IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 035007 https://doi.org/10.1088/2632-2153/adec3c

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
7 March 2025

REVISED
23 June 2025

ACCEPTED FOR PUBLICATION
4 July 2025

PUBLISHED
17 July 2025

PAPER

Considering the ethics of large machine learning models in the chemical sciences

Evan Walter Clark Spotte-Smith

Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States of America
Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States of America

E-mail: ewcspottesmith@cmu.edu

Keywords: ethics, foundation model, sustainability, generative AI, protein model, large language model, chemistry

Supplementary material for this article is available online

Abstract

Foundation models, including large language models, vision-language models, and similar large-scale machine learning tools, are quickly becoming ubiquitous in society and in the professional world. Chemical practitioners are not immune to the appeal of foundation models, nor are they immune to the many risks and harms that these models introduce. In this work, I present the first analysis of foundation models using the combined lens of scientific ethics and chemical professional ethics. I find that general-purpose generative foundation models are in many ways incompatible with the moral practice of chemistry, though there are fewer ethical problems with chemistry-specific foundation models. My discussion concludes with an examination of how the harm associated with foundation models can be minimized and further poses a set of serious lingering questions for chemical practitioners and scientific ethicists.

1. Introduction

In recent years, machine learning (ML) models have grown in complexity and scale (in terms of e.g. number of learnable parameters and training data size) at a rapid pace, culminating in the development of so-called 'foundation models' or 'large models'. These models, which are intended to either be directly applicable to problems in a wide domain space or to be easily generalizable through fine-tuning and transfer learning, include generative models like large language models (LLMs, e.g. GPT-3 [1], DeepSeek-r1) [2], image and video diffusion models (e.g. DALL-E [3], Stable Diffusion) [4], and vision-language models (VLMs, e.g. LLaVA) [5], among other, less common examples. Foundation models have infiltrated nearly every aspect of the digital world, from standalone applications like the popular chat interface ChatGPT to digital search, code generation, automated writing, and much more.

The chemical sciences have not been immune to the proliferation of foundation models. There have been a number of recent efforts seeking to evaluate the capabilities of generative models like LLMs for chemical tasks, from answering simple questions [6] and mining text [7] to performing complex experiments by interfacing with laboratory hardware [8, 9]. Efforts to create chemistry-specific foundation models have also appeared, including protein-folding models [10, 11] and 'universal' ML interatomic potentials (MLIPs) [12, 13].

In the recent rush towards ever larger ML models, there has been a growing chorus of concerns related to the ethics of these 'artificial intelligence' ('AI') models [14–16]. Less attention has been paid to the ethics of foundation models when applied in chemistry, though discussion in some other scientific domains such as medicine is more robust [17–21]. This work aims to fill that gap and initiate a conversation about responsible development in applied data science and scientific ML in and around chemistry.

I begin (section 2) by providing a theoretical and practical basis for ethical discussion. Drawing from the ideals of science as well as from several prominent ethical codes for chemists and chemical engineers, I discuss some basic principles that should guide the practice of chemical science (section 2). In section 3, I examine the compatibility of general-purpose generative foundation models (e.g. LLMs and VLMs) with

scientific and chemical ethics. With these more general considerations at hand, I shift focus in section 4 to examine ethical concerns that are unique or particularly relevant to applications of foundation models in the chemical sciences. Section 4 also considers the ethics of chemical foundation models. Finally (section 5), I provide a brief discussion and conclusion. I suggest how some of the risks and harms mentioned might be mitigated, using analogies from efforts in other areas of chemistry, and I emphasize the need for further research into chemical 'AI' ethics.

2. A basis for chemical ethics

2.1. The ideals of the (chemical) scientist

There have been many attempts to distill the ethical ideals of science and scientists. The different sets of ideals overlap considerably. In the interest of simplicity, I will focus on the six ideals described by the National Academies of Science, Engineering, and Medicine in their report *Fostering Integrity In Research* [22]:

1. *Objectivity*: the responsibility of a scientist to avoid letting their personal feelings or beliefs introduce bias into their findings
2. *Honesty*: a core responsibility which urges scientists to engage in truth-telling without significant omission
3. *Openness*: a scientist's responsibility to transparently present all relevant findings and details
4. *Accountability*: a scientist's obligation to be able to explain and justify their work, and to stand behind what they have done
5. *Fairness*: the obligation to consistently make judgments based on clear, equitable criteria
6. *Stewardship*: an overarching responsibility to maintain relationships within scientific organizations and between scientists and broader communities

For the purposes of this discussion, I will act on the assumption that these ideals are appropriate for (chemical) scientists and that they are generally worth aspiring to.

It is worth pointing out that the principles of the National Academies speak to the social element of science. While science is an epistemology that has been called a pursuit of 'reliable knowledge' [23], as a profession and practice, it is also a social activity. Essentially all practicing scientists 'know' or 'believe' ideas about the natural world on the basis of the theories, simulations, and experiments of other scientists [24]. In this way, scientific epistemology and scientific ethics cannot be disentangled; our knowledge is only as good as the scientists who we are reading, listening to, and interacting with. By upholding honesty, we seek to improve the reliability of our knowledge and that of those around us, while stewardship more directly preserves the social connections which enables science's decentralized knowledge generation.

In this discussion, I further assume that all chemical scientists should try to live and work according to these ideals. It is reasonable to challenge this assumption and ask, 'Why should they'? It would be insufficient to draw on legal or professional punishment. Laws are the ethics of the State, not of people, and fear of retribution, however valid a motivation, is not an ethical one. While the motivations for ethical behavior are outside of the scope of the present work, I should point out the work of Kovac [24, 25], who has invoked virtue ethics and specifically the virtue of *reverence* as a possible answer to this question.

