"Alongside us and not against us": Scientists on How Scientific AI Agents Can be Developed Responsibly

Malia Morgan[♥1] **Jimin Mun**[♥12]

Faeze Brahman[1] **Sarah K. Dreier**[3] **Maarten Sap**[12]

[1]Allen Institute for AI [2]Carnegie Mellon University [3]University of New Mexico

[♥]marks equal contribution.

Abstract

Scientists across disciplines are increasingly interested in leveraging scientific AI agents for accelerating discovery and automating research. However, approaches to AI agents' responsible (i.e., safe and ethical) development and use remain disconnected from scientists' actual practices, narrowly scoped to particular fields and failure types, and focused on current rather than emerging capabilities. To begin addressing this gap between interest and consensus on appropriate implementation, we conducted semi-structured interviews with $N = 17$ scientists across the natural, applied, and social sciences, examining (1) field-specific responsible research practices, (2) current AI use, (3) desired agent use cases and their risks and impacts, (4) approaches to agent evaluation and guidelines, and (5) expectations for the future of AI agents and the role of human researchers. We observe both explicit and implicit research norms, including practices around human participants, publication, and collaboration. Responsible AI use varies across researchers and fields, with notable instances of deliberate non-adoption and disagreement. Participants anticipated risks, including agent errors and long-term impacts ranging from wasted resources to deskilling, while also identifying beneficial uses (which are contingent on rigorous validation), including large-scale data processing, coding, and automating repetitive work. Scientists also described expectations for concrete safeguards, including task-specific unit tests and guardrails against mis-attribution and plagiarism. Outlooks on the future of AI-enabled science were divided, ranging from excitement to existential concern. We conclude with practical implications for developing and deploying current and future scientific AI agents in ways that build trust and empower scientific communities.

1 Introduction

A long-standing vision of Artificial Intelligence (AI) has been to accelerate science (Anthropic; Waltz and Buchanan, 2009). This goal has driven the development of scientific AI tools, ranging from domain-specific applications (e.g., AlphaFold, Jumper et al., 2021b) to general-purpose systems (e.g., Google's Co-Scientist,Gottweis et al., 2026; or Claude Science, Anthropic), attracting substantial public (Garisto and Glickman, 2026) and private investment (Google.org, 2026; OpenAI, 2026). These models, including general-purpose tools that are not specifically developed for science (e.g., Claude, ChatGPT), now touch nearly every stage of the research process, from ideation (Si et al., 2026; Zhou et al., 2024a) and literature review (Kittler, 2024) to writing (Mun et al., 2026), peer review (Thakkar et al., 2026), and execution of experiments (Kon et al., 2026). This widespread use demonstrates AI's potential to meaningfully transform how science is conducted and to offer benefits to help accelerate discovery and reduce researcher burdens

(Sajadieh et al., 2026; Liao et al., 2024b).

Beyond ongoing efforts to define and evaluate AI capabilities (Bragg et al., 2026), one substantial barrier to the adoption of AI for science have been scientists' concerns over the models' safety (Aránguiz Dias et al., 2026), reliability, and trust (Yoshikawa et al., 2023; Boiko et al., 2023). Validating these concerns, early adoption of generative AI tools in research has produced negative outcomes, including hallucinated citations (Naddaf and Quill, 2026) and the flooding of peer-review systems (Naddaf, 2026; Jones, 2026; Gibney, 2026; La Rosa and Nasser, 2026). As scientists increasingly incorporate AI tools in their research flows, scientific AI agents—i.e., AI systems given autonomous and independent access to tools to perform tasks, often without adequate safeguards and evaluations—could cause broader real-world harm through, among other things, research misconduct (Lund et al., 2023; Hosseini et al., 2024; Hao et al., 2026) and further exacerbate biases and resource disparities that hinder scientific progress (Fernández Pinto, 2023). As demonstrated by history of science, research misconduct can lead to detrimental harms (e.g., the Tuskegee Syphilis Study, Paul and Brookes, 2015; or the Stanford Prison Experiment of 1971, Zimbardo, 1973). Thus, when planning to adopt AI tools, it is as essential for a scientist to evaluate AI safety and ethics (Miteu, 2024) as it is for them to consider AI capabilities. However, what constitutes responsible research with AI agents remains unresolved, comparatively understudied, and narrow in scope (Tang et al., 2025b). Despite such limitations, the introduction of fully automated research is emerging across disciplines.[1]

In this work, we begin to fill this gap by studying scientists' perspectives about the responsible development and use of AI agents in science, which presents several challenges. First, while safety remains an important consideration for the companies building frontier models, they can nevertheless result in both catastrophic failures—such as release of toxic chemicals or misuse of radioactive instruments (Boiko et al., 2023; Zhou et al., 2024b)—and tangible real-world harms—including misrepresentation or use of private data (Tie et al., 2025; Kapania et al., 2024). Many studies of AI safety (Zhao et al., 2024a; Jiang et al., 2026; Li et al., 2024) have been narrow in scope, in disciplinary application (e.g., chemistry, biology), and in types of failures (e.g., malicious misuse and catastrophic failures). Second, responsible research practices vary across fields (Koppel et al., 2025); are shaped over time by historical, political, social, and legal contexts (Artal and Rubenfeld, 2017); and are often implicit. Such practices have not been well-documented or considered as part of the development and deployment of AI uses in science (Koppel et al., 2025). Third, lack of open engagement with the scientific communities and lack of transparency (Staufer et al., 2026) into the safety implementations and ethical practices reduce trust among scientific communities and limit adoption. Finally, because AI for science is a rapidly evolving field, anticipating its possible benefits and harms is an important but abstract exercise, required for addressing not only current but also future safety practices in AI use.

To address these challenges, we conduct semi-structured interviews with $N = 17$ scientists across natural, applied, and social sciences. This endeavor marks a step towards collaboration with scientists across diverse fields to shape responsible scientific AI agent development.[2] We construct our interviews to uncover responsible research practices across scientific fields, surfacing example use cases, a wide range of error scenarios, and various speculations of the future of AI agents in science for each field. Specifically, our interviews address the following research questions:

RQ1. What are existing practices around responsible research in each field?

RQ2. What practices do researchers employ when using generative and agentic AI?

RQ3. What are the desired use cases of AI agents, their respective risks or failures, and the foreseeable positive and negative impacts of those use cases and errors?

RQ4. How should responsible research evaluations of AI agents be implemented in practice?

RQ5. What is the anticipated future of AI agents in science and the role of researchers?

Our interviewees identified both implicit (e.g., publication and collaboration norms) and explicit (e.g., Institutional Review Board) responsible research practices. In aggregate, participants reported using AI for

[1] In one example, a fully-automated AI scientists (Tang et al., 2026) produced research that was accepted at a computer science workshop (Lu et al., 2026).

[2] In our interviews, we consider both AI agents that are specifically developed and marketed for science and general purpose models used by scientists to support their workflow.

coding, writing, ideation, and reviewing. Interviewees also described tasks for which they chose *not* to employ AI tools, along with practices they built to uphold responsible research when adopting AI (e.g., validation, "sandboxing" AI agents to specific parameters of use or access, and disclosure). In the absence of explicit institutional guidelines, these individual researcher decisions were often shaped by external pressures and informal sources, including media and colleagues. Some uses, notably writing, ideation, and reviewing drew strong disagreement across our interview pool. Participants identified beneficial and appropriate uses, such as automating repetitive tasks and processing and analyzing large datasets, as well as potential failures and misuses, including hallucination, bias, validation and task-scoping errors, and data manipulation. They also anticipated longer-term impacts of widespread AI use, ranging from wasted time and equipment to lost reputation, eroded trust in science, and deskilling. Outlooks on the future were mixed: Some interviewees were excited about the opportunity to reinvest human effort and overcome human limitations, while others described an existential crisis, fearing the loss of genuine understanding and of humans' ability to shape and understand the future through science.

We conclude with practical implications of our findings. First, we argue that AI developers and deployers alike should involve scientific communities as they shape AI tools, from identifying beneficial use cases and their acceptability criteria to co-developing community-centered evaluations. Second, we offer recommendations for building scientific trust in AI agents across five dimensions: validation, privacy, objectivity, accountability, and transparency. Finally, we discuss how scientific institutions and policymakers can prepare the scientific community for the future of scientific research through upskilling researchers on AI approaches and developing and enforcing responsible AI policies that uphold the ethics and the integrity of the scientific process.

2 Related Works

2.1 Scientific Agents

We consider three categories of AI tools used for scientific applications. The first is a domain-specific scientific AI agent, which includes targeted models (e.g. AlphaFold, Jumper et al., 2021a) and task-specific models (e.g. ChemLLM, Zhang et al., 2024a). The second encompasses general-purpose large language models (LLMs) that have been specifically trained for scientific tasks. These models may be targeted at specific domains (e.g. ChemCrow, Bran et al., 2023; or Coscientist, Boiko et al., 2023) but may also be designed for particular tasks (e.g. LitLLM for literature review, Agarwal et al., 2024; or SciAgents for idea generation, Ghafarollahi and Buehler, 2025). They may also be more general across domains or cover a wider section of the research process, like DeepScientist (Weng et al., 2026). Finally, many scientists and researchers are currently using general-purpose LLM models including Gemini (Codreanu et al., 2026) and Claude (Anthropic, 2026). These models are often easier to access and offer a broad set of capabilities across the research workflow.

2.2 Existing Responsible AI Practices

As scientists integrate AI tools, two major forms of verification and evaluation have emerged.

Community-Based Research This approach integrates current opinions from practicing scientists. Prior works (Chugunova et al., 2026; Liao et al., 2025; Van Noorden and Perkel, 2023; Arroyo-Machado et al., 2025), which include researcher surveys about their use of generative AI, have discovered an array of usages, including information seeking, administrative work, writing, and coding. Benefits have included an increase in equitability for non-native English speakers and speeding up coding or administrative tasks. Researchers expressed concerns, such as misinformation, plagiarism, hallucination, AI biases, a lack of explainability, a model's overconfidence in answers, and weakened trust (Morris, 2023; Tie et al., 2025). Prior interview works have found substantial concerns around accountability, transparency, and trust (Gautam et al., 2026a), as well as concerns about data privacy and authorship (Kapania et al., 2024). Other studies highlight the implicit difficulties resulting from different definitions of ethical conduct between fields (Lindkvist et al., 2026) and hard-to-measure systemic impacts, like the narrowing of a field's research scope (even as the number of publications expand; Hao et al., 2024). Institutional guidelines also reflect changing community-based practices, as many conferences and journals have implemented guidelines for disclosing AI use (Jones; Resnik and Hosseini, 2026).

While these prior works provide valuable insights into usage of generative AI in research, many are constrained by one or more of the following considerations: 1) narrow participation, by domain, place of employment, or stage in research career; 2) lack of specific focus on agentic AI, 2) lack of focus on ethics and safety concerns; and/or 4) sole reliance on a survey with inflexible responses. AI usage varies substantially across different domains, and the use of autonomous AI agents in science poses unique challenges alongside the challenges introduced by generative AI more generally (Chan et al., 2023). Furthermore, these works focused on general usage and sentiment rather than specifically addressing attitudes about ethics and safety concerns.

Technical Safety Approaches Existing work has highlighted that despite the a common desire among users for a full evaluation of ethical risk factors, there is still no unifying framework (Tang et al., 2025a). Despite a growing concern over the ethical implications of these systems, scientific safety benchmarks primarily focus on capabilities and red-teaming or dual use risks. Some focus on physical lab safety (e.g. SafeAgentBench with embodied agents (Yin et al., 2024), Agent-SafetyBench (Zhang et al., 2024b) and LabSafetyBench which pulls direct lab safety concerns from OSHA (Zhou et al., 2024b)). Other benchmarks evaluate dual-use capabilities (e.g. SOSBench based on existing regulations(Jiang et al., 2026), and SciSafe Eval which includes protein and genomic languages risks). (Li et al., 2024). Several benchmarks include basic ethics questions, including SciKnowEval Level 4 (Feng et al., 2024), SciMT-Safety (He et al., 2023), and SciSafetyBench (Zhu et al., 2025). However, none of these are designed around community-based implicit knowledge, most are scoped to specific scientific domains, and a large meta analysis of AI scientist frameworks found ethical governance remains one of the most important unsolved areas (Liao et al., 2024a). Example safety taxonomies focus on the risks brought by agency, instead of scientific use cases (Boddy and Joseph, 2025), or solicit input directly from AI experts rather than scientific users, leading to a focus on dual-use capabilities (Saeri et al., 2026). Some domain-specific models also include responsible research sections in their design or evaluation of the system, though they are again primarily directed at physical lab safety risks or the potential for dangerous dual use capabilities (Bran et al., 2023; Yoshikawa et al., 2023; Boiko et al., 2023; Zhang et al., 2024a; Zhao et al., 2024a).

This highlights a distinct gap between community based approaches–which surface themes like verification, hallucination, trust, and data privacy but offer limited guidance on model development and evaluation–and technical solutions, which are capable of being implemented to shape development but address themes like dual use and physical safety. This work aims to bridge that gap by engaging practicing scientists from a wide variety of domains and experience levels in discussion around specific and actionable concerns around responsible science and AI tools.

3 Methods

In order to collect a wide range of opinions from a diverse population of scientists, we recruited broadly and conducted semi-structured interviews with a three stage coding process guided by the responses themselves.

3.1 Recruitment

We reached out to two mailing list resources established by Ai2, a broad invitation to a group testing academic literature research tools, and a targeted domain-based invitation to a feedback on scientific AI systems group. We also employed a snowball method of recruitment, where all authors reached out to personal contacts and asked each contact to do the same. Participants self-identified as scientists, and we constrained acceptance to the study on being currently employed in a job that required research-based tasks and located in the US. The study received an IRB exemption memo through the lead author's institution. Invitations among eligible participants prioritized maintaining diversity across: 1) domains, 2) years of research experience, 3) self-identified gender, 4) source institutions (Table 1). Further details on the recruitment process can be found in Appendix D, and further information on participant demographics in Appendix C.

3.2 Interview Setup

Each interview followed a semi-structured design (Lazar et al., 2017), where the base questions and themes were held constant across participants and additional follow-up questions were tailored to the responses in the

interview, similar to setups in Chechelnitsky et al. (2026); Morris (2023). The base interview structure included questions about general responsible research processes, AI use practices, AI failure cases, and predictions for future AI use. (Appendix E). The interviews were conducted by one of the two first authors, recorded and transcribed by the native Google Meet Gemini-based transcription and Gemini 3.8 Flash, and each transcription was validated by an author.

