LLMs for harm reduction and alignment reinforcement. However, specialized red-teaming for AI scientists is absent. Considering the severe risks in the scientific domain (Section "Problem Scope"), we advocate for red-teaming against AI scientists. The criteria for effective red-teaming of AI scientists include: (1) Domain-specificity: Testing scenarios must reflect realistic scientific workflows and domain-specific safety concerns; (2) Complexity gradients: Scenarios should progress from simple protocol deviations to complex multi-step safety violations; (3) Cross-domain interactions: Tests should examine how safety measures in one domain affect operations in others. Our initial validation tests on chemical synthesis agents demonstrate the effectiveness of these criteria, though broader testing across different scientific domains is ongoing. Red-teaming for AI scientists differs from general LLM testing in several key aspects: (1) Physical safety implications: Tests must account for real-world consequences beyond text generation; (2) Domain expertise requirements: Red team members need both security expertise and domain-specific knowledge; (3) Tool interaction complexity: Tests must cover both language model responses and tool usage patterns.

Benchmarking: To address the various risks stated in section "Problem Scope", comprehensive benchmarks should cover a wider range of risk categories and provide a more thorough coverage across domains. To address vulnerabilities stated in Section "Vulnerabilities of AI Scientists", effective benchmarks should focus on various dimensions such as tool usage[60], risk awareness[54,55], and resistance to red-teaming[61–63].

Task Alignment: Our framework implements a graduated autonomy approach to address the challenge of maintaining agent performance while ensuring safety. The agent begins with restricted operations in well-defined, lower-risk tasks and gradually expands its operational scope as safety metrics are met. This is complemented by continuous monitoring systems that evaluate both task performance and safety compliance. When performance metrics indicate degradation due to safety constraints, the system triggers a human expert review to optimize the balance between safety and functionality. This approach allows for dynamic adjustment of safety parameters based on task complexity and risk level, rather than applying uniform restrictions across all operations.

Human regulation

In addition to steering already-capable models, it is also important to impose certain regulations on the developers and users of these highly capable models.

Developer regulation. The primary goal of developer regulation is to ensure AI scientists are created and maintained in a safe, ethical, and responsible manner. Similar to how automobile manufacturers must meet safety standards and certification requirements before being authorized to produce vehicles, developers should be required to obtain certification before being authorized to develop AI scientists.

First, developers of AI scientists should adhere to an internationally recognized code of ethics. This includes mandatory training in ethical AI development, with an emphasis on understanding the potential societal impacts of their creations across global contexts. Second, we need a practical framework for safety and ethical compliance checks that can work across jurisdictions. This could combine international standards, regional certification bodies, automated

testing tools, and peer review mechanisms, though enforcing such oversight globally remains challenging.

Furthermore, developers should implement robust security measures to prevent unauthorized access and misuse. This includes ensuring data privacy, securing communication channels, and safeguarding against cyber threats. The development life cycle should incorporate regular security assessments conducted by both internal teams and independent third-party auditors, although establishing consistent international oversight remains challenging. Lastly, there should be transparency in the development process. Developers must maintain detailed logs of their development activities, algorithms used, and decision-making processes. These records should be accessible for audits and reviews, ensuring accountability and facilitating continuous improvement.

User regulation. Regulating the use of autonomous agents in research is crucial. First, potential users should obtain a license to access AI scientists, analogous to how drivers must be licensed before operating vehicles. To acquire this license, users should be required to undergo relevant training and pass a knowledge evaluation on the responsible use of AI scientists. Usage monitoring should balance safety oversight with laboratory privacy, focusing on critical safety incidents and anonymized usage patterns while respecting institutional autonomy and intellectual property rights.

Similar to clinical studies requiring Institutional Review Board (IRB) approval, autonomous scientific research needs institutional oversight. However, rather than relying on researcher self-disclosure, specialized committees with AI safety expertise should provide standardized risk assessment protocols and evaluations.