2.2. Chemical ethics in practice

One would hope that most practicing scientists agree with the National Academies' principles and similar ideals of science, but most scientists have not explicitly agreed to live and work according to these ideals. Even if they had, ideals represent aspirations, and it is not reasonable to expect everyone to perfectly live up to ideals [26].

On the other hand, scientists, including chemists and chemical engineers, regularly agree to follow professional codes of ethics. It is thus perhaps more appropriate to ground ethical discussions in professional codes or to use such codes as a complement to broad scientific ideals. Professional codes of ethics have the additional benefit of being somewhat more specific in what they require of the members of the profession.

Here, I discuss professional chemical ethics, synthesizing guidance from the American Chemical Society's (ACS's) *The Chemical Professional's Code of Conduct* [27], the *Global Chemists' Code of Ethics* (also developed through the ACS) [28], the American Institute of Chemical Engineers *Code of Ethics* [29], the Gesellschaft Deutscher Chemiker (German Chemical Society) *Code of Conduct* [30], and *The Hague Ethical Guidelines* [31]. I apply an open coding methodology to identify common themes across the five different codes and use these themes to identify a 'general' set of ethical requirements for chemical practitioners, recognizing the significant limitation of using only five arbitrarily selected codes of ethics. Details of the open coding

approach, intermediate results, tabulated themes, and descriptions of key themes are provided in the supplementary information. The resulting synthesized ethical requirements are provided below.

The codes of ethics studied here are mainly focused on applied and experimental chemistry, as illustrated by their focus on such concepts as safety, chemical weapons, and security of chemicals and equipment. In the supplementary information, I consider how one could apply these codes to computational chemistry and data chemistry, the main foci of this work. Most chemical codes of ethics, including those analyzed here, also fail to address what happens when a chemical practitioner violates their agreed-upon norms. In the present work, I am mostly unconcerned with the repercussions of unethical behavior and wrongdoing. Instead, I aim to analyze which behaviors are ethical or unethical. I briefly discuss repercussions later in this work in section 5.

2.2.1. Stakeholders

In their professional capacity, chemical practitioners should seek to benefit, protect, and serve the stakeholders of chemical science. Chief among these stakeholders are the public, members of future generations, and the environment. Chemical practitioners should also benefit, protect, and serve their employers, clients, colleagues, and the scientific community.

2.2.2. Ethical requirements

Working in the interest of these stakeholders, chemical practitioners must:

- Advance scientific knowledge and understanding, including practical knowledge towards peaceful, benign, and beneficial applications
- Openly share knowledge and information with the public and within the scientific community
- Promote and advance peaceful, benign, and beneficial applications of chemical science while working to prevent misuse, e.g. towards the development of illegal compounds or harmful and destructive applications of dual use compounds and equipment
- In all things, work to protect the environment and work towards sustainability
- Disclose conflicts of interest and seek to minimize the impact of these conflicts of interest on their work
- Engage in the education of themselves, junior chemical practitioners under their supervision, and the broader public through formal instruction, professional development, and outreach
- Treat others fairly and with respect, minimizing the impact of bias and avoiding harassment, discrimination, bullying, and similar harmful behaviors
- Be honest and forthright in their professional and scientific communication, avoiding plagiarism, misrepresentation, fraud, and fabrication or falsification of results.
- Be aware of and abide by relevant laws and policies, particularly those related to safety and chemical security
- Protect their own health and safety and the health and safety of all relevant stakeholders
- Engage in critical oversight of one's own work, the work of others, and the work environment to ensure accuracy, safety, and security
- Report any wrongdoing, misuse, misconduct, or significant risks to relevant authorities
- Maintain and enhance, through their ethical behavior, the respectability of the chemical sciences

3. Squaring foundation models with scientific and chemical ethics

Chemical practitioners are already using generative foundation models in professional contexts, and these include non-chemistry-specific applications (e.g. image generation, brainstorming, code generation, or question answering) [32–34]. In this section, I consider how such non-chemistry-specific applications of general-purpose generative foundation models align with scientific and chemical ethics as described in section 2. That is, I seek to answer the question: is it ethical for chemical practitioners to lean on general-purpose foundation models in their off-the-shelf form? Ethical conflicts related to unique applications in and around chemistry, as well as chemistry-specific foundation models, will be explored in section 4 below.

3.1. Foundation model risks

Weidinger *et al* [35] developed a taxonomy of risks associated with LLMs. This taxonomy is not perfectly suited for my purpose, namely because it was written before the widespread adoption of multimodal VLMs (and, therefore, cannot hope to account for any additional risks that non-textual input and output introduce) and because it accounts only for LLM applications and not model training. Still, it provides a useful grounding for the present analysis.

The Weidinger taxonomy lists 21 risks in six areas: discrimination, hate speech, and exclusion; information hazards; misinformation harms; malicious uses; human–computer interaction harms; and environmental and socioeconomic harms. Many of these risks apply to applications by chemical practitioners in a generic sense. For instance, because chemistry is a relatively homogeneous field [36, 37] where students and practitioners holding minoritized identities face marginalization and a compromised sense of belonging [38] which can impact their academic and professional performance and retention [39–43], risks related to 'social stereotypes and unfair discrimination' and 'exclusionary norms' are potentially important within the context of the practice of chemistry. However, these risks are not any more relevant to chemists than they are to professionals in other settings (e.g. mathematicians or musicians). In the interest of brevity, I will not discuss these generic risks in detail. Rather, I will briefly mention some risks that are particularly relevant to the core work of chemical practitioners (e.g. understanding the natural world, developing technologies to benefit society) and their ethics.

I note that, in this text, I primarily use the term 'risk', as opposed to 'harm'. The former refers to problems or negative outcomes that may or may not occur, while the latter refers more narrowly to real negative outcomes. The risks that I discuss may have already resulted in real harms my intent is not to investigate or document these but merely to raise plausible concerns. I do, however, use the term 'harm' where it is used in the Weidinger taxonomy to allow readers to more easily connect my arguments with that previous work.