3.3 Qualitative Analysis Procedure

We use three-stage coding to analyze interview data (Braun and Clarke, 2006). First, we perform open-coding where the two first authors individually identify codes based on the transcript, with each author coding the transcripts of the interviews they did not conduct using a local tool Gururaja (2026). Then, based on discussion of the codes, the two first authors agreed on a set of codes and performed closed coding. After closed coding each interview, the two first authors wrote a qualitative field note about the most important themes and takeaways of that interview. There were 120 total codes, spanning categories including AI use, negative impacts, research tasks, and ethical guidelines. The same authors then applied axial-coding and thematic analysis to organize the codes and find high-level themes with input from the full research team (Braun and Clarke, 2006). The codes and themes can be found in Appendix F. In addition, the authors collected a set of inductive observations that can be found in Appendix A.

4 Results

Table 1 Demographics (gender, domain, and years of experience in research) of interview participants.

	Gender	Domain	YoE
P01	Male	Biology	21-30
P02	Female	Chemistry	40+
P03	Male	Political Science	21-30
P04	Male	Physics	1-5
P05	Female	Biology	1-5
P06	Male	Sociology	11-15
P07	Nonbinary	Chemistry	1-5
P08	Male	Medicine	11-15
P09	Female	Physics	40+
P10	Female	Biology	6-10
P11	Female	Computer Science	11-15
P12	Male	Sociology	1-5
P13	Nonbinary	Psychology	16-20
P14	Female	Political Science	16-20
P15	Female	Astronomy	6-10
P16	Female	Engineering	6-10
P17	Female	Anthropology	16-20

Using the codes and themes that emerged, we investigate our research questions on 1) existing practices around responsible research, 2) current AI adoption and practices, 3) possible beneficial usage, failure cases, and their impact, 4) implementation of responsible research for scientific AI agents, and 5) anticipated future of AI agents and science.

4.1 RQ1. What are existing practices around responsible research in each field?

We asked participants about safety and ethical practices that they followed and whether any external guidelines existed. Participants discussed practices that fell under six themes: human participant practices, animal subject practices, equipment usage, collaboration or publication norms, and broader community norms. Participants who dealt with research on human (e.g., social sciences, medicine) or animal subjects (e.g., biology) discussed safety of participants. As noted by P13, external guidelines such as IRB requires the researchers to question *"Do the benefits of carrying out the study, what we learn, outweigh the costs of putting people into those stressful positions?"* Participants in natural sciences, such as physics, chemistry, and biology, have also noted lab equipment-handling safety. Beyond IRB, many participants pointed to explicit policies enforced by publication venues (e.g., journals) and funding institutions (e.g., the U.S. National Institute of Health or European Union funding agencies).

Many responsible research practices are implicit. Many responsible research practices stemmed from well-established implicit scientific community norms. Participants noted that responsible conduct also involves following norms around publication, such as authorship and novelty, and collaboration, such as expected contributions by collaborators. Furthermore, participants discussed responsibilities around teaching and mentorship and resource usage such as environmental, funding, equipment use, and others' time.

At a high-level, participants also discussed responsibilities of researchers to consider the impact, especially in fields that are more applied. As noted by P03 in political science, *"So there's a real sensitivity there to how our work is used and what the ethical um considerations of what responsible use of that looks like of, you know, in an applied sense."* More broadly, participants expressed commitment to scientific trust through practice of validation, transparency, and reproducibility.

Responsible research practices create trust in science. Such implicit and explicit practices have value in creating trust in science. While many practices are implicit to provide freedom of exploration, there are strong normative standards in science. As P14 expresses, these norms and practices guide standards of scientific works and build trust within and outside of the scientific community – *"We create trust through institutions and their rules. And so all of these rules of like, you know, academic rigor and scientific method have been like the the sort of backbone of academic institutions."*

4.2 RQ2. What practices do researchers employ when using generative and agentic AI?

We asked participants to describe their current AI use and how people in their field are using AI. As AI agents are relatively new in their introduction and adoption in scientific research, we asked participants to describe both generative and agentic AI usage.

AI use and non-use. When reporting on their AI use, many participants also highlighted their use boundaries (non-use). Participants described AI usage in tasks such as coding, repetitive tasks (especially those that are administrative), information seeking, literature review, writing, visualization, ideation, and learning, especially for cross-domain information.

However, some of these task areas were divided in their acceptance across participants and across scientific communities, with some respondents using AI for tasks that others specified to be in their non-use category. Specifically, we observed more divided opinions on writing and ideation tasks. For example, P11 noted, *"AI is pretty good uh for giving interesting ideas, but how to say, I don't know this is really validated method or like classic methods, because somehow I feel that it just it's just it's pretty creative and fun."* In contrast, P01 emphasized their boundaries to not take ideas from AI, saying *"I cannot use AI to generate an idea. That's my clear-cut boundary."* Similar disagreement emerged for using AI as a writing tool. P10 discussed their non-use *"I think it's probably going to be a lot of writing, which I don't know if I necessarily agree with... I think writing is still an art form, and so I think there still has to be like a human element to it"*, whereas P17 showed more openness to adoption: *"For scientific writing anyways, like I think there's currently kind of a stigma... I think much of scientific writing is very formulaic... I'm not saying like everybody should just fully AI-written paper..., but there's a lot of things, like methods and results, that you know, there's no there's no intellectual work in there. It's just like reporting."* Additionally, use of AI for reviewing others' papers was surfaced as a concerning use in the field, with P16 noting that they *"get concerned when other people use it for those applications. Primarily, like journal review is the first thing that comes to mind"*, while providing feedback to own was a beneficial usage identified by one participant *"I use AI as the reviewer on my paper"* (P11), highlighting the importance of usage context. Such acceptability criteria and usage were actively evolving for many participants, and many held conflicting views even in their own use:

> *"And then I sort of like got a lot of help on that paper, like from Claude and like revising stuff... and then I was like, 'Wait a minute. Like, I don't think I'm okay with this.' And I had to go back and rewrite that paper 100% because I don't know what was AI and what was me... I myself don't feel like I have my own compass about what is appropriate and inappropriate use. And my own compass is being influenced by the diverse voices in my life that are across different disciplines about whether AI just in this situation, just this one tiny little situation, is okay or not."* - P14,

Political Science

Some participants reported a generational divide in adoption patterns in their field, noting *"I've seen this as a generational split, where it's younger scholars are much more willing to adapt these either for assisting in code generation, assisting in research and literature development, um even things like deploying surveys or survey experiments and devising what those are going to look like, and then even once the data would be collected from humans, what the subsequent analysis is going to look like."* (P03). Furthermore, we observe that respondents from the social sciences held particularly strong views or commonly discussed perspectives on AI usage:

> *"How do you start a conversation and ask like, 'Do you use it, and how do you use it?' versus um not wanting to come off or be perceived as like I don't know, 'I'm killing the environment and, you know, causing the the destruction of society because I'm using it.' Because it feels like there's really no middle ground to some people."* - P13, Psychology[3]

AI use practices vary widely. Similar to the wide-ranging views held by participants on usage and non-usage of AI, participants had varying AI use practices developed around their adoption. We noted themes of full-adoption to fully sand-boxed usage, validation practices built around AI use, and disclosure policies adopted around AI usage by journals and individuals. Many participants also noted iterative process in their usage to ensure quality and validate outputs. Highlighting the variation in usage practices, P04 described more full-adoption, using AI for *"at least I would say 60% of it. I use it some way or other, not like fully rely on it, but if I want to have speed then yeah, I mean it's involved. I mean, I use AI a lot, actually."*, whereas P03 emphasized using AI only in sand-boxed manner *"So I have not allowed the agent to go out and do a wide internet search or or filtering of that, mainly because trying to train the agent to have the same degree of validation and filtering that I'd want to have on the front end is it's just vastly easier for me to do at this stage."*. However, common across participants was adoption of validation process such as P08 noting *"when I explore some applications from the AI findings, I always, always go back into the original publication to check the details."* and P02 adding *"But I always ask for references. So, I usually go back in the literature and corroborate what I've been suggested."*. Additionally, one participant shared his disclosure policy, in which he and his collaborators directly and transparently communicate their AI use:

> *"like a couple weeks ago, I didn't really, like, demand people disclose to me, but I kind of said like, think this is where things are headed... it was sort of like a declaration from my end that I sent out about disclosure."* - P01, Biology

What shapes AI use practices? As we observed varying adoption and practices, our interview further uncovered what shaped these decisions. Participants discussed sources of pressure and information that shaped AI adoption had themes such as information from colleagues, fear of being outdated, and institutional adoption. Such fear of becoming outdated was highlighted by P02 (*"But on the other hand, if I don't understand the tool, I don't use it, I become a dinosaur."*) whose practice was informed by colleagues with expertise *"I talked to my my friends, mathematician and people in computer science what for what I want to do and uh that's what they recommended."* Some of the reasons for non-use included concerns of data privacy, validation difficulty, and, at a higher level, human's critical role in science or the value of learning:

> *"I'm like being trained and I want to like learn all these things sort of the traditional way, I guess, before I like I want to be a PI, so I want to be able to like learn all these skills so I can train other people as well to do these skills"* - P10, Biology

The concerns of data privacy were often linked to privacy of participant data and lack of transparency in how user data is protected by the AI companies. Some participants also referred to the recent media around the possible privacy issue involved in OpenAI's Navier-Stokes solution (Chang, 2026) as a source of information towards their decision making, as noted by P10, *"the whole thing with the math person just happened, where the So that's like another example"*.

Some of the guiding principles were developed through metaphors and analogies to human-practices, some thinking of AI agents as students. This metaphor shaped practices such as validation and iterative refinement,

[3]Although generally considered a social science, certain sub-fields of psychology closely overlap with natural or biological sciences.

with P03 describing *"if what has come out the other side of that is an error... then the agent needs to be corrected, just as the student would be, right?"* Furthermore, the responsible research practices also guided their adoption and practices around AI usage:

> *"I think it's it's important that, as scientists, like the basically there's a long tradition in sociology and elsewhere of scientists harming communities through their research, and I think that this is no exception to that. And so making sure that we're using what we've learned from the past to inform how we operate today with these new tools is really important"* - P12, Sociology

We discuss the factors that influence participants' choice of AI vendors and models in Appendix B.

4.3 RQ3. What are the desired use cases of AI agents, their respective risks or failures, and the foreseeable positive and negative impacts of those use cases and errors?

In order to ground specific responsible research interventions, we defined three categories of concern and asked about them both generally and with respect to a desired AI use case. The three categories included:

- *AI Failure*: a user has a benign and well-scoped intent, and there is failure attributable to the AI agent.
- *User Error*: a benign user makes an error, and the AI agent acts as the input dictates but not as intended.
- *Misuse*: a user provides a malicious input designed to circumvent the AI agent's safeguards.

To contextualize these failures, we began by asking participants to define a step in their research process that would be most useful to fully automate using an AI tool.

Desired AI Tool Use Cases There was a strong overall theme of repetitive or manual tasks (e.g. iterative testing of method development, administrative tasks, or data processing and collection). P02 noted that the best application would be *"one of these mundane things that we really don't like to do, I think we need more of that than anything else."* Other tasks mentioned included literature review and documentation of experimental procedures to assist with reproducibility.

Positive Impacts. Participants expressed optimism around two themes. First, that the tools will speed up tedious work, freeing up time to focus on more valued aspects of their research (e.g., field work, mentoring, and ideation). P02 noted that they *"gain tremendous amount. I come from a generation where we had to go for weeks in a library just to found a process, my gosh, you know. It's like light speed."*. Secondly, participants discussed tasks that are currently beyond human abilities. This primarily involved the processing of large datasets, with several participants highlighting the possibility of AI finding cross-domain project patterns that humans were not able to see because of the volume of work in the area.

Failure Cases. Though we asked about AI failure and user error separately, participants repeatedly noted the same scenarios and coupled the two together strongly. P03 when asked specifically about AI failures noted that *"That's on the human to take responsibility and oversight. That comes down to validation"*. Common failure cases participants were concerned about in current appropriate use cases included hallucinations, biases, data privacy, coding errors, out of domain errors, and hard to detect errors, where an output may look very plausible but contain a small inaccuracy. These plausible but incorrect outputs were mentioned frequently, with P09 noting that they were concerned about a model *"giving me the wrong result with a high degree of confidence."*. P01, who isolated literature search as a desired capability pointed that they still *"wouldn't necessarily want to rely it on for subtleties of interpretation"*. The final type of failure mentioned concerned the idea that out of the box or creative thinking was something with which AI tools fundamentally struggled. P03 was concerned about a *"regression to the mean effect that's going to occur, where they've been trained on a lot of data, they're going to give average or modal responses. And so very often that's great if what we wanted was an average or modal response. But if what we need is something more out of the box or something more heterogeneous or something addressing the extreme cases, it it's going to work against us to do that"*.

User Error. Closely tied to AI failure, participants cited the need to properly define the task, validate the response, or sandbox an agent's access. Participants were concerned about users posing biased questions with false presumptions, with P10 noting that it could lead *"to what you want to hear, leading to validating or showing you only papers or only data that is aligned with your hypothesis versus the true scope."*. Other considerations aligned with current use practices of AI, with validation and data privacy being by far the most frequently mentioned. The variance of user education also concerned participants, who discussed themes including AI literacy, generational gaps in usage, and the necessity of having an expert intuition when evaluating AI actions to be able to spot errors.

Malicious Uses. Broadly, participants were not as concerned with malicious uses of AI tools as they were user error or AI failure. We asked both about malicious public users and malicious scientific users, finding that scientific misuse centered on research misconduct (e.g. plagiarism and data manipulation). Participants were split on whether any AI generated text was plagiarism or only if the AI was quoting something else, and did point out that neither were new problems, just easier with AI. When considering public misuse, some cited dual use (participants from political science, chemistry, and biology, though participants also discussed a monetary barrier to obtaining actual supplies, not just an information barrier). Other concerns centered around misinformation, with P06 giving the example *"here's this piece of misinformation, we're going to make sure this sees as many eyeballs as possible in the lead-up to an election or um, you know, in the lead-up to social unrest or something like that"*.

Negative Impacts. Participants cited three types of negative impacts, including concerns about scientific integrity, the impact of research on broader society, and the well-being of scientists. On the practice of science, participants discussed a growing disparity in access:

> *"I think that, in general, I'm most concerned about kind of the exclusionary nature of access and how AI models can be biased and can um I I wouldn't really want the people that are running AI companies and the I don't want science to be able to be bought."* - P07, Chemistry

Other participants noted that an increasing number of papers with more shallow overall contributions may overwhelm the review system for conferences and journals, or divert resources incorrectly.

Impact of research on the broader community included concerns about the environmental impact (e.g. energy and water use), but centered on a rising distrust in the process of science. As P14 explained:

> *"we create trust through norms, right? We treat, We create trust through institutions and their rules. And so all of these rules of like, you know, academic rigor and scientific method have been like the the sort of backbone of academic institutions. And if AI is sort of not committed to those rules as as as its own agency, or if people are using AI in ways that we can't tell we we can no longer enforce those rules, we can't see if that's what's happening"* - P14, Political Science

.