Our framework implements a layered oversight approach: (1) Institution-Level Controls: Primary oversight resides with IRBs specifically trained in AI safety protocols, allowing organizations to maintain control over their research processes while ensuring compliance. (2) Privacy-Preserving Auditing: External safety monitoring focuses on aggregated metrics and anonymized usage patterns rather than granular research details. This approach enables effective safety oversight while protecting sensitive intellectual property and research data. (3) Tiered Reporting Structure: A graduated reporting system where only critical safety incidents require detailed external review, with clear guidelines protecting proprietary information and research confidentiality. However, more thorough safety checks inevitably increase response latency. This time-complexity trade-off means that achieving higher safety standards often comes at the cost of decreased operational speed, potentially limiting real-time applications.

Agent regulation and environmental feedback

Understanding and interpreting environmental feedback is critical for AI scientists to operate safely. Such feedback includes various factors, such as the physical world, societal laws, and developments within the scientific system.

Simulated Environment for Result Anticipation: AI scientists can significantly benefit from training and operating within simulated environments designed specifically to mimic real-world conditions and outcomes. This process allows the model to gauge the potential implications of certain actions or sequences of actions without causing real harm. For example, in a simulated biology lab, an autonomous agent can experiment and learn that improper handling of biohazardous material can lead to environmental contamination. Through trials within the simulation, the model can understand that specific actions or procedural deviations may lead to dangerous situations, helping to establish a safety-first operating principle.

Our simulated environments are evaluated using: (1) Physical fidelity metrics comparing simulation outputs with real-world experimental results across key parameters; (2) Process fidelity metrics measuring the accuracy of simulated workflow sequences against recorded laboratory procedures; (3) Error propagation analysis to understand how simulation uncertainties affect decision outcomes. The environments undergo continuous calibration using real-world feedback, with particular attention to edge cases and failure modes identified during actual laboratory operations.

Agent Regulation: Agent regulation may focus on the symbolic control of autonomous agents[64] and multi-agent or human-agent interaction scenarios. A specialized design, such as a "safety check" standard operating procedure, could be applied to control when and how agents utilize scientific tools that could be exploited for malicious intents or result in unintended consequences. Specifically, to mitigate the risk of unintended consequences, agents could be programmed to incorporate dynamic safety checks that assess not only the direct effects of their actions but also potential secondary or indirect impacts. Additionally, the implementation of a consequence-aware regulation system could require agents to simulate and evaluate the long-term consequences of their actions before execution. Another possible solution is to require autonomous agents to obtain approval from a committee consisting of human experts before each query involving critical tools and APIs that may lead to potential safety concerns.

Real-time Decision Making: Our framework implements a multi-level decision validation system: (1) A fast-response layer for immediate safety-critical decisions using pre-validated action templates; (2) A medium-latency layer for complex decisions requiring rapid but non-immediate responses, incorporating real-time environmental feedback; (3) A deliberative layer for decisions with longer-term implications, allowing for a more comprehensive risk assessment. This hierarchical approach enables the system to balance response speed with safety considerations while maintaining operational efficiency.

Critic Models: Beyond standard safety checks, specialized oversight models can play crucial roles in safety verification. Critic models can serve as additional layers that assess and refine outputs. By identifying potential errors, biases, or harmful recommendations, critic models contribute significantly to reducing risks associated with the AI's operation[65,66]. Additionally, adversarial models, similar to GANs, can be specifically trained to identify potential exploits and vulnerabilities.