3.1.1. Misinformation harms

'Misinformation harms' refer to risks caused by LLMs outputting erroneous or misleading information. Particularly relevant to chemistry are two closely related risks: 'Disseminating false or misleading information' and 'Causing material harm by disseminating false or poor information e.g. in medicine or law'.

Studying, researching, and leveraging practical applications of the chemical sciences often involve considerable risk to human and environmental health and wellbeing, motivating chemical ethics' strong emphasis on safety (see section 2.2.2 and the supplementary information). If a foundation model is used to produce information related to chemicals, any incorrect information provided could be dangerous or even deadly. As an illustrative—if absurd—example, an LLM used to generate recipes for human consumption described a process that would produce chlorine gas [44], which is highly toxic to humans [45]. Incorrect information is inherently problematic in a scientific context, as it threatens science's pursuit of 'reliable knowledge'. Incorrect code generated by an LLM, for instance, could lead to erroneous results in a scholarly publication. Though I am not aware of any specific reports of erroneous LLM outcomes influencing the chemical literature, participants of the NSF-sponsored 'Integrating Large Language Models into the Chemistry Laboratory Curriculum' workshop reported instances of LLMs producing 'fictitious scientific journal articles, nonexistent Python libraries, and erroneous arithmetic—some of which were initially convincing' [46].

3.1.2. Human–computer interaction harms

One common setting for general-purpose foundation model use is synthetic dialogue (e.g. with ChatGPT). Weidinger *et al* identify four risks associated with such 'conversational agents' [47]. Of these, one stands out for the work of chemical practitioners: 'Anthropomorphising systems can lead to overreliance or unsafe use'. As Kidd and Birhane argued in a recent perspective [48], the apparent confidence and expertise of conversational agents make human beings likely to trust them, which makes overreliance in critical applications plausible. Notably, this risk intersects with the misinformation risks described above; overreliance on conversational agents could not only lead a chemist to hold distorted, potentially incorrect beliefs [48] but could also lead to safety risks and the spread of misinformation.

3.1.3. Environmental and socioeconomic harms

Finally, of the environmental and socioeconomic risks associated with foundation models like LLMs, two are particularly relevant to chemistry: 'Environmental harms from operating LLMs', referring to the direct energy [49, 50] and water consumption [51, 52] of foundation models, and 'Disparate access to benefits due to hardware, software, skill constraints', which suggests that the benefits of foundation models will be concentrated in the hands of those with (particularly economic) privilege.

LLMs and other generative models require immense quantities of energy and water for training and inference, and the increased use of graphics processing units for ML has directly been implicated in rapidly increasing energy consumption by data centers [53]. In principle, the energy to power data centers could be provided by renewable sources, but in practice this is not often the case, and many data centers rely on fossil fuels for power [54]. In the recent explosion of interest in 'AI', many companies in the data and tech sectors are actually becoming less sustainable, increasing rather than decreasing their greenhouse gas emissions [55].

One might contest the severity of foundation models' environmental risk on either a local or a systemic level. On the local level, one might say that, by using a foundation model, they can avoid greenhouse gas emissions and other environmental harms. This argument might be made by a chemical practitioner using an LLM to make some prediction of a chemical property for which they would otherwise need to perform an experiment. Though using a foundation model has an environmental cost, there might be a net positive effect if other environmental harms (e.g. the use of toxic or environmentally harmful chemicals) are avoided. The potential for a *relative* improvement in environmental impact is further discussed in section 4.1.2.

On the systemic level, one might take the long view. It has been suggested that foundation models could dramatically accelerate human progress, including in science [56, 57]. Yes, right now foundation models are (somewhat literally) bleeding the Earth dry, and yes, they are worsening the ongoing climate catastrophe, but what if these models eventually lead to breakthroughs in climate technologies and sustainable energy? I concede that this might be true, and 'AI' technologies may have environmental benefits in the future. However, it is dangerous and irresponsible to ignore the concrete and real harms of environmental degradation and climate change today [58, 59] because of a speculative future.

While the risk of disparate access is real, I note that the potential inequity in foundation model access within chemistry is not dissimilar to the inequity in access to other major pieces of equipment. In fact, depending on how one needs to interact with foundation models, it is possible that these models are more accessible than other chemical equipment. For instance, there are zero synchrotron light sources on the African continent [60], significantly limiting African researchers' ability to perform advanced x-ray characterization, but it is facile to access the Internet in Africa and, by so doing, access ChatGPT's applications programming interface (API) or Web interface. For more advanced or in-depth applications, the relative equity in access between experimental equipment and foundation models is less clear, as training a generative foundation model often requires many high-performance computing nodes with graphics processing units [61], and such large-scale computing resources may be inaccessible or excessively expensive for those in low-resource environments.

3.2. Threatening science's ideals

3.2.1. Objectivity

Computational and data-driven models are often praised for their objectivity [62]. However, generative foundation models are informed by human subjective biases—including hegemonic biases against marginalized groups [63–72]—and reproduce these biases in their outputs. Though ML models, lacking agency, cannot hold feelings or beliefs themselves, their behavior (i.e. output at inference) is deeply impacted by the biased feelings and beliefs of the humans whose thoughts are reflected in the models' training data. In this way, generative foundation models intrinsically lack objectivity. But how do generative foundation models impact the objectivity of their users—for the present work, (chemical) scientists?

Regardless of how biased foundation models are, they have no power to coerce users to act based on model biases. However, foundation model use may make it more challenging for a scientist to approach objectivity. As Kidd and Birhane explain [48], human users are susceptible to internalizing model biases. Because users typically interact with generative models like LLMs and VLMs when they are open to learning something new or lack necessary information, their beliefs are particularly open to influence at the time of interaction [73]. If users view models as objective sources of accurate information—as they might if they are using a generative model to answer questions—then it is possible that the model bias will be transmitted to the user and influence their behavior without them even realizing that such transmission is taking place.