Impacts on scientists' well-being included themes around deskilling, labor replacement, increased stress, and a loss of community. P02 explained the impact of the growing uncertainty in the field–*"A lot of my students, they don't know what's going to happen after they graduate. And I think the worst is the social and psychological impact on human beings."*. Comments about a loss of community included frustration over mentors using AI to give advice, colleagues using it to respond to emails, and students using it to ask questions instead of reaching out to their peers. P13 pointed out that the extreme case breaks community norms, saying *"yes, I can direct it, but I would never want to ask my student to like, 'Hey, you know, just sit in front of this computer and have your one-on-one meeting with it as opposed to me, and have it provide you feedback and guidance on what you want to do.' I think that's where it feels unethical to me."*.

4.4 RQ4. How should responsible research evaluations of AI tools be implemented in practice?

We asked participants four questions on the implementation of evaluations, distinguishing between checklists of guardrails on specific tasks (tied to the ideal tasks in §4.3), when AI developers should consult scientists during

development, technical safeguards established by AI developers, and human-AI use guidelines administered by researchers.

Task-Specific Implementation Across specific tasks, participants emphasized validation and data privacy. Implementation of validation techniques varied by domain, including unit tests for coding and data analysis tasks and comparison of data processing AI methods to human-implemented methods. For literature review tools, participants mentioned guardrails that included checking the licenses of papers, visualizing knowledge cutoffs, validating claims are actually in the paper, and verifying that results are not skewed by the sources tools have access to. Data privacy concerns cited transparency in where an AI tool is getting data from and what files it is touching (e.g. human participant data and confidential in-progress research ideas). P17 pointed out that data privacy is an evolving and complicated system, saying that *"we think the community needs to have input for access to that data, if not ownership of that data, and how that data gets interpreted"*. Concerns around AI tool use for physical labs tasks included safeguards around variables like temperature or material toxicity, validation of DNA sequences, pathogenic checks of microbes, verifying that any experimental procedure minimizes waste and environmental impacts, and considering the safe working limits of a human. Finally, P03 discussed several limitations of checklists and specific task guardrails, pointing out that *"writing down what that full list is seems relatively foolish and something people are just going to spend their time trying to get around. And getting them to understand transparency and responsibility and accountability seem like the smarter things to teach and get them to stop and think about...I realized the checklists break down very quickly."*

AI Developer Responsibilities Not all participants agreed that guardrails should be added by the AI developer, with P04 arguing *"maybe we don't need it...the expert can customize the the scientific agent or AI scientist as per their own need."* P03 made a more specific argument, saying that: *"The AI has been told you're not allowed to talk about certain things that we may want to study: political violence, human rights violations...it affects the researcher's ability to engage in or even replicate the same thing"*. Similarly to per task checklists, participant remarks in favor of guardrails centered around transparency (e.g. watermarking), validation, and data privacy. Other concerns mentioned included intrinsic biases within AI agents, with P13 asking *"what is it about how a person asks questions that shapes how the agent then produces information to them? Like, does it provide different information to men versus women because of the way that because of the way that they ask a question"*.

Researcher Responsibilities When discussing human-AI use guidelines, the most frequent themes were verification, disclosure, a frustration at a lack of current clear institutional guidelines, and a distrust in companies. P10 linked verification back to trust and the scientific process–*"people should say that they've used it and in what ways, because it won't be reproducible if you ask something that can't, If I can't be given the same data and produce the same results because the AI has given us the results and we don't know what's going on in the background, then it's not good science."*.

AI Developer and Scientist Collaboration These themes centered on understanding field-specific needs, using their external guidelines, collaborating with researchers in the field including by attending conferences, and a frustration over the lack of communication, with P14 commenting that it *"doesn't feel like a two-way street"*. There was also a prominent theme concerned with AI literacy in scientists, with P06 arguing that *"knowing the limits of the tool um and knowing the potential of the tool would be really important for academics making that decision because again, a lot of them are still kind of caught, you know, five years back in where these tools are"*.

4.5 RQ5. What is the anticipated future of AI agents in science and the role of researchers?

To understand the anticipated future of science under the continued development of scientific AI agents, we asked participants to imagine 5 to 10 years into the future and describe the possible roles of agents and researchers. There was also a split between participants who were mainly concerned about AI's impact on science and those who were cautiously excited.

Many participants expressed pessimism of the future, and discussed possible existential threat for their role in science (e.g., replacement) as well as to the institution of science itself, as expressed below:

> *"So once that starts to produce knowledge, then we're all [expletive]. Well, then what is our purpose, right? It becomes very existential. You know, like we we are here to create and produce, shift, move society in a forward direction."* - P13, Psychology

Some participants also expressed lack of ability to shape or stop the future using the term *"the cat is out of the bag"* (P01, P02). However, fewer participants expressed genuine excitement as well about possible AI's contribution in science, overcoming human limitations, improved communication, and reinvesting human resources:

> *"But I just wish we I think we could do so much more and so much better science through AI that I think um those the those are very valid concerns, but I I do feel like we could be embracing AI to both help solve that problem and make our science better and more impactful."* - P17, Anthropology

Participants envisioned future role of scientists were mainly in the managerial role and expanded investment of human scientists to do more human-facing work in fields with human components.

> *"I think it would become more of a managerial overseer of AI agents type role, and ideally understanding all of the steps that AI would replace within that scientific process, so you are going through each of the steps with the AI, just kind of watching it, holding its hand somehow."* - P05, Biology

5 Discussion

Our study revealed key considerations towards responsible AI for science such as empowering scientific communities for shaping AI tools, building scientific trust in AI agents, and preparing for the future of co-existence with AI and human researchers. In this section, we present practical recommendations for developers, deployers, and policy makers.[4]

5.1 Empowering Scientific Communities in Shaping AI Tools

We asked participants how they currently use AI agents and which uses they consider appropriate and beneficial. They described a wide range of uses, including coding, literature review, automating repetitive tasks, large-scale data processing and analysis, writing, ideation, and reviewing, along with the practices they use to validate outputs. Unprompted, participants also described tasks for which they deliberately chose *not* to use AI and voiced disagreement with how others in their field use it. They were similarly divided on AI's broader impact on science: some expected it to increase equity, while others feared it would deepen disparities by letting well-resourced researchers buy scientific output through access to more expensive and capable models.

Acceptability and Impact of Task Automation These divisions suggest that AI agents can be a double-edged sword for science. They can boost productivity and enable previously infeasible progress (Wang et al., 2023), but they can also erode established practices and cause unintended harms (Do et al., 2023). Developers and deployers must therefore carefully consider impact (Prunkl et al., 2021), user needs and acceptability (Mun et al., 2024), and how AI fits into scientists' workflows (Wang et al., 2019). Because many participants' practices were still evolving, developers also need to understand where the boundaries of acceptability lie (Mun et al., 2025). These considerations can shape development decisions in several ways: prioritizing tools that many scientists want, such as coding agents (O'Brien, 2025) and large-scale data processing tools; reconsidering objectives, for example by providing author-centered feedback on papers rather than replacing peer review (Mun et al., 2026); and designing tools that meet community acceptability criteria (Jung et al., 2025).

[4]While our work focuses on AI agents, many of these recommendations extend to generative AI and additional AI tools, so we use the term AI more broadly.

Collaboration Towards Responsible Scientific AI Evaluation Robust evaluations that reflect scientific norms and practices can further support acceptability. However, many norms around responsible research are implicit and specific to particular communities, fields, or even individuals, so surfacing them requires in-depth collaboration with scientists. Static, one-size-fits-all evaluations developed without scientists' input risk missing real-world harms (Nguyen et al., 2026). For instance, participants who conduct human participant research wanted finer control over data practices, privacy, and attribution, along with evaluation guarantees for each. Participants also wanted transparency about which sources literature review agents omit, which files coding agents access, and the environmental cost of proposed experimental designs, as well as agents that push back during hypothesis ideation. These are specific, actionable concerns that developers could build benchmarks for or measure through community-centered evaluations (Hamna et al., 2026), yet current scientific AI safety benchmarks do not address them. Moreover, differences in acceptable use and concerns among participants, even within the same field, point to the need for multiple evaluations tailored to specific use cases (Hutchinson et al., 2022).

Implications Our findings thus suggest that responsible development of scientific AI agents requires sustained engagement with scientific communities. First, scientists should participate in tool design and evaluation, defining use cases, acceptability, and impact throughout development and deployment (Suresh et al., 2024; Cao et al., 2025). Second, participation should extend beyond the most visible stakeholders to the wider scientific community, rather than centering fields with large open datasets (Hao et al., 2026) or industry priorities (Tian et al., 2024). Third, because use and acceptability vary across communities, developers should account for this variability, potentially through community-specific agents (Kuo et al., 2026) and evaluation practices (Hamna et al., 2026).

5.2 Building Scientific Trust in AI Agents

Many participants expressed concern that AI could erode trust in the scientific enterprise. Across their answers, we identified five themes related to scientific trust: validation, privacy, objectivity, accountability, and transparency. Below, we discuss each theme and offer practical recommendations for building scientific AI that upholds these practices.

Validation Participants described validation as both a key part of their current AI workflows and a desired feature of future scientific AI agents. They emphasized users' responsibility for validating AI outputs, often through workflows grounded in expert intuition, yet found errors difficult to detect without system support. Agent development should therefore prioritize human oversight (Mitchell et al., 2026; Dhanorkar et al., 2026), for example through high-fidelity explanations (Dwivedi et al., 2023), robust multi-agent validation (Pan et al., 2026; Zheng et al., 2026), and transparent, readable documentation of agent work (Mittal et al., 2026; Chan et al., 2024). Furthermore, extending usability to entry-level researchers would require designing systems that help them validate outputs beyond their expertise while still supporting their learning (Fan et al., 2026).

Privacy Privacy was another concern shared across participants. It was especially critical for those working with human participant data, for whom it was a main reason for limiting AI use. Many participants also worried about their own data and intellectual property, citing risks of plagiarism and unauthorized use of their data for model training. Addressing these concerns requires methods that preserve data privacy while maintaining usability (Zhou et al., 2026), granular user control over data (Li et al., 2026; Shanmugarasa et al., 2025), and transparency about how AI providers use data (Zhang et al., 2026). Such safeguards are essential for protecting research participants and ensuring correct attribution of intellectual work.

Objectivity Objectivity has long been a challenge for scientific integrity (Fanelli et al., 2017), and many participants worried that AI could amplify existing biases. They described how user and AI biases can interact to undermine integrity, for instance when users' false presuppositions meet AI's tendency toward agreeableness and sycophancy (Cheng et al., 2026). Participants were also concerned about biases instilled through decisions made by companies and developers, whose demographic composition is highly uniform (Jensen, 2021). Addressing these concerns requires understanding, communicating, and mitigating AI biases and their interaction with human biases. This includes methods to rigorously analyze AI biases and behaviors (Chen et al., 2026; Bai

et al., 2025), transparency about training methods and data (Jiao et al., 2025), mitigation at multiple stages of the pipeline (Echterhoff et al., 2024), and epistemic friction that encourages users' critical thinking (Nitchals, 2026). More fundamentally, current training paradigms that optimize for user retention and preferences (Sharma et al., 2024) are often at odds with the objectivity and critical pushback required for scientific thinking.

Accountability and Transparency Participants often discussed accountability and transparency alongside validation. While some held that users, as experts, should take responsibility for system outputs, others questioned whether they should be liable for errors caused by agent failures. Participants also discussed disclosure practices for transparently assigning contributions when AI is involved. Authorship has traditionally carried responsibility for a paper's contents (McNutt et al., 2018), but AI agents blur this link between authorship and accountability. Current development and deployment practices compound the problem: developers often over-claim capabilities and limit access to system components and information (Paris et al., 2026), and poorly calibrated AI outputs, such as unreliable uncertainty expressions, make it harder for users to take responsibility (Gautam et al., 2026b). Responsible AI adoption in science thus requires a more robust accountability structure (Franklin et al., 2022), driven by scientific communities and supported by developers and deployers. This includes transparent, open evaluation and reporting of capabilities and failures (Burnell et al., 2023) and rigorous methods for analyzing failures and their causes (Rabanser et al., 2026).

5.3 Preparing for the Future

Our interviews also surfaced researchers' uncertainty about how AI will shape the future of science and scientists. Many participants feared being replaced and voiced existential concerns about the role of human scientists, while others were excited about AI-driven scientific progress. Participants also expressed frustration at the lack of guidance and their limited influence over how AI develops, and noted a reluctance to discuss AI openly in their communities, which could leave them less prepared for an AI-integrated future.

Upskilling Researchers Participants proposed continued learning and training in AI use as one path forward, ideally with frontier AI developers engaging directly to educate scientists and evolve alongside them. Many also emphasized that human learning and understanding are central to science. As AI becomes more involved, educating future scientists and upskilling experienced researchers will require rethinking how education is structured: identifying where AI can support learning and where it should be kept out to preserve it (Zagami, 2026).

Developing Policies and Enforcement Participants identified developing and enforcing guidelines as a key challenge for the future. Technical measures, such as AI watermarking and detection (Zhao et al., 2024b), and non-technical measures, such as disclosure policies (Jones), must be developed in tandem. The accountability of scientists, developers, and deployers for AI-assisted authorship also requires continued open discussion to establish policies with a transparent accountability structure (Raji et al., 2020). Finally, some participants raised environmental impact as an ethical consideration for responsible AI adoption, which policy can address by requiring accurate and transparent reporting of AI's environmental costs (Xiao et al., 2025).

6 Limitations

We note that while our work aimed to cover diverse fields and population of scientists, our interviews were conducted only with US-based participants and heavily represent Western-centric view of science. Therefore, future works should extend similar inquiries into larger community of scientists across diverse countries, representing wider disciplines that might not be as prevalent in the US. Furthermore, our work covered only limited demographics, excluding additional demographics that might influence experience with AI such as disabilities (Morr et al., 2024) or experience with discrimination, which might influence perception (Kingsley et al., 2024). Due to the limitations of our study format, online, over video conferencing, we interpret our participants' answers without their full lived context.

7 Conclusion

Interest in AI agents for science is growing, yet their responsible use and development have received little attention and limited collaboration with scientists. As a first step toward closing this gap, we interviewed 17 scientists across the natural, applied, and social sciences. Our study surfaces responsible research practices, patterns of AI use and non-use, beneficial and appropriate use cases, concrete failure cases, and implications for agent evaluation and the future of AI-enabled science. Based on these findings, we offer concrete recommendations for AI developers, deployers, and policymakers to support responsible scientific AI that empowers scientific communities and sustains trust in science.