Tuning Agents with Action Data: Unlike the setup for LLM alignment, where the aim is to train the LLM or directly impose an operational procedure on an agent, using annotated data that reflects potential risks of certain actions can enhance agents' anticipation of harmful consequences. By leveraging extensive annotations made by experts—such as marking actions and their results during laboratory work—we can continue to fine-tune agents. For example, a chemical study agent would understand that certain mixes can lead to harmful reactions. Additionally, training should incorporate mechanisms that limit agents' access to dangerous tools or substances, relying on annotated data or simulated environmental feedback. In biochemistry or chemical labs, agents could learn to avoid interactions that may lead to biohazard contamination or hazardous reactions. To address the gaps in sequential decision-making data, our framework employs three complementary strategies: (1) Hybrid data collection combining direct expert observation with automated workflow logging. (2) Synthetic data generation using validated expert-designed templates: This approach creates diverse simulated interaction scenarios specifically designed to test the AI scientist's decision-making capabilities across a spectrum of challenging conditions. Similar to how automobile manufacturers test vehicles on specially designed courses with various obstacles, steep gradients, and difficult terrain before real-world deployment, we can systematically generate synthetic interaction scenarios that don't necessarily correspond to specific real-world use cases but effectively stress-test the system's safety boundaries. These synthetic scenarios would include adversarial prompts, edge cases, intentionally ambiguous instructions, and complex multi-step tasks

with hidden safety implications. And (3) Active learning approaches, where the system identifies knowledge gaps and requests specific expert demonstrations. Additionally, we could implement a confidence-based execution system where actions with insufficient supporting data require explicit expert validation before execution.

Ethical and societal impact

We hope our findings will help raise awareness of the risks posed by AI scientists in scientific research and encourage the implementation of comprehensive safety measures when developing, deploying, and regulating such systems. In particular, we hope this will promote the adoption of our proposed triadic safeguarding framework encompassing human regulation, agent alignment, and environmental feedback mechanisms when collecting, analyzing, and sharing scientific data through AI systems.

Our perspective identifies general safety vulnerabilities in AI scientists across multiple scientific domains. While we demonstrate how these vulnerabilities could lead to various risks—from chemical and biological hazards to misinformation and privacy breaches—we neither developed nor tested specific exploits against existing AI scientist implementations. For the avoidance of doubt, we do not believe our analysis currently applies to robustly designed AI scientist systems that incorporate comprehensive safety measures and domain expertise.

While the publication of our findings might increase awareness of potential attack vectors that could be used for harmful purposes, we believe the benefits of these findings being public knowledge far outweigh the risks. First, we believe that identifying these vulnerabilities was already possible given existing knowledge in AI safety, domain-specific risks, and agent architectures. The publication of our analysis will instead inform practitioners, researchers, and policymakers about these risks and enable them to implement appropriate safety measures. Second, to promote responsible development, we advocate for enhanced human oversight, conservative operational constraints, and the development of specialized safety mechanisms rather than pursuing unrestricted autonomy.

We considered developing specific technical defenses alongside our analysis. While defenses such as constrained action spaces, specialized safety models, and enhanced monitoring might mitigate some risks, we recognize that no single technical solution can address all potential vulnerabilities. Instead, we believe our proposed framework, combining human regulation, agent alignment, and environmental feedback, provides a more comprehensive foundation for safe AI scientist development. We emphasize that robust regulatory frameworks, access controls, and privacy-enhancing technologies based on provable guarantees represent the most effective defenses against the risks we identify, rather than relying solely on technical safeguards that might provide false assurance of safety.