The risk of foundation models introducing biases or amplifying the existing biases of scientists can be mitigated, for instance if they, as foundation model users, engage in mindfulness and reflection [74] as they receive model outputs and cross-reference with reliable sources [75]. At the same time, foundation model use is associated with reduced critical thinking [76], weakened reasoning, and reduced depth of engagement with information [77], which would seem to suggest that foundation model users are unlikely to take on the added cognitive effort required for bias mitigation. In sum, the ideal of objectivity is violated by generative foundation models themselves, and such models threaten the objectivity of scientist users.

3.2.2. Honesty

As I alluded to in section 2.1, the ideal of honesty can be violated in two ways: by making untrue statements and by omitting true statements.

While models like LLMs and VLMs have been described as 'lying' or 'hallucinating' [78, 79], this misrepresents the models' behavior. To lie, one must say something as true while knowing or believing it to be false [80]; to hallucinate, one must make a flawed perception of reality [81]. Generative foundation models have no internal notion of truth and cannot directly perceive the world in either a flawed or unflawed way. It is therefore more precise to describe LLMs, VLMs, and the like as 'bullshit' machines [82], using

Frankfurt's terminology [83]. That is, they produce outputs with no concern for truth. Although they do not lie, by bullshitting, these models frequently produce flawed or untrue outputs.

Using a tool that produces statistically likely text (as opposed to directly obtaining accurate information from a credible source or using a tool that reliably produces accurate and precise results) is dubious in the context of a truth-seeking enterprise like science and makes the effort of approaching the ideal of honesty more fraught. If scientists using these models are not careful, they could develop flawed understandings or beliefs and communicate falsehoods, thus failing to engage in truth-telling. This could occur in multiple settings, including querying foundation models to answer some question about the world and generating code which could produce results that appear to be correct but are based on flawed algorithms.

3.2.3. Openness

Openness, as the term is used in the National Academies' scientific ideals, 'refers to the value of being transparent and presenting all the information relevant to a decision or conclusion' [22]. Openness is related to, predicated on, but distinct from honesty. Transparency allows others to reproduce and build on the results of the past. When data, code, and methods are openly, transparently disclosed, external parties can also critically evaluate a scientific work, identifying mistakes or misconduct (including failures in truth-telling, e.g. fabricated data). Transparency and openness are key to science's ability to self-correct and to maintaining trust within the scientific community as well as between science and the broader public.

Transparency is a serious issue for generative foundation model use. First, models themselves vary significantly in their openness. On the one hand, there exist models such as LLaVA [84] that release their training corpus, source code, model parameters, training procedure, and relevant benchmarks. Users, researchers, and other interested parties can, in principle, interrogate these 'open source' or 'fully open' models to understand model behavior or to reproduce results. On the other extreme, some models (e.g. GPT-3) [1] are essentially opaque to users; users can interact with the model through an API or public interface but are given little information about the model's implementation or training. Between these two extremes are models that are moderately transparent, such as the 'open weight' models that disclose pre-trained parameter weights but withhold other important details (e.g. the training corpus or training methodology) [85].

Even if a scientist, using a foundation model to generate some text or code or make some prediction, aimed to be completely transparent in communicating their findings, the models that they use might frustrate openness and downstream effects like reproducibility. If a researcher using an opaque LLM reported what model they used, when they used it (i.e. what version they used), and what prompt they used, their findings may still not be reproducible if that model is no longer available (e.g. if only newer versions are available through the model's API).

The 'black box' nature of large deep learning models also somewhat challenges efforts towards openness. Let us say that a scientist uses a fully open LLM, and let us further say that the scientist discloses that they used this model and describes how they have interacted with the model. In this case, the model itself, being open source, is open to interrogation. In principle, others can try to train or fine-tune a new model using the original methods, or try to reproduce the scientist's results, though this may be limited by the frequently high computational costs of training generative models. Even in this ideal case, the process by which the model works is opaque even to the user. This is in contrast with most scientific tools, such as spectrometers and microscopes, but it is not at all unique to generative foundation models, as most deep learning models can be described as 'black boxes' that are difficult to understand or explain.

3.2.4. Accountability

As noted above, generative foundation models are 'black boxes'. The opaqueness of model behavior, even if the model code, training data, and training procedure are transparently disclosed, makes explaining and justifying what one has done challenging for scientists aspiring towards accountability.

It is also possible that less scrupulous scientists could use foundation models to avoid accountability, using excuses along the lines of, 'It was not me, it was the machine!' At least within the context of scientific publishing, accountability diffusion does not seem like a serious problem. A number of journals and publishers have already provided ethical guidance reaffirming that authors must remain individually and collectively accountable [86, 87], even if they use generative foundation models. As long as publication outlets consistently apply and enforce their guidelines, accountability can be maintained.

As we will see below, when generative foundation models are used to direct scientific processes (e.g. automating experiments), additional problems related to accountability arise.

3.2.5. Fairness

There are three elements to the ideal of fairness as I have paraphrased it in section 2.1: that criteria used for judgment and decision-making are clear, that they are equitable, and that the same criteria are used consistently.

Considerable effort has been made recently towards improving the 'reasoning' capabilities of generative foundation models. The resulting 'reasoning' models, including the o1 and o3 models from OpenAI and the R1 models from DeepSeek, can provide verbose explanations justifying their outputs. Ostensibly, this would suggest that judgments and decisions obtained from such 'reasoning' models are clear. Unfortunately, the clarity provided by such explanations is misleading. These models can convincingly explain why incorrect solutions are correct [88]. Moreover, as Sarkar explains [89], what we call foundation model 'explanations' are not really explanatory (Sarkar prefers the term 'exoplanations'), in that the text generated for the models' justification does not actually reflect how the model obtained its judgment or answer. Because the 'explanations'/'exoplanations' of 'reasoning' models are meaningfully disconnected from the model's final answers, it is also unlikely that the stated criteria and explanations used by foundation models are consistent. Indeed, model inconsistency has been widely reported in the natural language processing community [90–93].