Acknowledgments

The authors of this paper would like to thank Hannah Betts, Anya Porter, Victoria Graf, Allyson Ettinger, and Rory Donovan-Maiye for invaluable feedback on the interview structure and guiding questions. We would also like to thank Daniel Chechelnitsky for essential guidance on interview preparation and analysis, and Sireesh Gururaja for assistance with qualitative coding technology. Additionally, we thank Kynnedy Smith, Mingqian Zheng, and Karina Halevy for feedback on early drafts, and the rest of the Sapling Lab for assistance and support. We would also like to acknowledge Taira Anderson, the Communications team, and the Legal team at Ai2 for their support during this study. Finally, we would like to thank the participants in this study for their considerable time and thoughtful responses throughout the process.

This material is based upon work supported by the National Science Foundation under Award No. 2413244.

This research was in part supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. 140D0426C0023. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Generative AI Use

The authors of this paper used generative AI tools, including Gemini 3.8 Flash and integrated Google Meet tools, to transcribe each interview. Every transcription was verified line by line by one of the first authors. Claude Opus 5.5 was used to create the formatting of Latex tables in this paper, but did not touch table content containing participant demographics or quotes. Same Claude model was used to suggest edits for some sentences and phrases originally written by the authors that did not include any participant responses. Additionally, AI coding tool, Claude Code with Opus 5.5, was used to edit Python code for open and closed coding interface, where the outputs were tested and verified by two members of the research team.

Ethical Considerations Statement

The research was conducted under the regulations and procedures of the lead author's institution. An IRB exemption memo was obtained. All participants gave written consent for the recording, transcription, and analysis of the interview, and gave additional verbal consent at the beginning of the interview. All questions during the interview were optional, and all transcripts were anonymized prior to any analysis. Questions on the intake form that were required in order to ascertain eligibility were not optional, but either contained free-text options or prefer not to say options. All data was stored with the lead author, following data regulations from their institution. Interview participants were compensated with $50 after the completion of the interviews.

References

S. Agarwal, G. Sahu, A. Puri, I. H. Laradji, K. D. Dvijotham, J. Stanley, L. Charlin, and C. Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.

Anthropic. Claude Science, an AI workbench for scientists, is now available. URL https://www.anthropic.com/news/claude-science-ai-workbench. Anthropic News.

Anthropic. Ai for science program. URL https://www.anthropic.com/news/ai-for-science-program.

Anthropic. Early progress on accelerating drug development with claude, October 2026. URL https://www.anthropic.com/research/early-progress-drug-development.

G. Aránguiz Dias, K. Jafari, A. Griffith, C. Aránguiz Dias, G. R. Kim, L. Saadeddin, and M. J. Kochenderfer. The doctor will (still) see you now: On the structural limits of agentic ai in healthcare. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 7916–7935, 2026.

W. Arroyo-Machado, J.-H. Ma, T.-P. Chen, T. P. Johnson, S. Islam, L. F. Michalegko, and E. Welch. Generative ai and academic scientists in us universities: Perception, experience, and adoption intentions. *PLOS One*, 20:e0330416 – e0330416, 2025. URL https://api.semanticscholar.org/CorpusID:280974278.

R. Artal and S. Rubenfeld. Ethical issues in research. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 43: 107–114, 2017.

X. Bai, A. Wang, I. Sucholutsky, and T. L. Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025.

S. Boddy and J. Joseph. Regulating the agency of llm-based agents. *arXiv preprint arXiv:2509.22735*, 2025.

D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

J. Bragg, M. D'Arcy, N. Balepur, D. Bareket, B. Dalvi Mishra, S. Feldman, D. Haddad, J. Hwang, P. Jansen, V. Kishore, et al. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. In *International conference on learning representations*, volume 2026, pages 110136–110223, 2026.

A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa. URL https://doi.org/10.1191/1478088706qp063oa.

R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, et al. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138, 2023.

C. Cao, Y. Wu, X. Z. Fang, Z. Liang, L. Mamykina, L. Sbaffi, and X. Xu. Medai-scits: Enhancing interdisciplinary collaboration between ai researchers and medical experts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713926. URL https://doi.org/10.1145/3706598.3713926.

A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 651–666, 2023.

A. Chan, C. Ezell, M. Kaufmann, K. Wei, L. Hammond, H. Bradley, E. Bluemke, N. Rajkumar, D. Krueger, N. Kolt, et al. Visibility into ai agents. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pages 958–973, 2024.

K. Chang. An N.Y.U. mathematician clashed with OpenAI over a $1 million proof. *The New York Times*, Sept. 2026. URL https://www.nytimes.com/2026/09/10/science/tristan-buckmaster-openai-math-navier-stokes.html.

D. Chechelnitsky, S. Gururaja, S. Olojo, W. H. Deng, G. Attanasio, C. Zerva, and M. Sap. Translating with feeling: Centering translator perspectives within translation technologies. *arXiv preprint arXiv:2604.00758*, 2026.

L. Chen, Y. Zhang, J. Feng, H. Chai, H. Zhang, B. Fan, Y. Ma, S. Zhang, N. Li, T. Liu, et al. Ai agent behavioral science. *Humanities and Social Sciences Communications*, 2026.

M. Cheng, C. Lee, P. Khadpe, S. Yu, D. Han, and D. Jurafsky. Sycophantic ai decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352, 2026.

M. Chugunova, D. Harhoff, K. Hölzle, V. Kaschub, S. Malagimani, U. Morgalla, and R. Rose. Who uses ai in research, and for what? large-scale survey evidence from germany. *Research Policy*, 55(2):105381, 2026.

M. Codreanu, A. Imas, J. Mateos-Garcia, J. Emmens, E. Muiruri, A. Turrell, J. Jacobs, A. Kasirzadeh, A. Trisovic, Y. Chen, et al. Ai in science: Early insights. *arXiv preprint arXiv:2609.28504*, 2026.

S. Dhanorkar, S. Passi, and M. Vorvoreanu. Human oversight of agentic systems in practice: Examining the oversight work, challenges, and heuristics of developers using software agents. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 6438–6465, 2026.

K. Do, R. Y. Pang, J. Jiang, and K. Reinecke. "that's important, but...": How computer science researchers anticipate unintended consequences of their research innovations. In *CHI*. Association for Computing Machinery, 2023. URL https://wildlab.cs.washington.edu/Publications_files/Do_CHI2023.pdf.

R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM computing surveys*, 55(9):1–33, 2023.

J. M. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He. Cognitive bias in decision-making with LLMs. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.739. URL https://aclanthology.org/2024.findings-emnlp.739/.

Y. Fan, L. Yan, I. Molenaar, S. Greiff, and D. Gašević. A metacognitive approach to learning and performance in human-ai interaction. *Communications Psychology*, 4(1):150, 2026.

D. Fanelli, R. Costas, and J. P. Ioannidis. Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, 114(14):3714–3719, 2017.

K. Feng, X. Shen, W. Wang, X. Zhuang, Y. Tang, Q. Zhang, and K. Ding. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

M. Fernández Pinto. Methodological and cognitive biases in science: Issues for current research and ways to counteract them. *Perspectives on Science*, 31(5):535–554, 2023.

M. Franklin, H. Ashton, E. Awad, and D. Lagnado. Causal framework of artificial autonomous agent responsibility. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 276–284, 2022.

D. Garisto and K. Glickman. White house rolls out ai funding-and signals a new era for us science. *Nature*, 2026.

S. Gautam, H. Liu, Y. Choi, and M. Lease. How researchers navigate accountability, transparency, and trust when using ai tools in early-stage research: A think-aloud study. *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, 2026a. URL https://api.semanticscholar.org/CorpusID:287777068.

S. Gautam, H. Liu, Y. Choi, and M. Lease. How researchers navigate accountability, transparency, and trust when using ai tools in early-stage research: A think-aloud study. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 1494–1512, 2026b.

A. Ghafarollahi and M. J. Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025.

E. Gibney. How ai slop is causing a crisis in computer science. *Nature*, 2026.

Google.org. Google.org Impact Challenge: AI for Science. https://www.google.org/impact-challenges/ai-science/, 2026. Accessed: 2026-10-16.

J. Gottweis, W.-H. Weng, A. Daryin, T. Tu, P. Sirkovic, A. Myaskovsky, G. Glowaty, F. Weissenberger, A. Orlandi, D. Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pages 1–3, 2026.

S. Gururaja. interview-analysis. https://github.com/gsireesh/interview-analysis, 2026.

H. Hamna, G. Bhat, S. Mukherjee, F. M. Lalani, E. Hadfield, D. Siddarth, K. Bali, and S. Sitaram. Building benchmarks from the ground up: Community-centered evaluation of llms in healthcare chatbot settings. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791172. URL https://doi.org/10.1145/3772318.3791172.

Q. Hao, F. Xu, Y. Li, and J. Evans. Artificial intelligence tools expand scientists' impact but contract science's focus (just accepted by nature, to be online soon). *arXiv preprint arXiv:2412.07727*, 2024.

Q. Hao, F. Xu, Y. Li, and J. Evans. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 649(8099):1237–1243, 2026.

J. He, W. Feng, Y. Min, J. Yi, K. Tang, S. Li, J. Zhang, K. Chen, W. Zhou, X. Xie, et al. Control risk for potential misuse of artificial intelligence in science. *arXiv preprint arXiv:2312.06632*, 2023.

M. Hosseini, L. M. Rasmussen, and D. B. Resnik. Using ai to write scholarly publications, 2024.

B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1859–1876, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533233. URL https://doi.org/10.1145/3531146.3533233.

B. Jensen. Ai index diversity report: An unmoving needle, 2021.

F. Jiang, F. Ma, Z. Xu, Y. Li, Z. Rao, B. Ramasubramanian, L. Niu, B. Li, X. Chen, Z. Xiang, et al. Sosbench: Benchmarking safety alignment on six scientific domains. In *International Conference on Learning Representations*, volume 2026, pages 43762–43791, 2026.

C. Jiao, Y. Pan, E. Xiao, D. Sheng, N. Jain, H. Zhao, I. Dasgupta, J. W. Ma, and C. Xiong. Date-lm: Benchmarking data attribution evaluation for large language models. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2507.09424.

N. Jones. As researchers aim for universal AI disclosure guidelines, the devil is in the details. URL https://www.science.org/content/article/researchers-aim-universal-ai-disclosure-guidelines-devil-details. News article.

N. Jones. Arxiv preprint server clamps down on ai slop. *Science*, 2026.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021a.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021b.

J.-Y. Jung, D. Saxena, M. Park, J. Kim, J. Forlizzi, K. Holstein, and J. Zimmerman. Making the right thing: Bridging hci and responsible ai in early-stage ai concept selection. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, page 2992–3012, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714856. doi: 10.1145/3715336.3735745. URL https://doi.org/10.1145/3715336.3735745.

S. Kapania, R. Wang, T. J.-J. Li, T.-S. Li, and H. Shen. 'i'm categorizing llm as a productivity tool': Examining ethics of llm use in hci research practices. *Proceedings of the ACM on Human-Computer Interaction*, 9:1 – 26, 2024. URL https://api.semanticscholar.org/CorpusID:268793750.

S. Kingsley, J. Zhi, W. H. Deng, J. Lee, S. Zhang, M. Eslami, K. Holstein, J. I. Hong, T. Li, and H. Shen. Investigating what factors influence users' rating of harmful algorithmic bias and discrimination. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 12(1):75–85, Oct. 2024. doi: 10.1609/hcomp.v12i1.31602. URL https://ojs.aaai.org/index.php/HCOMP/article/view/31602.

H. Kittler. Will ai revolutionize literature reviews? *Journal of the European Academy of Dermatology and Venereology*, 38(12):2213–2214, 2024. doi: https://doi.org/10.1111/jdv.20354. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jdv.20354.

P. T. J. Kon, Q. Ding, J. Liu, X. Zhu, J. Peng, J. Xing, Y. Huang, Y. Qiu, J. Srinivasa, M. Lee, et al. Exp-bench: Can ai conduct ai research experiments? In *International Conference on Learning Representations*, volume 2026, pages 120555–120604, 2026.

L. Koppel, A. M. Lindkvist, and G. Tinghög. Scientific normative dissonance: A large-scale survey of researchers' subscription to scientific norms and counternorms across academic fields. *Journal of Economic Behavior & Organization*, 237:107140, 2025.

T.-S. Kuo, S. Liu, Q. Z. Chen, J. Seering, A. X. Zhang, H. Zhu, and K. Holstein. Botender: Supporting communities in collaboratively designing ai agents through case-based provocations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3790500. URL https://doi.org/10.1145/3772318.3790500.

G. R. M. La Rosa and M. Nasser. Ai in peer review: the elephant in the editorial room, 2026.

J. Lazar, J. H. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2 edition, 2017.

T. Li, J. Lu, C. Chu, T. Zeng, Y. Zheng, M. Li, H. Huang, B. Wu, Z. Liu, K. Ma, et al. Scisafeeval: a comprehensive benchmark for safety alignment of large language models in scientific tasks. *arXiv preprint arXiv:2410.03769*, 2024.

Z. Li, Y. Wu, Y. Li, X. Gui, and Y. Luo. Privacy control in conversational llm platforms: A walkthrough study. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2026.

Z. Liao, M. Antoniak, I. Cheong, E. Y.-Y. Cheng, A.-H. Lee, K. Lo, J. C. Chang, and A. X. Zhang. Llms as research tools: A large scale survey of researchers' usage and perceptions. *arXiv preprint arXiv:2411.05025*, 2024a.

Z. Liao, M. Antoniak, I. Cheong, E. Y.-Y. Cheng, A.-H. Lee, K. Lo, J. C. Chang, and A. X. Zhang. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions, Oct. 2024b.

Z. Liao, M. Antoniak, I. Cheong, E. Y.-Y. Cheng, A.-H. Lee, K. Lo, J. C. Chang, and A. X. Zhang. LLMs as research tools: A large scale survey of researchers' usage and perceptions. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=p0BwJk3R1p.

A. M. Lindkvist, L. Koppel, D. Andersson, D. Västfjäll, and G. Tinghög. Is research ethics discipline-specific? a survey of researchers' and ethics reviewers' views on research misconduct and questionable practices. *Research Policy*, 55(4):105435, 2026. ISSN 0048-7333. doi: https://doi.org/10.1016/j.respol.2026.105435. URL https://www.sciencedirect.com/science/article/pii/S0048733326000260.

C. Lu, C. Lu, R. T. Lange, Y. Yamada, S. Hu, J. Foerster, D. Ha, and J. Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

B. D. Lund, T. Wang, N. R. Mannuru, B. Nie, S. Shimray, and Z. Wang. Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5):570–581, 2023.

M. K. McNutt, M. Bradford, J. M. Drazen, B. Hanson, B. Howard, K. H. Jamieson, V. Kiermer, E. Marcus, B. K. Pope, R. Schekman, et al. Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences*, 115(11):2557–2560, 2018.