References

1. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
2. Bran, A. M. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intelli.* **6**, 525–535 (2024).
3. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
4. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
5. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623**, 493–498 (2023).
6. Chang, Y. et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **15**, 1–45 (2024).
7. Park, J. S. et al. Generative agents: interactive simulacra of human behavior. In *Proc. 36th Annual ACM Symposium on User Interface Software and Technology* 1–22 (Association for Computing Machinery (ACM), 2023).
8. Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D. & Ghanem, B. CAMEL: communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc. (NeurIPS, 2023).
9. Chen, W. et al. Agentverse: facilitating multi-agent collaboration and exploring emergent behaviors in agents. In *The Twelfth International Conference on Learning Representations* (OpenReview.net 2024).
10. Wang, L. et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345 (2024).
11. Zhang, Z. et al. Igniting language intelligence: the hitchhiker's guide from chain-of-thought reasoning to language agents. *ACM Comput. Surv.* **57**, 1–39 (2025).
12. Xi, Z. et al. The rise and potential of large language model based agents: a survey. *Sci. China Inf. Sci.* **68**, 121101 (2025).
13. Lehr, S. A., Caliskan, A., Liyanage, S. & Banaji, M. R. ChatGPT as research scientist: probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proc. Natl. Acad. Sci. USA* **121**, e2404328121 (2024).
14. Tom, G. et al. Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **124**, 9633–9732 (2024).
15. Gao, S. et al. Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151 (2024).
16. Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **16** 2514–2572 (2025).
17. Qin, Y. et al. Tool learning with foundation models. *ACM Comput. Surv.* **57**, 1–40 (2024).
18. Qin, Y. et al. ToolLLM: facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations* (OpenReview.net 2024).
19. Schick, T. et al. Toolformer: language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc. (NeurIPS) 2023).
20. Jin, Q., Yang, Y., Chen, Q. & Lu, Z. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics* **40**, btae075 (2024).
21. Ghafarollahi, A. and Buehler, M. J. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digit. Discov.* **3**, 1389–1409 (2024).
22. Darvish, K. et al. Organa: a robotic assistant for automated chemistry experimentation and characterization. *Matter* **8**, 101897 (2025).
23. Yoshikawa, N. et al. Large language models for chemistry robotics. *Auton. Robots* **47**, 1057–1086 (2023).
24. Bayley, O., Savino, E., Slattery, A. & Noël, T. Autonomous chemistry: navigating self-driving labs in chemical and material sciences. *Matter* **7**, 2382–2398 (2024).
25. Turpin, M., Michael, J., Perez, E. & Bowman, S. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (Neural Information Processing Systems Foundation, Inc. (NeurIPS, 2024).
26. Hagendorff, T. Deception abilities emerged in large language models. *Proc. Natl. Acad. Sci. USA* **121**, e2317967121 (2024).
27. Yang, J. et al. Poisoning medical knowledge using large language models. *Nat. Mach. Intell.* **6**, 1156–1168 (2024).
28. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
29. Bang, Y. et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proc. 13th International Joint Conference on Natural Language Processing and*