Finally, generative foundation models do not equitably make decisions. The same intrinsic biases and prejudices that threaten or violate the objectivity of foundation models (see section 3.2.1 above) make the models unfair.

The use of an unfair model, as described here, does not in and of itself make a scientist's behavior unfair. There may be applications of generative foundation models where the aforementioned problems regarding clarity, consistency, and equity are not significant or relevant. With this said, the ideal of fairness is threatened whenever generative foundation models are used to answer questions, make judgments, or make decisions, which is a major goal of state-of-the-art 'reasoning' models and, as will be described below, a major application of foundation models in chemistry.

3.2.6. Stewardship

In this section, I have argued that foundation models lack objectivity, bullshit rather than consistently producing truthful statements, are variably opaque, and are in many respects unfair. These flaws, combined with the risks identified in section 3.1, undermine the ideal of stewardship.

Science functions on an assumption of honesty [22]. We as scientists often believe what we read in academic journals or hear in seminars and conference presentations without confirming through independent observations because we assume that those involved in the studies that we are reading about or listening to are truth-telling individuals who are being transparent and not omitting significant details.

The risk of misinformation associated with generative foundation models calls this trust into question. If one believes that a piece of scientific work was generated in full or in part by a foundation model, one might reasonably suspect that there are flaws in the information presented. Note that this line of inquiry and skepticism does not require one to assume that any researcher intends to harm others. It only requires the belief that researchers are willing to cut corners to publish, which, given increasing demands to 'publish or perish' [94, 95] in an increasingly tight academic job market [96], is plausible.

If the mutual trust that undergirds scientific inquiry is disrupted, then 'every scientist for themselves' could become the norm, with more and more time being spent trying to verify claims rather than using established studies to move one's own work forward. Indeed, in the most extreme (though unlikely) case of widespread distrust, science could cease altogether to be a community effort if researchers feel that they could more effectively produce reliable results in isolation than by consulting a flawed and unreliable literature.

By the same token, the use of foundation models has the potential to damage public trust in science if the public has reason to believe that statements made by scientists are becoming less reliable. Such a loss of trust—or, said another way, such a breach of the public's confidence—would be damaging for society as a whole, as public trust in science is associated with better public health outcomes [97] and stronger action related to climate change [98].

Even outside of eroding trust via bullshit, foundation models can risk the ideal of stewardship. LLMs and VLMs frequently produce plagiarized text [99, 100]. By repeating ideas without appropriate attribution, foundation models can damage relationships within the scientific community, erasing researchers' contributions and implicitly eliminating them from scientific conversations. Finally, the unsustainable—and rapidly growing [53, 55]—energy and water use required for generative foundation models worsens relations between scientist users and the environment.

3.3. Professional chemical ethics of generative foundation models

While it is deeply troubling in just how many ways and to what extent foundation models challenge science's ideals, it is hardly surprising that they are not in full alignment. As I noted above, ideals are aspirational and do not necessarily represent reasonable expectations. Even so, foundation models also go against or threaten several of the ethical norms of the chemical profession(s).

The preceding discussion has already alluded to threats to or violations of several of the ethical requirements laid out in section 2.2.2. Generative foundation models do not prevent chemical practitioners from '[treating] others fairly', but the intrinsic biases present in foundation models and reflected in model output (section 3.2.1) hardly '[minimize] the impact of bias'. Use of generative foundation models further risks violation of the requirement for chemical practitioners to 'Be honest and forthright in their professional and scientific communication, avoiding plagiarism, misrepresentation, fraud, and fabrication or falsification of results'. As I have discussed (section 3.2.2), models like LLMs, while not lying *per se*, nonetheless make frequent untrue statements. They also plagiarize in their outputs (section 3.2.6). It is therefore possible for any chemical practitioner using generative foundation models in their scientific communications to (perhaps unintentionally) engage in forms of dishonesty. The massive resource burden of generative foundation models (section 3.1.3) additionally makes their use incompatible with the norm of '[protecting] the environment' and '[working] towards sustainability'.

Generative foundation models may also pose some risk to 'the respectability of the chemical sciences'. However, this depends on public perception of such foundation models and the information that they provide. As I discussed in section 3.2.6, a belief that scientific statements are unreliable could break the trust between chemical scientists and the public.

This is not to say that generative foundation models are unethical with respect to every aspect of chemical professional ethics. Foundation model use has little impact on a number of professional ethical norms. For instance, take the requirement that chemical practitioners 'Disclose conflicts of interest and seek to minimize the impact of these conflicts of interest on their work'. Use of ML models introduces no significant conflicts of interest in the general case, and even where conflicts of interest do arise, the nature of such conflicts is not necessarily specific to foundation models. A chemist researching VLMs while having a major stake in a company developing or making use of VLMs has a conflict of interest in much the same way as a biochemist who founds a biotechnology company in the area of their research.

In some ways, generative foundation models can even be said to be in alignment with chemical professional ethics. The large and growing number of peer-reviewed publications at the intersection of generative foundation models and chemistry would suggest that not only do the researchers performing and communicating such research believe that it helps to 'Advance scientific knowledge and understanding', but that others in the scientific community agree.

It is also possible to consider that adherence to chemical professional ethics in some areas might mitigate the ethical risk and harms of foundation models in other areas. Notably, by '[engaging] in critical oversight' of work making use of or examining generative foundation models, researchers can potentially mitigate misinformation risks. As a concrete example, if a chemical practitioner used a generative model to draft a document (say, a safety data sheet or a scholarly manuscript) for publication, the practitioner themselves should critically examine the text that has been generated and review it for mistakes. External parties (e.g. other chemical practitioners engaged in peer review) should also critically review the document before it is published. Even if the original text is deeply flawed, the final document resulting from this process of oversight and revision could be free from errors.