M. Mitchell, A. Ghosh, and S. Passi. Ai agents push humans out of the loop. *arXiv preprint arXiv:2608.23642*, 2026.

G. D. Miteu. Ethics in scientific research: a lens into its importance, history, and future. *Annals of medicine and surgery*, 86(5):2395, 2024.

S. Mittal, S. S. Kim, S. Counts, J. Neville, D. Das, and P. Laban. Babysitting ai agents: A user-centered taxonomy and study of agentic delegation failures. *arXiv preprint arXiv:2610.34748*, 2026.

C. E. Morr, B. Kundi, F. Mobeen, S. Taleghani, Y. El-Lahib, and R. Gorman. Ai and disability: A systematic scoping review. *Health Informatics Journal*, 30(3):14604582241285743, 2024. doi: 10.1177/14604582241285743. URL https://doi.org/10.1177/14604582241285743. PMID: 39287175.

M. R. Morris. Scientists' perspectives on the potential for generative ai in their fields. *ArXiv*, abs/2304.01420, 2023. URL https://api.semanticscholar.org/CorpusID:257921814.

J. Mun, L. Jiang, J. Liang, I. Cheong, N. DeCario, Y. Choi, T. Kohno, and M. Sap. Particip-ai: A democratic surveying framework for anticipating future ai use cases, harms and benefits. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 997–1010, 2024.

J. Mun, W. B. A. Yeong, W. H. Deng, J. S. Borg, and M. Sap. Why (not) use ai? analyzing people's reasoning and conditions for ai acceptability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1771–1784, 2025.

J. Mun, C. Jung, X. Zhou, H. Kim, and M. Sap. Goodpoint: Learning constructive scientific paper feedback from author responses. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=u7u3hVx1Io.

M. Naddaf. More than half of researchers now use ai for peer review—often against guidance. *Nature*, 649(8096): 273–274, 2026.

M. Naddaf and E. Quill. Hallucinated citations are polluting the scientific literature. what can be done? *Nature*, 652 (8108):26–29, 2026.

T. Nguyen, J. Garcia-Gathright, H. Washington, A. Chouldechova, H. Wallach, and J. Wortman Vaughan. *Validating and Refining Measurements for Generative AI Evaluation Via Stakeholder Engagement*, page 6554–6586. Association for Computing Machinery, New York, NY, USA, 2026. ISBN 9798400725968. URL https://doi.org/10.1145/3805689.3812403.

C. Nitchals. Designing for disagreement: epistemic friction, sycophancy, and the ethics of human–ai advice—a critical integrative review. *AI and Ethics*, 6(5):522, 2026.

G. O'Brien. How scientists use large language models to program. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713668. URL https://doi.org/10.1145/3706598.3713668.

OpenAI. Accelerating Scientific Discovery with ChatGPT for Academic Researchers. https://openai.com/index/chatgpt-for-academic-researchers/, July 2026. Accessed: 2026-10-16.

J. Pan, Y. Liu, R. Miao, K. Ding, Y. Zheng, Q. V. H. Nguyen, A. W.-C. Liew, and S. Pan. Explainable and fine-grained safeguarding of llm multi-agent systems via bi-level graph anomaly detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30491–30506, 2026.

T. Paris, A. Moon, and J. L. Guo. Don't trust the process: When verifiability undermines ai accountability. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 5348–5370, 2026.

C. Paul and B. Brookes. The rationalization of unethical research: revisionist accounts of the tuskegee syphilis study and the new zealand "unfortunate experiment". *American journal of public health*, 105(10):e12–e19, 2015.

C. E. A. Prunkl, C. Ashurst, M. Anderljung, H. Webb, J. Leike, and A. Dafoe. Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence*, 3(2):104–110, Feb. 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00298-y. URL https://doi.org/10.1038/s42256-021-00298-y.

S. Rabanser, S. Kapoor, P. Kirgis, K. Liu, S. Utpala, and A. Narayanan. Towards a science of AI agent reliability. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=4LTwRqhmhk.

I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372873. URL https://doi.org/10.1145/3351095.3372873.

D. B. Resnik and M. Hosseini. Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary? *Accountability in research*, 33(2):2481949, 2026.

A. K. Saeri, J. Graham, M. Noetel, P. Slattery, D. Ah-King, E. Aittokallio, I. Akindehin, A. A. Mahdi, E. Alhajjar, R. A. Lipcsey, et al. Prioritization of risks from artificial intelligence: A delphi study of 272 international experts. *arXiv preprint arXiv:2606.04490*, 2026.

S. Sajadieh, L. Fattorini, R. Perrault, Y. Gil, V. Parli, L. Santarlasci, J. Pava, N. Maslej, R. Altman, E. Brynjolfsson, et al. Artificial intelligence index report 2026. *arXiv preprint arXiv:2606.15708*, 2026.

Y. Shanmugarasa, S. Pan, M. Ding, D. Zhao, and T. Rakotoarivelo. Privacy meets explainability: managing confidential data and transparency policies in llm-empowered science. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2025.

M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. Bowman, E. Durmus, Z. Hatfield-Dodds, S. Johnston, S. Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144, 2024.

C. Si, T. Hashimoto, and D. Yang. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 74195–74240, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/78d84ad7c99601a422301e4bdc4274ce-Paper-Conference.pdf.

L. Staufer, K. Feng, K. Wei, L. Bailey, Y. Duan, M. Yang, A. P. Ozisik, S. Casper, and N. Kolt. The 2025 ai agent index: Documenting technical and safety features of deployed agentic ai systems. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 1536–1576, 2026.

H. Suresh, E. Tseng, M. Young, M. Gray, E. Pierson, and K. Levy. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1609–1621, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658992. URL https://doi.org/10.1145/3630106.3658992.

J. Tang, L. Xia, Z. Li, and C. Huang. Ai-researcher: Autonomous scientific innovation. *Advances in Neural Information Processing Systems*, 38:9481–9520, 2026.

X. Tang, Q. Jin, K. Zhu, T. Yuan, Y. Zhang, W. Zhou, M. Qu, Y. Zhao, J. Tang, Z. Zhang, et al. Risks of ai scientists: prioritizing safeguarding over autonomy. *Nature Communications*, 16(1):8317, 2025a.

X. Tang, Q. Jin, K. Zhu, T. Yuan, Y. Zhang, W. Zhou, M. Qu, Y. Zhao, J. Tang, Z. Zhang, et al. Risks of ai scientists: prioritizing safeguarding over autonomy. *Nature Communications*, 16(1):8317, 2025b.

N. Thakkar, M. Yuksekgonul, J. Silberg, A. Garg, N. Peng, F. Sha, R. Yu, C. Vondrick, and J. Zou. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, 8(3):326–336, 2026.

W. Tian, R. Cai, Z. Fang, Q. Xie, Z. Hu, and X. Wang. Research funding in different sci disciplines: A comparison analysis based on web of science. *Quantitative Science Studies*, 5(3):757–777, 08 2024. ISSN 2641-3337. doi: 10.1162/qss_a_00315. URL https://doi.org/10.1162/qss_a_00315.

G. Tie, P. Zhou, and L. Sun. A survey of ai scientists. *arXiv preprint arXiv:2510.23045*, 2025.

R. Van Noorden and J. M. Perkel. Ai and science: what 1,600 researchers think. *Nature*, 621(7980):672–675, 2023.

D. Waltz and B. G. Buchanan. Automating science. *Science*, 324(5923):43–44, 2009.

D. Wang, J. D. Weisz, M. Muller, P. Ram, W. Geyer, C. Dugan, Y. Tausczik, H. Samulowitz, and A. Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov. 2019. doi: 10.1145/3359313. URL https://doi.org/10.1145/3359313.

H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Y. Weng, M. Zhu, Q. Xie, Q. Sun, Z. Lin, S. Liu, and Y. Zhang. Deepscientist: Advancing frontier-pushing scientific findings progressively. In *International Conference on Learning Representations*, volume 2026, pages 47981–48037, 2026.

T. Xiao, F. F. Nerini, H. D. Matthews, M. Tavoni, and F. You. Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. *Nature Sustainability*, 8(12):1541–1553, Dec. 2025. ISSN 2398-9629. doi: 10.1038/s41893-025-01681-y. URL https://doi.org/10.1038/s41893-025-01681-y.

S. Yin, X. Pang, Y. Ding, M. Chen, Y. Bi, Y. Xiong, W. Huang, Z. Xiang, J. Shao, and S. Chen. Safeagentbench: A benchmark for safe task planning of embodied llm agents. *arXiv preprint arXiv:2412.13178*, 2024.

N. Yoshikawa, M. Skreta, K. Darvish, S. Arellano-Rubach, Z. Ji, L. Bjørn Kristensen, A. Z. Li, Y. Zhao, H. Xu, A. Kuramshin, et al. Large language models for chemistry robotics. *Autonomous Robots*, 47(8):1057–1086, 2023.

J. Zagami. Ai refusal in higher education: the right to refuse, the duty to understand and the diagnostic value of non-use. *Higher Education Research & Development*, 45(7):2287–2294, 2026. doi: 10.1080/07294360.2026.2694409. URL https://doi.org/10.1080/07294360.2026.2694409.

D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, W. Ouyang, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024a.

S. Zhang, E. He, S. Tao, Y. Yang, Y. Ma, A. Wang, X. Yi, and H. Li. A scoping review and guidelines on privacy policy's visualization from an hci perspective. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2026.

Z. Zhang, S. Cui, Y. Lu, J. Zhou, J. Yang, H. Wang, and M. Huang. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*, 2024b.

H. Zhao, X. Tang, Z. Yang, X. Han, X. Feng, Y. Fan, S. Cheng, D. Jin, Y. Zhao, A. Cohan, et al. Chemsafetybench: Benchmarking llm safety on chemistry domain. *arXiv preprint arXiv:2411.16736*, 2024a.

X. Zhao, P. V. Ananth, L. Li, and Y.-X. Wang. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=SsmT8aO45L.

Z. Zheng, Y. Li, C. Qian, Y. Fu, Y. Fu, L. Sheng, J. Shao, and D. Liu. Stepguard: Learning step-level guardrails with scalable supervision and safety–utility balancing. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 24644–24667, 2026.

J. Zhou, N. Mireshghallah, and T. Li. Operationalizing data minimization for privacy-preserving llm prompting. In *International Conference on Learning Representations*, volume 2026, pages 56919–56943, 2026.

Y. Zhou, H. Liu, T. Srivastava, H. Mei, and C. Tan. Hypothesis generation with large language models. In L. Peled-Cohen, N. Calderon, S. Lissak, and R. Reichart, editors, *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139, Miami, FL, USA, Nov. 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4science-1.10. URL https://aclanthology.org/2024.nlp4science-1.10/.

Y. Zhou, J. Yang, Y. Huang, K. Guo, Z. Emory, B. Ghosh, A. Bedar, S. Shekar, Z. Liang, P.-Y. Chen, et al. Labsafety bench: Benchmarking llms on safety issues in scientific labs. *arXiv preprint arXiv:2410.14182*, 2024b.

K. Zhu, J. Zhang, Z. Qi, N. Shang, Z. Liu, P. Han, Y. Su, H. Yu, and J. You. Safescientist: Toward risk-aware scientific discoveries by llm agents. *arXiv preprint arXiv:2505.23559*, 2025.

P. G. Zimbardo. On the ethics of intervention in human psychological research: With special reference to the stanford prison experiment. 1973.

A Inductive Observations

There were several observations made across interviews of themes that did not necessarily fit a single research question. We observed multiple participants express frustration over a lack of regulations or guidelines, with P17 noting that *"I feel like everybody is kind of figuring that on their own, you know, is, in theory, like is would your data be safe in AI or not, you know, if you use the given guardrails that you have. Like, there's currently really no guidelines around that at all, both at the university level or within my field or journals"*. P14 similarly noted that there is *"no guidance whatsoever at my university about what software is safe, how to do that so that that data is all protected and not just going back into some other place"*.

Participants also discussed a loss of something human about the process of science, with P01 saying *"in the perfect world, like, great, the outputs are perfect. But then there's this other question of: did the person who obtained the perfect outputs get anything out of this"*, highlighting a general sentiment of concern around losing an understanding or development of a learning process.

There were many comparisons of an AI tool to a grad student or human participating in research, with P03 saying *"that agent is acting just like a student would in, say, a chemistry lab or a science lab."*, and P13 commentating that evaluating AI agentic error is like *"I think it's one of those scenarios where like if you ask a student, "Hey, can you go find papers on this topic?" and they're new, you know, they bring back what they think I'm looking for. Like, "Oh, this is Yeah, this is, you know, in the area, but like now I want you to specifically focus in like this other very very narrow thing."*

Finally, participants described how difficult it was to talk about AI usage, feeling pressures in both directions with P13 noting that *" this is something that weirdly I don't broach with my colleagues as much...I don't talk to my colleagues enough about how they use it in part because it, it feels like like for some people it's a dirty word to use."*, while P11 reflected that *"I'm sometimes intimidated that I'm so outdated, or am I using AI in a wrong way, so I'm not satisfied with their outcome, or I I should give up uh feeling uncomfortable from the AI's writing tone and just take this flow"*.

B Influences on AI Vendor Selection

We asked participants to describe what has influenced their AI selection. Some common themes were ease of access, including institutional purchasing or paid for by labs, model testing, where participants have tested a model with their own use case, and tech-stickiness where the decisions they had made became sticky with the cost of switching. Participants further noted ethical or value alignments with the company and discussed generative AI embedded into the systems such as Google search. However, with the embedded system, some participants showed frustration over lack of control. Open-source models were also preferred by some participants for advantage of more comprehensive control, especially with the non-transparent guardrails put in place by commercial labs.

C Participant Demographics

Interview participants spanned a wide range of AI usage practices and opinions. We report these here by participant for context, and report demographic summaries as well.

C.1 Demographic Summaries

- **Age Range**: 18-25(3), 26-35 (6) 36-45 (3), 46-55 (2), 56-70 (1), 65+ (1)
- **Ethnicity**: Asian (4), White (11), Indian (1), Mexican American
- **First language**: English (11), German (1), Hindi (1), Mandarin (1), Korean (1), Spanish (1)
- **Highest degree obtained**: PhD (9), Master's degree (5), Bachelor's degree (3)

Table 2 Participant's current relationship with AI tools

	AI Use	AI Opinion	AI Payment
P01	summarization, transcription	"concern", "awe"	institution, personal
P02	literature search, ideation	"pessimistic"	personal
P03	code, summarization, teaching	"positive"	institution
P04	full process, excluding ideation and data analysis	"amazing"	institution
P05	background, debugging	"skeptical", "cautious"	institution
P06	code, statistical analysis	"interesting"	personal
P07	proofread, code	"narrow application", "unregulated"	personal
P08	literature review, visualization	"convenient"	institution
P09	none	"wary"	-
P10	literature search, domain specific	"skeptical"	institution, free
P11	code, information retrieval	"demotivating"	personal
P12	code, writing, citation	"pragmatically optimistic"	-
P13	literature search, editing	"helpful", "necessitates boundaries"	-
P14	transcription, data processing	"mixed"	-
P15	integrated search summaries	"concern"	-
P16	code	"optimistic", "useful"	personal
P17	literature review, brainstorming, administrative, code	"hopeful"	-

D Recruitment

D.1 Recruitment email

Hello,

Are you a scientist? Are you interested in responsible AI practices? We're looking for your input for a study from Ai2 and CMU on responsible practices and AI use in scientific fields.