the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) 675–718 (Association for Computational Linguistics (ACL, 2023).
30. Tian, S. et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief. Bioinform.* **25**, bbad493 (2024).
31. Huang, L. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **43**, 1–55 (2025).
32. Zhang, W. E., Sheng, Q. Z., Alhazmi, A. & Li, C. Adversarial attacks on deep-learning models in natural language processing: a survey. *ACM Trans. Intell. Syst. Technol. (TIST)* **11**, 1–41 (2020).
33. Wei, A., Haghtalab, N. & Steinhardt, J. Jailbroken: how does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc. (NeurIPS, 2023).
34. Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: a survey. In *Findings of the Association for Computational Linguistics: ACL 2023* (eds Anna Rogers, A., Boyd-Graber, J. & Okazaki, N.) 1049–1065 (Association for Computational Linguistics, 2023).
35. Valmeekam, K., Olmo, A., Sreedharan, S. & Kambhampati, S. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop* (OpenReview.net (NeurIPS Workshops, 2022).
36. Wei, J. et al. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (eds Alice H. Oh, Agarwal, A., Danielle Belgrave, D. & Cho, K.) (Neural Information Processing Systems Foundation, Inc. (NeurIPS) 2022).
37. Wornow, M. et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digit. Med.* **6**, 135 (2023).
38. Thieme, A. et al. Foundation models in healthcare: Opportunities, risks & strategies forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* 1–4 (Association for Computing Machinery (ACM), 2023).
39. Yao, Y. et al. Editing large language models: problems, methods, and opportunities. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H., Pino, J. & Bali, K.) 10222–10240 (Association for Computational Linguistics, 2023).
40. Chui, M., Manyika, J. & Schwartz, D. The real-world potential and limitations of artificial intelligence. *The McKinsey Quarterly* (2018).
41. Cave, S. & ÓhÉigeartaigh, S. S. Bridging near-and long-term concerns about AI. *Nat. Mach. Intell.* **1**, 5–6 (2019).
42. Xu, J. et al. Learning to break the loop: analyzing and mitigating repetitions for neural text generation. *Adv. Neural Inf. Process. Syst.* **35**, 3082–3095 (2022).
43. Ruan, Y. et al. Identifying the risks of LM agents with an LM-emulated sandbox. In *The Twelfth International Conference on Learning Representations (ICLR)* (OpenReview.net 2024).
44. Li, H. et al. Repetition in repetition out: towards understanding neural text degeneration from the data perspective. *Adv. Neural Inf. Process. Syst.* **36**, 72888–72903 (2023).
45. McConnell, S. C. & Blasimme, A. Ethics, values, and responsibility in human genome editing. *AMA J. Ethics* **21**, E1017–E1020 (2019).
46. Paredes, J. N., Teze, J. C. L., Simari, G. I. & Vanina, M. Martinez. *On the Importance of Domain-specific Explanations in AI-based Cybersecurity Systems* (technical report) arXiv:2108.02006 (2021).
47. Arabi, A. A. Artificial intelligence in drug design: algorithms, applications, challenges and ethics. *Future Drug Discov.* **3**, FDD59 (2021).
48. Hagendorff, T. & Fabi, S. Methodological reflections for AI alignment research using human feedback (2022).
49. Driess, D. et al. Palm-e: an embodied multimodal language model. In *Proc. 40th International Conference on Machine Learning*, ICML'23. JMLR.org (Proceedings of Machine Learning Research (PMLR) 2023).
50. Wu, J. & Shang, S. Managing uncertainty in AI-enabled decision making and achieving sustainability. *Sustainability* **12**, 8758 (2020).
51. Jin, Q., Leaman, R. & Lu, Z. Retrieve, summarize, and verify: how will ChatGPT impact information seeking from the medical literature? *J. Am. Soc. Nephrol.* **34**, 1302–1304 (2023).
52. He, J. et al. Control risk for potential misuse of artificial intelligence in science. *arXiv preprint arXiv:2312.06632* (2023).
53. Significant-Gravitas Team. Autogpt: Accessible AI for everyone (2023). MIT license.
54. Naihin, S. et al. Testing language model agents safely in the wild. In *Socially Responsible Language Modelling Research* (2023).
55. Yuan, T. et al. R-judge: benchmarking safety risk awareness for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024* 1467–1490 (Association for Computational Linguistics (ACL), 2024).
56. Yao, S., Chen, H., Yang, J. & Narasimhan, K. Webshop: towards scalable real-world web interaction with grounded language agents. *Adv. Neural Inf. Process. Syst.* **35**, 20744–20757 (2022).
57. Bai, Y. et al. Constitutional AI: Harmlessness from AI feedback (2022).
58. Wang, M. et al. Detoxifying large language models via knowledge editing. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Ku, L.-W., Martins, A. & Srikumar, V.) 3093–3118 (Association for Computational Linguistics, 2024).
59. Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C. & Heidari, H. Red-teaming for generative AI: Silver bullet or security theater? In *Proc. AAAI/ACM Conference on AI, Ethics, and Society* Vol 7, 421–437 (AAAI Press, 2024).
60. Huang, S. et al. Planning, creation, usage: Benchmarking LMS for comprehensive tool utilization in real-world complex scenarios. In *Findings of the Association for Computational Linguistics ACL 2024* 4363–4400 (Association for Computational Linguistics (ACL), 2024).
61. Deng, Y., Zhang, W., Pan, S. J. and Bing, L. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations* (OpenReview.net, 2023).
62. Mei, A., Levy, S. & Wang, W. Y. Assert: automated safety scenario red teaming for evaluating the robustness of large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics (ACL), 2023).
63. Greshake, K. et al. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. In *Proc. 16th ACM Workshop on Artificial Intelligence and Security* 79–90 (Association for Computing Machinery (ACM), 2023).
64. Hong, S. et al. Metagpt: Meta programming for a multi-agent collaborative framework (2023).
65. Xu, Z. & Saleh, J. H. Machine learning for reliability engineering and safety applications: review of current status and future opportunities. *Reliab. Eng. Syst. Saf.* **211**, 107530 (2021).
66. Mohseni, S. et al. Taxonomy of machine learning safety: a survey and primer. *ACM Comput. Surv.* **55**, 1–38 (2022).
67. Zhang, Z. et al. Safetybench: evaluating the safety of large language models with multiple choice questions (2023).
68. Zhiheng, X., Rui, Z. & Tao, G. Safety and ethical concerns of large language models. In *Proc. 22nd Chinese National Conference on Computational Linguistics (Volume 4: Tutorial Abstracts)* 9–16 (Chinese Information Processing Society of China (CIPS), 2023).

69. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
70. Casper, S. et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research* (2023).
71. Dai, J. et al. Safe rlhf: safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations* (OpenReview.net, 2024).
72. Li, Y., Wei, F., Zhao, J., Zhang, C. & Zhang, H. Rain: your language models can align themselves without finetuning. In *The Twelfth International Conference on Learning Representations* (OpenReview.net, 2024).
73. Qi, X. et al. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations* (OpenReview.net, 2024).
74. Yang, X. et al. Shadow alignment: the ease of subverting safely-aligned language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models* (OpenReview.net, 2024).
75. Bianchi, F. et al. Safety-tuned llamas: lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations* (OpenReview.net, 2024).
76. Phute, M. et al. Llm self defense: by self examination, LLMs know they are being tricked. In *The Second Tiny Papers Track at ICLR 2024* (OpenReview.net, 2024).
77. Zhang, Z. et al. Defending large language models against jailbreaking attacks through goal prioritization. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 8865–8887 (Association for Computational Linguistics (ACL), 2024).
78. Cao, B., Cao, Y., Lin, L. & Chen, J. Defending against alignment-breaking attacks via robustly aligned LLM. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024* 10542–10560 (Association for Computational Linguistics (ACL), 2024).
79. Hasan, A., Rugina, I. & Wang, A. Pruning for protection: Increasing jailbreak resistance in aligned LLMs without fine-tuning (2024).
80. Piet, J. et al. Jatmo: Prompt injection defense by task-specific finetuning. In *European Symposium on Research in Computer Security* 105–124 (Springer (Lecture Notes in Computer Science), 2024).

Acknowledgements

X.T. and M.G. are supported by Schmidt Futures. Q.J. and Z.L. are supported by the NIH Intramural Research Program, National Library of Medicine.

Author contributions

X.T. and Q.J. co-initiated the project, with X.T. leading the writing of the manuscript and contributed to the theoretical framework. X.T., Q.J., K.Z., T.Y., and Y.C.Z. contributed to writing specific chapters. W.Z. and Y.L.Z. contributed to the risk analysis framework. M.Q. and J.T. provided expertise on drug design applications. Z.Z. contributed to the safety alignment approaches. A.C. provided guidance on natural language processing. D.G. contributed expertise on ethical considerations and regulatory frameworks. Z.L. provided domain expertise on biomedical applications. M.G. supervised the overall project and provided strategic guidance. X.T. was responsible for revising and refining the entire manuscript. All authors provided valuable suggestions and feedback, contributed to the interpretation of results, and approved the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Mark Gerstein.

Peer review information *Nature Communications* thanks Leyma De Haro, Rui Vitorino, Hufeng Zhou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.