4. Chemical foundation model use

A number of thorough reviews have been written on the uses of general-purpose generative foundation models in the chemical sciences and chemistry-specific foundation models [101–105]. Here, rather than repeating this work, I will briefly and non-exhaustively survey some areas where generative models (mainly LLMs) can be and have been used, namely, literature review and text generation; chemical classification and regression tasks; laboratory automation; and education. I will then discuss chemical foundation models. In doing so, I point to ethical problems and how these models, in their chemical applications, align or fail to align with scientific ideals and chemical professional ethics.

4.1. Applications of general generative models in the chemical sciences

4.1.1. Literature review and text generation

Many applications of generative foundation models in the chemical sciences share significant similarities with applications in other areas. For instance, there has been considerable interest into using LLMs for chemical information querying and question-answering tasks [106]. LLMs appear to be particularly useful in

generating structured data from unstructured text [7, 107], which is beneficial in e.g. constructing databases of synthesis recipes. Alone, LLMs are not well equipped to answer complex questions, particularly when calculations are required, but models that can generate and execute code through external tools are considerably more effective [6].

Information extraction and question answering are domains where misinformation risk is high. While providing LLM 'agents' with external tools decreases the likelihood of erroneous model outputs [108, 109], such false statements cannot be entirely avoided [110, 111].

The risk of misinformation threatens the scientific ideals of objectivity, honesty, and stewardship, as well as the chemical ethical norm related to honest communication of scientific results. These risks are elevated when the text generated by an LLM or similar model is used in materials for publication, a phenomenon which is becoming more common in the scientific literature [112, 113]. Relatedly, the risk of plagiarism from e.g. LLMs and VLMs becomes relevant when such models are used to generate text for publication or dissemination, further straining the ideal of stewardship by failing to recognize the work of other researchers and going against a specific, explicit ethical prohibition.

Tasks which rely on an LLM 'interpreting' or 'summarizing' text and generating unstructured responses are also open to model biases. This presents an additional challenge to scientific objectivity—even a well-meaning scientist who works hard not to let their personal biases impact their judgment and behavior could be undermined if the text generated by their foundation model tool is heavily influenced by intrinsic biases.

Finally, fairness is an issue when foundation models are used for literature review and writing. As I discussed in sections 2.1 and 3.2.5, the ideal of fairness relates to decision-making. Whenever decisions must be made the criteria must be clear, equitable, and consistent. Literature review and writing are processes which require significant decision-making at all stages. In a literature review, a researcher (or LLM) must decide what information is relevant and what is irrelevant, what ideas to emphasize, how to resolve possible conflicts or contradictions in the literature, and more. In scientific writing, decision points are even more plentiful, from who to cite and what tone to use to how to present an unexpected data point that goes against a statistical trend. The abundance of intellectually and epistemologically critical decisions makes applications of LLMs and similar foundation models problematic; after all, generative foundation models are intrinsically unfair, lacking clear judgment criteria, being inconsistent in applying such criteria, and failing to make judgments equitably (section 3.2.5).

4.1.2. Classification and regression tasks

Recent efforts have considered if general-purpose foundation models can act as substitutes for specialized chemical and materials ML models. The idea is simple: if, by being trained on a huge corpus including the open literature, an LLM or similar model has some underlying 'understanding' (i.e. representation) of chemical concepts, then it should be possible to fine-tune these models to make arbitrary predictions about chemical entities. It appears that LLMs are able to compete and in some cases even exceed small, single-purpose ML models when available data are scarce [114]. In particular, LLMs seem to excel at classification and struggle somewhat more on regression and generation tasks [106]. When ample data are available, however, LLMs do not provide significant benefits.

Applying generative foundation models for chemical classification and regression tasks may be more ethically sound than using such models for text generation, as described in section 4.1.1 above. In particular, the environmental impact of generative foundation models could be less severe (in relative terms) for classification and regression than for other applications of the same models because, in the absence of generative foundation models, chemical practitioners would still train and use an ML model and incur the associated environmental costs. Though it is likely that such specialized ML models are considerably less expensive for both training and inference than generative foundation models (a review of chemical ML's energy and water demands has not, to the best of my knowledge, been reported), the cost of data generation necessary to train specialized models to an acceptable level of accuracy could at least in some cases outweigh this difference in environmental impact for the models themselves.

Given that, in this case, practitioners are explicitly and knowingly making predictions about potentially unknown quantities, the model results for classification or regression are likely to be scrutinized. This is in contrast to question answering tasks, where the user queries a model to obtain some ostensibly true information which they believe to be known and available (e.g. in the scientific literature). Thus, generative foundation model-based classification and regression may be somewhat less vulnerable to misinformation risks. However, if a generative foundation model makes an erroneous prediction, the user will typically be less able to explain why the model erred than if they had used a more specialized classification or regression model (see section 3.2.3); this is especially true for users of opaque models. This limits the possibility for accountability and potentially frustrates efforts towards oversight.

4.1.3. Laboratory automation

As part of a growing movement towards automated high-throughput experiments and 'self-driving laboratories' [115, 116], LLMs have been used to direct complex chemical processes. To accomplish this, a model is given access to an API for a laboratory automation system, which could control a robot or other instrument or could send instructions to an existing automated laboratory. Thus far, this approach has been demonstrated on well-studied systems. Boiko *et al* [8] applied their LLM-based laboratory automation system, Co-scientist, to optimize Suzuki and Sonogashira cross-coupling reactions, while Bran *et al*'s ChemCrow [109] demonstrated the ability to synthesize non-trivial products such as N,N-diethyl-meta-toluamide (DEET). In the area of optimization, LLM-based approaches appear to be competitive with the more traditional Bayesian optimization [8]. It is unclear how models like ChemCrow and Co-scientist, which in addition to laboratory APIs can access the chemical literature through the Internet, will be able to generate and execute experimental plans to study truly novel chemical or materials systems for which no literature recipes exist.