We are recruiting interview participants in various scientific fields including:

applied sciences (agricultural science, computer science, environmental science, engineering, medicine)

natural sciences (biology, chemistry, physics, meteorology, geology, astronomy)

social sciences (psychology, economics, anthropology, sociology, political science)

If you are interested in being interviewed for this study, please fill out our intake form (<5 mins). Note: No previous experience with AI is required to participate.

Compensation

If you are selected, the interview should take about 60 minutes and you will be compensated for your time through Tremendous for $50.

Thank you for helping us promote responsible AI in science!

D.2 Intake Form

Thank you for your interest in participation!

We are a team of researchers from Ai2 and CMU conducting a study on responsible practices in scientific AI use. We are recruiting interview participants in various scientific fields. We'll follow up if there's a fit.

If selected for an interview, it will take about 60 minutes. You will be compensated $50 from Tremendous.

Note: No previous experience with AI is required for this study.

First Name

Last Name

Email

What best describes the gender you identify as?

- Nonbinary
- Female
- Male
- Prefer not to say

What field do you currently work in?

- Agricultural Science
- Anthropology
- Astronomy
- Biology
- Chemistry
- Computer Science
- Economics
- Engineering
- Environmental Science
- Geology
- Medicine
- Meteorology
- Political Science
- Psychology
- Physics
- Sociology
- Other

If you answered "Other" to the previous question, how would you define your field?

What is you current position title?

How many total years of research experience do you have? This may include any form of active research participation, including time as a student.

What institution or company are you currently affiliated with?

What is the highest degree you have attained?

- Bachelor's degree
- Master's degree
- PhD
- Other professional degree (please use open response field below)
- Other (please use open response field below)

If you answered "Other" to the previous question, how would you describe your degree?

At what institution did you attain the above degree?

What country do you currently reside in?

What are some types of research tasks you perform on a daily basis?

D.3 Recruitment process

In total, we received 112 responses through the intake form, 60 of which were eligible for participation, and 31 of which were invited to complete interviews. Table D.3 shows the recruitment pipeline by domain.

Table 3 The recruitment pipeline

		Intake Form	Eligible	Reach Out	Interviewed
Applied Sciences		**21**	**8**	**8**	**3**
	Agricultural Science	1	0	0	0
	Computer Science	10	2	2	1
	Environmental Science	0	0	0	0
	Engineering	3	1	1	1
	Medicine	7	5	5	1
Natural Sciences		**34**	**22**	**15**	**9**
	Biology	14	6	5	4
	Chemistry	5	3	3	2
	Physics	9	7	2	2
	Meteorology	0	0	0	0
	Geology	0	0	0	0
	Astronomy	6	6	5	1
Social Sciences		**54**	**30**	**9**	**6**
	Psychology	3	2	1	1
	Economics	2	0	0	0
	Anthropology	4	1	1	1
	Sociology	8	6	4	2
	Political Science	37	21	3	3

E Interview Script

Thank you so much for participating in this interview! I will be taking notes and we will be recording this interview on google meet so please feel free to turn your camera off now if you would not like your face to be recorded. This recording will only be accessed by the research team for transcription. This interview should take no more than 60 minutes. For your own privacy, please avoid bringing up any personally identifiable information during the interview. In this interview, we are going to discuss your research work and opinions on scientific AI tools.

All the questions we will ask do not have a definitive right or wrong answer. Instead, we encourage you to describe, based on your own experience, how you conduct research and use AI tools in the process.

Please don't hesitate to interrupt me and ask for clarifications. Most importantly if there is a question you do not want to answer at any point let me know and we can skip it.

Now, before we get started, do you have any questions? If everything is okay, we will begin recording now.

E.1 Research field and practices

We will first begin by getting to know your field and research process. Could you briefly describe your field of study and what areas you were trained or work in?

What research topics do you focus on now?

Can you describe your current research process from ideation to publication and beyond? What steps does a project go through? For instance, could you describe your most recent project and describe what steps it went through?

E.2 Process of responsible research

We would also like to learn more about the norms around responsible research in your field. What safety research practices are in your field? What are some things that are not acceptable to do in research in your community?

What ethical research practices are in your field?

What external guidelines help shape these?

E.3 AI experiences

How would you describe your general sentiment about AI?

How much do you personally pay for AI tools to use in your research and where does it come from?

E.4 AI use in research

Now, we will ask you a few questions regarding AI use in your field.

In this interview, we are mainly interested in two types of AI systems: generative AI and scientific AI agents. Generative AI refers to an AI system that generates new text, images, videos, etc. given an input. AI agents are a subset of generative AI systems that are equipped with abilities to use external tools and take actions such as running code, searching on the web, or retrieving documents from a database. They can act as autonomous agents capable of making decisions. Examples of generative AI systems may include but are not limited to ChatGPT, Gemini, Claude, etc. Examples of AI agents may include Codex (which can autonomously write and execute code), or a literature review agent that can search a database and summarize relevant papers. A fully automated AI agent for a scientific research step will be termed "AI scientist". These scientific AI agents might not exist yet completely, but they are in active development and might be present in the near future (3-5 years). Some AI agents might have access to tools with a physical existence in the real world, such as an automated pipette machine or programmable robotic arm. These AI agents may also include highly domain-specific systems, for example a protein-folding AI model or a weather prediction AI model.

What questions do you have about these definitions?

E.5 Gen AI / AI agent use experience

What types of generative AI tools or AI agents do you currently use in or for your research?

(optional) A lot of existing tools could also use generative AI in the background. For example, have you ever used AI summaries while reading a paper?

How are AI agents currently used by others in your field?

If you use generative AI or AI agents, what influenced your choice of which to use?

What qualities do you wish the generative AI or AI agent system that you use had, though you can use the AI system anyways?

E.6 Negative Consequences

Next we will ask about some possible negative consequences of using AI agents. For these questions, we are asking specifically about AI systems that have access to external tools or methods and are not exclusively generating text.

We are interested in three broad categories of negative consequences. The first is internal AI agent failure. This category includes scenarios where the user of an AI agent has a benign intent that is within the intended scope of the AI agent, and there is a failure entirely attributable to the AI agent where it did not act as intended. It may include negative consequences resulting from a single interaction, but may also include an AI agent's systemic impact on the discipline.

The second category is user error. This category covers all scenarios where a user with good intentions uses an AI agent out of scope or with user error, which results in the AI agent acting as the input dictates but not as the user intended, which has negative consequences.

The third category is malicious misuse, where a user provides a malicious input designed to circumvent the AI agent's safeguards, and the AI agent fails to refuse the malicious query.

Do you have any questions about this distinction?

E.6.1 Failure

In your view, what would constitute an appropriate / ethical AI practice in your field?

What unintentional negative consequences could you envision resulting from AI safety failures amid these appropriate / ethical uses?

E.6.2 User Error

What would constitute incorrect AI practices in your field?

What would be the consequences of those incorrect AI practices?

What could be a negative consequence if an AI system with extensive knowledge and capability in your field were also available to the public?

E.6.3 Malicious Misuse

What would constitute unethical AI practices in your field?

What would be the consequences of those unethical practices?

E.7 AI agent automation risks and ethics

Let's assume that an AI scientist has been deployed in your field that is capable of automating steps in your research process. This is an AI agent designed to do scientific research tasks fully autonomously, without supervision of a researcher. Now let's go through the steps of your research process.

If you were to use an AI agent to automate one or more of the steps in your research process which would be most beneficial?

For those steps, if they were done by scientific AI agents, what could constitute an AI system failure? Please provide concrete examples.

User error?

Malicious misuse?

E.8 Safety checklist

If you were to create a checklist for assessing safety of AI agents for this research step, what would you include?

If you were to create a checklist for assessing ethics of AI agents for this research step, what would you include?

E.9 Safety Implementation

Where or how should AI agent developers consult researchers in your field when developing and evaluating scientific AI agents and their safety? What definitions should individual scientific fields have control over? How should researchers in your field be involved in development and evaluation of scientific AI agents and their ethical guidelines?

For this question, we are defining AI guardrails specifically as safeguards or limitations built into the AI agent by the developers who built it. How do you think AI guardrails around AI agents should be implemented in your field? What is the responsibility of AI agent developers?

For this next question, we define human-AI use guidelines as best use practices or regulatory guidelines that govern how researchers use an AI agent. What do you think should be included in human AI use guidelines (for example for publishing) around AI agents? What is the responsibility of researchers using the AI agents?

E.10 AI future agents

Now we will ask you about your predictions about the future of your field and AI.

How do you predict that AI agents will be used in your field in 5-10 years?

How do you think your roles would change if these AI agents were used?

What concerns do you have about these uses and their possible negative consequences?

E.11 Summary

Of the things we discussed today, is there one that stands out to you as the most pressing or important?

E.12 Demographics

Finally, we would like to collect some demographic information. These questions are completely optional, and based entirely on your self-identification.

What age bracket do you fall into? 18 - 25, 26-35, 36-45, 46 - 55, 56 - 70, 70+

How would you describe the ethnicity you identify as?

What was your first language?

Thank you so much for your time, please let us know if you have any other questions/follow ups!

F Codes and Themes

We report here our identified codes and themes. See Table 4 for all the codes, themes, definitions, and example quotes. Additionally, see Table 5 for organizational codes, which were used to organize participant responses to the theme of the questions asked.

- **research-practices**: Theme that concerns current research practices, community norms whether implicit or explicit, and actions take by scientists today. It may contain references to how AI use impacts these practices, but covers research practices more broadly
- **tasks** Theme that captures application areas either for non-use or AI-use
- **AI-use-practice** Theme that captures codes about current AI use practices and how they are shaped and influenced
- **ai-information-source** Theme that captures information sources that participants use to inform themselves or to make decisions about AI
- **scientific-trust** Theme that captures either practices towards scientific trust or values that participants hold about scientific trust
- **ai-impact-scientist** Theme that captures the impact AI use has on scientists, rather than scientific work.

- **benign-user-failures** Theme that captures somewhat both AI failure and user error; we combined these two because they are tightly coupled (e.g., hard-to-detect-errors can be both a failure of an AI system or user validation error, failure that is caught by user might no longer be a failure)
- **misuse** Theme that captures malicious or unethical use of AI
- **negative-impact** Theme that captures negative impacts of AI that are broader that impacts on scientists / scientific trust
- **responsible-implementation** Theme that concerns active or proposed methods for implementing responsible research practices around AI use
- **scientist-value** Theme that encompasses both the things that scientists value about scientific work, and the unique value a human scientist brings to their work
- **benefits-ai** Theme that captures benefits of AI usage
- **future-role** Theme that captures some of the possible future roles of researchers – note that these can be expressed as roles or tasks

Table 4 Codebook

Code	Theme	Definition	Example quote	Participant
coding	tasks	AI use application; specifically for either code generation (with full access to compute) or code edits, etc.	So, one person that I know would write their own code, and if they had some change that they had to make to a certain segment of their code, they would copy and paste that code into a generative AI and ask it to make the edits for them to save time, and specifically wanted to avoid using an AI agent to access the code directly.	P05
lit-reivew	tasks	AI use application; specifically for reviewing literature, including search, summary, etc. but there is a distinction between search enabled vs. summary as noted by one participant	I have had some experience using those agents to do literature summaries, but I have already then curated that literature on the front end, meaning I've already done the selection of the kinds of materials, or the filtering of, "These are a relevant series of 10 articles on this topic or in the latest research in this area. Now give me a summary of these 10."	P03
writing	tasks	AI use application; writing either draft or editing	We actually did a meeting where we took a brief informal survey amongst the 10, 15 people in our little group, um, and like it was 90% people are using it to code and like 9% using it to help write.	P01
visualization	tasks	AI use application; creating graphs	Something like editing code seems fine. Or accessing data files on a computer to generate a graph or a PowerPoint slide, something like that. This is in the realm of science, I guess, cuz I'm biased, but	P05

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
presentation	tasks	AI use application; Creating presentation such as powerpoint	people were producing PowerPoint slides with generative AI. I think Claude can do images, so maybe giving it, um, a graph that the scientist had produced, and some information, and, "Make me a slide using these colors and this graph to show this point."	P05
ideation	tasks	AI use application; Iterating over possible ideas and creating research questions or strategies	I cannot use AI to generate an idea. That's that's that's my clear-cut boundary that I Okay. do not use AI to generate ideas.	P04
information-seeking	tasks	AI use built on asking questions and seeking factual information	Well, occasionally I do a quasi-Google search through ChatGPT, so like, "Oh, tell me the basics of some mathy thing," for example, and I sort of read it with some grain of salt.	P01
repetitive-task	tasks	Tasks that are repetitive	Yeah. You know, one of these mundane things that we really don't like to do, I think we need more of that than anything else.	P02
lecture-prep	tasks	AI is used to prepare for lectures or teaching materials	And they've been more useful in translating research things to teaching materials or lecture materials, right?	P03
data-processing	tasks	Data processing application of AI; specifically to process larger amount of data	as we discussed, my area of study, our limitation at the moment, as I again, as I alluded to, is like data availability and machine-readable data availability. And there are mountains more of data out there, and so I think AI agents being able to help us prepare that data will really expand our like empirical model development, which will be a big transformation in our field.	P16
data-analysis	tasks	Participant discusses use of AI for data analysis	Oh, I know that they've also used it to analyze data sometimes, well like using like a pilot like Claude Science.	P07
ai-detection	tasks	Participant discusses use of AI for AI detection or broadly detecting AI in general	I'm also using Pangram just to see what's AI.	P01
reviewing-articles	tasks	Participant discusses using AI (or not) to review their or others papers	Because I know also that some people are responding to their review comments, or I imagine that they are, with LLMs, and it's like what even what's the point of all of that then, really? I don't know.	P16

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
cross-domain-information	tasks	Discussion of AI as an information dissemination tool	I think it will be helpful for um connecting ideas from disparate fields in ways that we wouldn't we don't have the time to do, haven't historically had the time to do	P13
physical-lab-tasks	task	Participant discusses using AI tools to automate physical research or wet lab processes	If I could put a camera on my shoulder that watches exactly what I am doing and transfer it in a document that I can give to my next generation who has to repeat what I would be doing was extremely helpful.	P02
mentoring-feedback	task	Participant discusses some application of AI tools in the mentoring or teaching of students	yes, I can direct it, but I would never want to ask my student to like, "Hey, you know, just sit in front of this computer and have your one-on-one meeting with it as opposed to me, [laughter] and have it provide you feedback and guidance on what you want to do." I think that's where it feels unethical to me.	P13
transcription	task	Participant discusses AI tools for transcription processes or recording	AI running transcriptions of qualitative data is going to be a huge like it will elevate the possibilities of qualitative work, um, to make it more affordable, etc.	P14
research-plan	task	Participant discusses using AI tools to form a research plan or timeline	I follow a kind of a workflow while I'm using it. It's mostly about you first define your goal. Um then so that skill set has like various levels, so the first level is like goal. Then you define a goal or aim, what's your aim. Then the second is like, "Okay, do a extensive research, verify everything." Then go in a plan mode, and then you plan, and then once it plans, the user verifies it, and then the user verifies it, and once it it likes the plan So it's not that AI is uh So when I say that I plan, it means I give my own inputs of the plan, and then the AI reviews that plan basically with all the different uh verifications passes, right?	P04

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
uniquely-human	scientist-value	The idea that research/science has uniquely human aspects that cannot be replaced	if a human can't make sense of the solution, like scien- to me, like science is inherently a human enterprise. It is like a collective body of knowledge that is evaluated and co-produced. It's like inherently human. So like I think that having a scientific result requires like a human community understanding of it, and that's something that, like, basically I I would say that you can't have just AI agent science that is not understood by human scientists.	P12
expert-intuition-nuance	scientist-value	Discussion of research taste; priorities; implicit learned standards	I have to say I have 40 years of lab experience. So, for me, it's very easy to see uh when something makes sense and other things is usually just a hallucination.	P02
learning	scientist-value	Participant discusses learning as a value that they prioritize or a reason for non-use; slightly more positive value than deskilling but could be combined	But the joy of I think the joy for for me was always I like to learn. And human beings want to have challenges, they want to to learn. It's their social fabric. They just don't want to party, they they need something.	P02
reproducibility	scientific-trust	Participant discusses reproducibility of the scientific work either in context of AI- or human-generated works	So reproducibility, keeping track of your code, it's I don't know, checking making Data sharing has become a thing over the last like that's now like required often by journals.	P01
validation	scientific-trust	Validation practices mentioned by participants	Okay. That's on the human to take responsibility and oversight. That comes down to validation in the way At least the way I've been teaching this, the way I've been thinking about it is at the end of the day, that that that agent is acting just like a student would in, say, a chemistry lab or a science lab. And if what has come out the other side of that is an error, something was synthesized wrong, that experiment didn't come out right, or something else, then then the the the agent needs to be corrected, just as the student would be, right? And you would then have to go back and say, "Okay, there's a cost to this. We either have to restart the experiment, or we have to refine the instruction um toward that."	P05

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
disclosure	scientific-trust	State how much AI was used; Participant discusses either disclosure policies or disclosure practices	And uh for me, I think if you are using AI to generate anything, you should also disclose this.	P08
transparency	scientific-trust	Researchers discuss the need for transparency	"Um, for me, it's mainly been these issues about about responsibility, transparency, and and what that looks like. Um, I think we're still in this thing where some people are, whether they're students, whether they're researchers, whether they're scholars, whether they're businesspeople, where they're using these tools, and they're not telling people they're using them. "	P03
misattribution	scientific-trust	Concerns covering plagiarism, lack of attribution for ideas or data	maybe the point that uh need to be discussed is that um definitely then like intellectual contribution, like what's the human's unique intellectual contributions, and also who will take really the accountability if that the result is very like bad, or misleading, or somehow it can lead to that the plagiarism, because I feel like I don't I don't want to say that we need a penalty or punishment, but we need more hard guideline guideline than just like current level	P11
use-pre-existing-guidelines	responsible-implementation	Participant states the use of pre-existing guidelines as checklists	And so I guess like if you're designing something that's going to an agent that's going to do like mass spectrometry analysis, talking to the people who develop the kind of current mass spec analytic software, and see what things they've already figured out and what sort of like checks need to be there, and then talking to researchers that do mass spec.	P07
unittest	responsible-implementation	Participant states the use of unittests to verify AI outputs in a more manageable and verifiable chunks	I imagine it would be good to do like simple unit tests. Like, here are five papers and they're 20 pages long, and we understand what's said in them and we know everything there is about these 20 papers, and let's like see how well such-and-such system, version one, version two	P01

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
distrust-checklist	responsible-implementation	Participant elaborates on limits of checklists for evaluation	I realized the checklists break down very quickly.	P03
enforcement	responsible-implementation	Participant discusses enforcement of AI policies	perhaps like there's a growing gap right now between what should happen and the capability of enforcing what should happen.	P14
scientist-developer-partnership	responsible-implementation	Participant discusses the importance of including scientists in the development of AI tools; Participant notes researcher and AI developer collaboration to build safe scientific AI either as consulting them or collaborating on projects	I mean, I think that for for my research that domain experts need to be involved almost it all the way through in in agent development in the decision-making. It's been amazing how the generalized models have been able to extract information and like um produce like reproduce the knowledge that's available kind of at the basic level, but for for these decision-making workflows, I think scientists and researchers have to be working together with the engineers.	P16
deployer-responsibility	responsible-implementation	Participant discusses the responsibility of the deployer, e.g., those that market and deploy AI as a product such as Meta, OpenAI, etc.	And so that person can develop that and say, "I have this thing that treats cancer." But the actual dosing of it, the pricing of it, the sale of it, the dissemination of it, that's going to happen by a pharmaceutical company and a pharmaceutical sales rep and someone who trains the oncologist and the doctor who prescribe that.	P03
community-norms	research-practices	Implicit norms of conducting research within the community	Yeah, there are norms. I guess they're rarely like written down. Um, and by definition, most people are in the norm, and occasionally you get people on that edges.	P01

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
collaboration-norms	research-practices	Participant discusses norms of collaboration	You do that. I do this." And for the most part, people are good faith actors, but occasionally people do things that again, not legally wrong, ethically wrong, yeah, I think debatably people could write uh could act in unethical ways like, you know, stabbing people in the back or like withholding information from a friend/collaborator that they probably should know, but like for whatever reason the person withheld it. Like you hear stories about that type of stuff. People flaking out, but again, that's not legally wrong. I don't even know if it's ethically wrong.	P01
publication-norms	research-practices	Participant discusses norms around publications, including journal guidelines, etc.	when I review papers in my field, um, some journals have required these types of disclosures now, but not all of them, and compliance is sort of a patchwork at the moment.	P12
human-participant-practices	research-practices	Responsible and ethical research practices to protect human participants and their data; this might include IRB	for the analysis one, I think it's really, you know, hard to say to someone else, and just like I can just say because they're my friends, that, "Are you really comfortable just like sharing that uh interview data, and is it really even like allowed thing for the IRB approval document?" I'm pretty sure that person would not say, "For the analysis, I will feed all that interview data to the AI," but that's just I'm kind of sharing my concern, but it's just also up to that person's decision,	P11
equipment-use-practices	research-practices	Responsible and ethical research practices around equipment use	Like MRI is has some dangers, of course, at like metal and stuff	P01
instutitonal-guidelines	research-practices	Participant discusses institutional guidelines; either current or needing one for AI use	Responsible Research Conduct trainings? We actually have one for AI now that they have us do every year	P10

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
history	research-practices	Participant references the history of a scientific field	a lot of this has come from bad practice in the past, but certainly today, you know, a recognition of, one, is this data really useful to the community? How can we give back in a way so that, you know, balancing uh questions that they are of scientific relevance and generalizability, but also that they are um of value to participants	P17
lab-practice	research-practices	Responsible research processes that are shaped by the lab	But I work with very hazardous chemicals. So, uh it this is more lab practice, I would say, how how you design a procedure.	P02
research-impact	research-practices	Participant discusses impact of their research or research in general	I think that applies, you know, broadly in in the political science as a discipline, and then even specifically within the areas I'm working in, because the things we look at involve things that have humanitarian angles to them. Where are refugees predicted to be? When are Why are they predicted to be there? But then also where and when those people are has implications for, say, targeting of opposition or supporting of other groups. So there's a real sensitivity there to how our work is used and what the ethical um considerations of what responsible use of that looks like of, you know, in an applied sense.	P03

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
correct-attribution	research-practices	correctly attributing where the data has come from and communities that have contributed to the data	I think of like genomic data that may be from protected populations. I think of uh news reports, human rights reports that may come from either a paid source, because they're behind a paywall and we need them to train something, I at least want to document the features and characteristics of those for people. And so I've tried to train my team in those run of how are we sensitive of other people's investment in intellectual property? And that could be uh source material that goes into training an AI model, it could be labels that some group of people have actually spent a lot of time to curate and put together that we then need to validate those models. In other words, those people deserve credit for having put in the time to do the labelings, and we should cite them or give them credit, or all the other things, because I think there's kind of a a vast group of people that have provided a lot of data to go into these models that that need to be respected for what they've done and our ability to validate them.	P03
replacement	negative-impact / future-role	Fear of replacement due to automation; simply discussion of replacement	We will have much less chemists. We'll have much more automation, and yeah, it's a shrinking of workforce, especially industry. They have the money to implement these sort of things. It's everywhere in the stack.	P02
resource-use	negative-impact	Agent output leads to unproductive resource use, especially funding, lab equipment, researcher time, etc.	I would be worried about waste of resources. Every experiment is costly in some way, and especially in the bio and cell culture space, it each experiment takes a long time to conduct, and even more so in mouse work.	P05

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
environment	negative-impact	Impact on environment	I actually looked into if I train a machine learning model on this hardware, on this power grid, for this long, how much energy does that use? And um at least in 2024, that was like a super difficult question to answer. Um, and I think that that's I think that that's unethical, because now we have people who are going to use very energy-consuming tools sort of willy-nilly, where maybe if they knew here's what your energy usage was, they might be a little bit less um might be a little bit more conservative in their usage.	P15
physical-harm	negative-impact	Agent output leads to physical harm	And the consequences of AI failure there, if it were used in practice, I think could, you know, could be catastrophic, could be like a bridge failure or a dam dam failure or something.	P16
political-unrest	negative-impact	Participant discusses a possible negative impact of AI causing political unrest	"And that that's It's It's one that that I have I have thought a lot about, um given that, you know, we are developing AI models that are oriented towards detecting and finding features of political violence, political conflict reports, human rights reports, and the rest, is, well, where and how are then those, if those are understood in a particular way, are you making it harder or easier for people to engage in political protest, repression, filtering of news and reports in another society?	P03

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
ai-disparity	negative-impact	AI driven disparity; institutions that can afford better models and people who can afford better models will excel but those who cannot will not be able to	the first inequality is like people that are using it versus people who are not. But the second inequality is um universities are purchasing AI according to whatever budgets they can afford. And so rich universities can get great AI, and poor universities will have none or bad agents. ...And if what you get is what you pay for, it's going to create university inequalities that are really messed up. ...there's no mechanism to enforce equities, like to enforce equity, then um like I think that rich universities and scholars that have huge budgets are going to be able to do things that other scholars can't. And I I think that the dynamism of higher education, and therefore also like the United States productivity and innovation, all that, it cannot be concentrated into Ivy Leagues or a smaller and smaller set of universities. It's just like the wrong way to go for society.	P14
shallow-ideas	negative-impact	Related to the overwhelm-system and uniquely-human codes, participant discusses the depth/quality of research ideas generated by AI	I hope that like the actual, I think like basically scientific questions and more more importantly like the directions that are the the scientific directions that we pursue, I hope are not outsourced or we're relying on agents for that. I think that, um, to me, that is pointing towards just like a monoculture of ideas and is potentially problematic	P12
future-complexity	negative-impact	Participant discusses the ways that using AI now will bring temporary benefits but make a downstream task (ie technical debt) harder	serious consequences to those when we as a as a society get used to that way of doing things. If like we want code to be short, readable, and easy to understand and maintain, because that means that like someone else in the future can fix it if there's a problem. And so if we are in the period of time where people are generating tons of technical debt, that's like a big problem for like not even future generation, but just like for ourselves in 10 years as we try to like untangle this Gordian knot.	P15

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
intrusion	negative-impact	Participant describes AI tools as feeling intrusive	if you open up Zoom, there is this ""Work happy with AI in Zoom Workplace"" thing. Um, I can't get rid of this, and every time I open up Zoom, my eye is drawn to the right side of the screen with the pretty bright colors, and I and I want to be like, ""Never show me this again. Like, I don't want this. Zoom is a is a phone call to me. Like, I don't want you to do work for me.	P15
data-manipulation	misuse	Discussion of data manipulation usually from malicious intent	""I'm going to write a paper that looks really plausible that I think the science might be sketchy on, but if I can use Claude to make me sound really smart, it's going to make it harder for people to realize that I don't know what I'm talking about."" Um, and I do think that there is like there are actual incentive structures to doing this.	P15
overwhelm-system	misuse	Discussion of AI tools making low-effort or low-contribution work easier, concerns about reviewer or conference load	Like, it's already a very much like push-to-publish sort of scenario, so I think that if uh AI is, in theory, making it more accessible for people to get out more publications quicker, um then I have concerns about, oh, if reviewers will use AI, and how I think the quality of the research might decrease and the quantity will increase	P07
dual-use	misuse	Participant discusses dual use of AI system	If you have this all-knowing model that can, you know, tell everybody the best way to get your rumors out there, in the most efficient way possible, right? Or here's this piece of misinformation, we're going to make sure this sees as many eyeballs as possible in the lead-up to an election or, you know, in the lead-up to social unrest or something like that.	P06
offloading-responsibilty	misuse	Participant discusses offloading of things that should be done by humans; this is usually discussed in regards to responsibilities such as thinking, teaching, writing, etc.	that the accountability issue, if something really happened, then it's very shallow, but who would take the responsibility in what with what way	P11

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
misinformation	misuse	Participant mentions an increase in scientific misinformation – unlike data-manipulation, this code covers more widely accessible information/public impacts and generally refers to text	Ooh. To some extent, um, I worry, right? If if you have this all-knowing model that can uh, you know, um tell everybody the best way to get your rumors out there, um, in the most efficient way possible, right? Or here's this piece of misinformation, we're going to make sure this sees as many eyeballs as possible in the lead-up to an election or um, you know, in the lead-up to social unrest or something like that.	P06
mangerial	future-role	Future role of researcher as a supervisor, idea of AI as a tool	I think it would become more of a managerial overseer of AI agents type role, and ideally understanding all of the steps that AI would replace within that scientific process, so you are going through each of the steps with the AI, just kind of watching it, holding its hand somehow.	P05
bias	benign-user-failures	Participant discusses bias of the model or user that might influence the systems' outputs in unintended or intended ways	I think it raises these issues about what sort of biases that agent might learn. I think it would be hard to keep that agent ethical.	P03
hallucination	benign-user-failures	Participant discusses AI hallucinations	I guess like the most obvious failure, which I would hope wouldn't happen for a literature search AI, but is kind of like falsifying information, or like having citations that don't actually exist, which I know has happened with some AIs.	P07
sycophancy	benign-user-failures	Bias induced in an answer as a result of the opinion of the user, a model's tendency to agree	prompts that lead that are like leading to what you want to hear, leading to validating or showing you only papers or only data that is aligned with your hypothesis versus the true scope	P10
uncertainty-expression	benign-user-failures	Discussion of overconfidence, lack of uncertainty expression	Okay. I think for my personal experiences, I feel the AI, with at least the AI model I am using, is very easily to get persuaded. So if I I got some question on the answer of the AI, and I think the AI model always not cert- become uncertain by itself.	P08