Granting an LLM control of physical laboratory tools presents many novel ethical concerns. First, it poses LLMs as dual-use technologies in the context of chemistry [117]. Just as they could be used for benign purposes like synthesizing DEET, they could be used to harmful ends. Boiko *et al* found that LLMs such as GPT-4 could be directed to synthesize compounds that are dangerous or illegal if given an appropriate prompt [8], bypassing the model's 'guard rails'. While this does not introduce a new capability for experienced chemical practitioners (i.e. a chemist likely already has the necessary knowledge to synthesize an illegal drug or chemical weapon), it does considerably lower the barrier for non-experts to undertake harmful chemical projects. In fact, even if a chemical LLM 'agent' like ChemCrow is not connected to laboratory automation equipment, the ease with which it proposes feasible synthesis recipes potentially lowers the barrier to synthesizing harmful and/or illegal chemical products. Because chemical practitioners have an obligation not only to promote and advance beneficial applications of chemical technologies technologies but to prevent misuse, additional efforts must be taken to minimize the risk of harm caused by LLMs and related models as dual-use technologies.

LLM-based laboratory automation further presents significant ethical questions around safety. In the contemporary laboratory, chemical practitioners are generally required to receive safety training and certification before using specialized equipment or handling dangerous chemicals [118, 119]. Junior practitioners are also typically supervised by more senior practitioners to ensure that they are conducting their work safely and appropriately. These requirements do not guarantee safety but provide a form of accountability and oversight, in line with scientific ideals and chemical professional ethics.

As Boiko *et al* and Bran *et al* have shown [8, 109], LLMs can perform chemical procedures entirely without human intervention. But since these models lack agency, they cannot undergo safety training, receive certifications, or be held accountable for unsafe behavior. Considering this, one could ask: what safeguards are in place to protect the safety of humans and the laboratory environment when LLM 'agents' are in the lab?

One answer would be that appropriately trained chemical practitioners should oversee all procedures conducted by tools like ChemCrow and Co-scientist. In this case, the LLM is essentially treated as a junior practitioner that has not yet sufficiently demonstrated their trustworthiness. This is a reasonable course of action but would somewhat defeat the purpose of laboratory automation. One might argue that, by fine-tuning models on safety procedures and equipment documentation, adherence to safety can be guaranteed, but this confidence is misplaced. As I have already discussed, LLMs unavoidably produce erroneous outputs (which, in this context, could mean *unsafe* outputs) and can go against their training (as demonstrated by Co-scientist being 'tricked' into producing harmful and illegal substances).

4.1.4. Education

Some groups have advocated for the use of generative foundation models (mainly LLMs) in educational settings [120]. In the context of the chemical sciences, proposed applications include: (1) assisting instructors in course and lesson planning [121, 122]; (2) automating or accelerating grading and evaluation [46, 122, 123]; (3) playing the role of 'tutors' or 'virtual teaching assistants' by providing students with direct guidance and instruction [46, 122, 124]; (4) helping students to study [46]; and (5) assisting students in completing course assignments [46, 122, 125–127]. These diverse educational settings raise different ethical questions and have different possible benefits and harms, so I will consider them one by one.

Instructor assistance: Du *et al* [122] in their perspective on LLMs in chemistry education, identify two ways that instructors might use LLMs for preparation: LLMs could assist in content creation, and they could 'simulate pedagogical scenarios', with LLM 'agents' taking on the role of students. The benefit of LLMs for content creation appears dubious and limited at best. Clark *et al* [121] found that LLMs could not generate useful content themselves (as assessed by instructors), even on the subject of historical experiments that are

likely well covered in LLM training data, but the models could assist in areas like generating lesson outlines and suggesting additional learning materials. The possible benefit of LLM simulation of instructor-student interactions is more difficult to evaluate. LLM simulations could, in principle, help to train instructors and make them better prepared in their interpersonal interactions; however, to the best of my knowledge, such methods have only been prototyped [46] and have not been widely deployed or closely studied in the intended setting.

Because instructors are presumably experts in or at least knowledgeable of the subjects that they teach, they are unlikely to be convinced by any bullshit produced by an LLM assistant. I therefore judge the misinformation risk of LLMs for course and lesson preparation to be low. However, the ethical risks and harms of model social biases (section 3.2.1) become salient when LLMs are used for educational simulations. Given LLMs' penchant for reproducing hegemonic prejudice in their outputs, it is possible that the synthetic students represented by LLMs will either reflect dominant social groups (failing to model the diverse perspectives of the student body) or else will present minoritized students using stereotypes, thereby potentially biasing the instructors using the LLM for training and preparation. Applications based around instructor assistance thus potentially threaten the scientific ideals of objectivity and fairness and the professional ethical requirement to minimize the impact of bias.

Grading and evaluation: Multiple recent publications [46, 122] have suggested that LLMs could assist in designing student assessments, automate grading, and provide feedback to instructors.

Detailed feedback is essential for student learning [128], but providing feedback is costly, requiring a significant time investment from instructors [129, 130]. This is especially true for large classes, which are common in the chemical sciences [131, 132]. It is understandable why instructors with limited time might turn to technologies like LLMs to automate student evaluation and feedback. Unfortunately, there is no obvious and significant benefit to using LLMs for student assessment. While it is true, as Du *et al* point out [122], that automated grading could eliminate *human* bias, it is doubtful that LLM-based grading would reduce grading bias overall, given the intrinsic biases embedded in LLMs. Similarly, LLM grading would eliminate human errors but would introduce errors based on erroneous LLM output [123]. The authors suggesting that LLMs could accelerate grading have further failed to make a convincing comparison to existing automated grading systems that do not rely on LLMs [133, 134].