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
security	benign-user-failures	Computer security	There could be, if it's scraping a website, it may inadvertently scrape, like, malicious code that is included in the website, right? And and potentially execute that, you know, without, quote, unquote, thinking.	P06
data-privacy	benign-user-failures	Participant discusses concern over data privacy	people have spoken about even putting your data into these AI models that they will take it, and like if someone else searches up this question, would they be able to like share your data that's not published yet? So that could lead to like scooping. And even though that's like purely academic, uh could like ruin people's careers.	P10
out-of-domain	benign-user-failures	The task is out-of-domain for AI usage such as countries or specific task is not what the AI is trained on	But if what we need is something more out of the box or something more heterogeneous or something addressing the extreme cases, it's going to work against us to do that.	P03
ai-blocked	benign-user-failures	Participant discusses limited access for AI agents on some websites, especially with regards to doing lit reviews	I found this title of this article, which seems to be aligned with what you're looking for, but I can't access the main text. So, you know, clarify this is the the DOI for it, you know, just double check it beforehand." Um So it's hard to say whether that's a failure or not because they don't have because it doesn't have access. But at the same time, it's like at what point like where is the line for when like full access is um like the basis of of of of what we're expecting from it?	P13
hard-to-detect-errors	benign-user-failures	Participant mentions that some errors are difficult to detect; AI generates something that is highly plausible where validation is difficult	do I have to be able to understand every step in the code and be directing it? Like that is a space where I feel like if if AI does something that's wrong, I wouldn't be able to catch it	P14
user-responsibility	benign-user-failures	Participant sees AI more as a tool; Putting responsibility on the users to check the outputs	the user needs to be particularly attuned to like failure modes and specific contexts where, like, anticipating where things could go wrong I think is is part of responsible use here	P12

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
false-presupposition	benign-user-failures	A user's prompt contains a false assumption about their query	Like I have a some idea that that Earth is flat, and I keep pushing, "Earth is flat. Earth is flat." Then maybe over time, it will generate its own citation of paper that the Earth is flat. So that's something that uh I mean, I do not see it happening a lot, but I see that that is one of the possibility that can happen where user can push its own bias uh while interacting with the AI agent.	P04
benefits	benefits-ai	Researcher discussed positive impact of AI tools, including speed up, more discoveries, etc	So sure, I mean for large data and testing and modeling and all this, it's a big improvement.	P02
speed-up	benefits-ai	One key benefit of AI use; speeds up discovery and daily work of researchers	And it seems to speed up, but where it goes... I wish we could put on the brakes and think about it.	P02
equity	benefits-ai	AI benefits of bringing more equity towards science	Look, this is also about equity. In Ivy League schools, professors get like um copy editors and people that can help them with writing. And now effectively these copy editors and people that can help you with writing are available to everybody all the time. And and then and therefore it's sort of like an equity issue. And so if AI is helping you to write something like, absolutely, that's okay.	P14
large-data	benefits-ai	Participant discusses benefits of AI in terms of processing large data and making possible what hasn't been able to be done before	If um drug trials have combined um these different compounds before, but they haven't tried compounds C and D together before, like maybe that is a space where um, you know, something with more, you know, bandwidth than the human brain, so to speak, um would be able to to, you know, take in um that large of a corpus of information and distill it.	P06
human-limitation	benefits-ai	Participant discusses limitations of humans	Cuz we don't have the mental cap-uh capacity to include all these parameter in our thought process, lots of things. Yeah, we have some instincts how things work, which is sometimes very hard.	P02

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
reinvest-human-resource		Participant discusses reinvesting human resources such as time to something more meaningful to them	so that I can do I can do better planning, I can use my energy to to to um to come up with right set of questions, and uh which which helps uh people broadly in a in a more general sense than me sticking with my computer and trying to say align my X-rays to a sample, right?	P04
domain-specific	AI-use-practice	AI tools designed for a specific scientific domain and purpose	people in my field use a lot of AlphaFold. That was like super revolutionary for like our entire field, and it's been really awesome to work with, especially as it has its new updates with putting in like ligand specificity, RNA specificity. Like it's been really cool.	P10
open-source	AI-use-practice	Open-source AI tools	For the most part, I've stayed in the space of trying to work with and deploy openly available models, either open open-source, open-weight, or ones that we can run and operate within the resources we have available within the university system.	P03
regular-commercial	AI-use-practice	Commercial, closed models, paid or free	part of it is pricing. Like, I have access to, um, like the Claude Max plan. Um, so that is a lot of compute for, I guess, a reasonable expense.	P12
integration	AI-use-practice	AI tools that are integrated into another platform, ie paper summaries or google search	probably the only thing that I use for writing code is if I um can't remember how to do something in Python, I ask Google. And then Google's um, it's actually it's pretty good. I think it's it's um based on AI. It it'll tell me how to write things.	P09
ease-of-access	AI-use-practice	Participants' AI use choice based on ease of access; including, familiarity, institutional adoption, integration into familiar systems e.g., google search	It was like the first thing—and like I just It's just instinct, so, "Okay, chatgpt.com, whatever."	P01
institutional-preference	AI-use-practice	institutional adoption / preference of a certain model leading to AI use or choice of which AI to use	then I know that like sometimes there's different connections for like I think uh my institution like prefers Claude, so things.	P07

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
tech-stickiness	AI-use-practice	Participant discusses that they made a choice of which system and have developed practices around that vendor so there is a change cost or they simply have inertia to keep using it	It was like the first thing—and like I just It's just instinct, so, "Okay, chatgpt.com, whatever."	P01
ai-non-use	AI-use-practice	How people are choosing to not use AI in research	the tool who can analyze the data and who can interpret the data in their way with their perspectives, but if you just let AI do, that's not your analysis, that's that's AI's and analysis, and I know that AI is now super intelligent, but I feel very uncomfortable on that	P11
full-adoption	AI-use-practice	Full adoption of AI; giving it full access	So there are some colleagues who full throat who are full-throated activists and advocates of it, who like it will they will ask it to create a specific aim page for a grant, pulling literature that it asks them to run. Um And they have automations that they do everyday. They um ask it to synthesize their email at the very beginning of the day. So I have some colleagues who use it in like in all of the ways, and then I have other colleagues who don't use it at all.	P13
sandboxed	AI-use-practice	Limited, sand-boxed adoption of AI	So no like writing to other stuff out over the internet. Definitely no internet connectivity.	P09
variation	AI-use-practice	Participant notes variation in the adoption and use of AI	I can speak mostly like to my department, um, but I think that, um, there's a huge spectrum. Some people completely, uh, resist the use in any capacity. Um, others use it like embrace it I guess as fully as I described that I do. And there's like a I think a lot of people in between.	P12
colleagues	AI-use-practice	Pressure from colleagues or recommendations from colleagues	I talked to my my friends, mathematician and people in computer science what for what I want to do and uh that's what they recommended.	P02

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
model-test	AI-use-practice	Adoption choice based on a certain test by the participant for either accuracy, feel, expected answers, etc.	my use of Claude is mainly because I maybe at an earlier point tested a couple and felt like the Claude output was significantly better. And also, out of like, gosh, I don't know about the ethics of these of companies, but at least there is some sort of verbal nod to ethics by Anthropic, even if it's superficial and doesn't carry any weight.	P14
generational	AI-use-practice	Participant discusses something about AI use being generational or there being a generational divide	I have a lot of uh fellow scholars, and and I think um I've seen this as a generational split, where it's younger scholars are much more willing to adapt these either for assisting in code generation, assisting in research and literature development, even things like deploying surveys or survey experiments and devising what those are going to look like, and then even once the data would be collected from humans, what the subsequent analysis is going to look like.	P03
skepticism	AI-use-practice	Participant expresses skepticism about AI answers	Um, as far as using it, I just think culturally there should be, especially as scientists, we're used to being skeptical and questioning the truth and being able to really think about and and break down a system. And I think that should not stop.	P05
iterative	AI-use-practice	Participant mentions an iterative process of using AI; adjust inputs based on output, provide feedback, etc.	the way we use AI iteratively, um you don't you don't always know if you're giving it all the information or correct information, and you don't know that what is coming back, you know, you can get some here's what I've searched, you know, here's some articles, whatever, here are best practices, but because we are limited in the way that, you know, AI is better because it is pulling from a much bigger purview of of data in that regard.	P17
cat-out-of-bag	AI-use-practice	Captures sentiment that the AI progress is already here and cannot be undone	Oh, the cat is out of the bag.	P02

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
human-comparison	AI-use-practice	Participant explicitly compares the ai agent to a human in how they reference or use it	I don't think it's anything different than already exists in science right now. Um, we're always working with a sort of limited set of information and limited ability for people, with peer review with everything, to really check data.	P17
task-clarity	AI-use-practice	Discussion of needing task clarity; input to AI needs to be clear	And as I know, if you cannot properly uh formulate your input, how you say, trash in, trash out, that's how it works.	P02
ai-training	ai-information-source	Training around AI	as a starting point, thinking about ways to train universities and to create opportunities. Like that's not happening, and like I will also say, some of the AI trainings are stupid. [laughter] They're like They're not training people on what they really need to be learning AI for, because some of the people that are developing and making money out of these AI trainings are like, you know, in 2 years ago, and now they're like training people how to run like a good search with key terms, you know, search terms. And like that's that's not even the cutting edge. Like understanding like even what an agent is is its own debate, because there is not that kind of training.	P14
AI-media	ai-information-source	Researchers discuss outside perception of AI issues or reference a news story	I can't help but think about all these like, you know, math proofs that keep coming out with these like Lean verified solutions, and it's like if the math community has identified these as important problems, and then there's like a Lean proof of it, that means something	P12
distrust-companies	ai-information-source	Researchers express concern over large AI companies or technological experts	AI companies didn't care, and they're the ones with the power, right? They have all the money, they have all these data centers, and they forgot to put a Google Docs link into the data	P10

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
ai-literacy	ai-information-source	Participant discusses collaborators or the field not fully understanding AI tools	two-pronged approach of AI-proofing your teaching in your classes, making sure students are learning stuff independently, but then also like AI integration of like like basically teaching students to do it themselves and check the AI results.	P14
mental-health	ai-impact-scientist	Loss of purpose, concern for the future of science	then what is our purpose, right? It becomes very existential. You know, like we we are here to create and produce, shift, move society in a forward direction.	P13
community-loss	ai-impact-scientist	Researchers lose contact with scientific community	he was like, 'Oh, that was really helpful. Chat wasn't understanding.' I was like, 'Cuz it's like a human thing to like touch the tubes and understand the labeling system that you want that will be useful for you to have in your notebook and to have it in the freezer.' Um so that was like an interesting use cuz he loves ChatGPT. I was just like I'd just ask a fellow friend for those sorts of things.	P10
reputation-loss	ai-impact-scientist	External pressure to publish, fears of how others see AI usage, reputation loss from incorrect science	How do you start a conversation and ask like, ""Do you use it, and how do you use it?"" versus um not wanting to come off or be perceived as like I don't know, ""I'm killing the environment and, you know, causing the the destruction of society because I'm using it."" Because it feels like there's really no middle ground to some people.	P13

Continued on next page

Table 4 Codebook (continued)

Code	Theme	Definition	Example quote	Participant
become-outdated	ai-impact-scientist	Fear of becoming outdated or impact of becoming outdated	I'm sometimes intimidated that I'm so outdated, or am I using AI in a wrong way, so I'm not satisfied with their outcome, or I I should give up uh feeling uncomfortable from the AI's writing tone and just take this flow, but yeah, some people use very heavily, and some people are saying like, if you just let AI do everything, then what is the enjoyable part of research? Then I pretty agree on that, but as a, you know, PhD students, I'm feeling also want to make more productivity, so that I'm sometimes very chastened by like, you know, I need to publish more, but my group is so small, like no funding that much, so that should I do a little more actively using AI and to do a lot of work	P11
deskilling	ai-impact-scientist	Researchers lose skill and a chance to learn	I would hope that like we're still training scientists, that people are still wanting to be scientists, that that is like an opportunity and avenue that people get to choose to go down, and that those opportunities aren't taken away because of AI.	P10
difficulty-of-discussion	ai-impact-scientist	Participant mentions not being able to talk to colleagues about AI use, conflicts around the use, discussions/discord among collaborators. Distinct from reputational harm this is about the discussion itself and not future impacts	I would say I I don't think that our field has converged on norms in terms of what's appropriate and inappropriate. Um, I believe that there are some people in my field that think it is absolutely unethical, um, and inappropriate for people to be using AI for any sort of writing functions. And then there are other people that are letting AI sort of write entire papers. Um, and I don't I I don't know if that really I feel like maybe lost your question, just cuz I got hung up on that piece, but, um, it's hard, because I feel like we can't start from a premise of thinking that there's any sort of acceptable use norm.	P14

Table 5 Organizational Codes

Code	Theme	Definition
research-process	organizational-code	Participant specified research process
external-guidelines	organizational-code	Explicit norms of conducting research enforced by a written guideline
field-use	organizational-code	How others in the field are using AI
ai-use	organizational-code	How people are using AI in research
appropriate-use	organizational-code	Discussion about appropriate use
agent-failure	organizational-code	Participant discusses failure of AI system
agent-misuse	organizational-code	Participant discusses AI misuse; someone with a malicious intent either unethical researcher or public
user-error	organizational-code	Participant discusses user error—benign intent and incorrect input
negative-impact	organizational-code	Negative impact of AI outputs
checklist	organizational-code	Discussion of checklist
guardrails	organizational-code	Implementations that are the responsibility of the AI developer
developer-responsibility	organizational-code	Discussion of developer responsibilities
general-ai-sentiment	organizational-code	General AI sentiment discussion
future-researcher-role	organizational-code	Participant lists future changes that may occur in the role, which parts they expect to still be doing in ten years
beneficial-ai-automation-area	organizational-code	Step in research process that is potentially beneficial
example	organizational-code	Possible benchmark item
public-misuse	organizational-code	Discussion about public-misuse