As I discussed in section 3, there are significant risks and ethical problems associated with these models in general, and some, such as unsustainable energy and water use, are presently unavoidable. If there is no or very limited benefit to using LLMs for student evaluation, LLM use in this area cannot be ethically justified.

There may be some benefit to instructors applying LLMs to evaluate themselves. Currently, instructors (at least within higher education) mainly receive feedback from student evaluations given at the end of the course [135, 136]. This paradigm has a host of problems, from reflecting significant prejudice against marginalized instructors (e.g. women and people of color) [137, 138] to low participation rates, selection bias [139], and ease of over-interpretation [140]. Some instructors also resist student feedback, viewing students as insufficiently knowledgeable to judge their education [141]. Even if these varied and significant issues could be addressed, end-of-term evaluations fundamentally do not allow instructors to improve their teaching as they go along through a course. LLMs could provide rapid feedback at any point during a course while avoiding some of the problems associated with student evaluations. Viewed as a supplemental means for instructor development and self-improvement, LLM-based instructor evaluation could aid chemical practitioners in fulfilling their obligations to engage in education and critical oversight.

Virtual teaching assistants: Much of the focus on LLMs in chemical science education has been placed on virtual teaching assistants. According to proponents, these LLM 'agents' could help answer student questions [124], help students interact with technical documentation, and even oversee students in during hands-on laboratory experiments [46, 122].

Question-answering is more problematic for students than it is for instructors. Students, by definition, are non-experts who may lack basic understanding of the course material. They are therefore ill-equipped to distinguish between true and false information. Misinformation risk, coupled with the human–computer interaction risk of overreliance (section 3.1.2), could lead to a disaster for student understanding. Indeed, there is considerable risk not only of LLMs providing erroneous information but doing so in a manner that students are likely to believe. Yik and Dood [124] found in their study on LLM descriptions of reaction mechanisms that the model explanations were often subtly wrong but provided with a high level of sophistication, lending them an air of authority. Providing LLM 'virtual teaching assistants' presents a plausible hazard to student education, going against a key chemical ethical norm.

Most alarmingly, LLMs have been suggested as virtual *laboratory* assistants, providing students guidance on laboratory protocols and safety and monitoring student activities [46, 122]. Teaching safe laboratory

procedures is essential to students in the chemical sciences; indeed, teaching safety skills is a core goal of any laboratory course. At the same time, safety and laboratory procedures are areas where small mistakes (e.g. adding water to acid vs. adding acid to water) can lead to significant health risks. By their design, LLMs cannot be guaranteed to provide only accurate safety information, creating an unacceptable risk to student health and wellbeing. In particular, LLMs are prone to producing outputs that are incorrect but which appear trustworthy and convincing, a phenomenon known as 'adversarial helpfulness' [88]. Adversarial helpfulness makes it likely that LLMs will produce subtly but dangerously incorrect guidance that students are likely to believe, as in the example of water and acid given above. While human instructors are also fallible, they can receive certifications demonstrating their ability to adhere to appropriate protocols and can be held accountable if mistakes occur. On the other hand, an LLM, lacking agency, cannot be held accountable for any erroneous safety information that it provides.

Study assistance: Similar to virtual teaching assistants that provide detailed but incorrect explanations, LLM-based tools for studying have a high risk for misinformation. Because of the limited background of chemical science students, this misinformation is likely to lead to flawed student understanding and hamper student education. Moreover, as I noted in section 3.2.1, use of LLMs is associated with a reduction in critical thinking and engagement with information. This suggests that, even if an LLM provided accurate information, students may not use LLMs effectively to gain a deep understanding of course material.

Tools for student assignments: Finally, LLMs have been suggested as a tool for students to use to complete course assignments, including developing models [127] and writing laboratory reports [46, 122, 125, 126]. Particular emphasis has been placed on the models' abilities to generate code and aid in writing.

A common argument in favor of LLMs as education tools is that, because LLMs can effectively generate code, students will not need to learn computer programming and can instead focus on understanding the core chemical material [122]. This argument is flawed and misleading. It implies that, by using LLMs instead of conventional programming, students will need to learn fewer extraneous skills. In reality, students will still need to learn a new non-chemistry-specific skill, namely the skill of effectively prompting LLMs. It is widely recognized that the performance of LLMs is heavily dependent on the prompting strategy used [142, 143], including in chemical contexts [124], and Subasinghe *et al* [126] found that, of the students surveyed in their study using LLMs for data visualization, 90% found the experience 'challenging'. The use of LLMs may not even eliminate the need to learn to program, as students may need to double-check LLM-generated code to ensure correctness. Validating this point, Subasinghe *et al* found that LLMs, when prompted to analyze and visualize a dataset, sometimes generated code and associated plots using fabricated data. This not only produces an incorrect result but might inadvertently cause a student to commit academic misconduct—a clear violation of the ideals of honesty and accountability as well as several professional norms.

With regard to writing, Du *et al* [122] suggest that LLMs 'aid in cultivating critical thinking in students', but this is in contrast with the initial evidence that links LLM use with a *reduction* in critical thinking [76, 77]. LLMs, used to generate text for an essay or manuscript, will also sometimes plagiarize, which is explicitly prohibited in the framework of chemical professional ethics described in section 2.2.2 and further undermines the ideal of stewardship by erasing the contributions of previous authors. There may be some permissible applications of LLMs in student writing (e.g. correcting grammar), and these applications could even lead to positive ethical outcomes. For instance, access to LLM-based writing assistance could benefit studenst who are learning and writing in a language in which they are not fluent. Even still, the ethical risk of assigning these tasks to LLMs is weighty.

4.2. Chemical foundation models

4.2.1. What are chemical foundation models?

The term 'foundation model' is somewhat ambiguous, and multiple definitions have been put forward [144, 145]. In the meaning that I use here (see supplementary information), the key features of a foundation model are that they are trained (or pre-trained) on vast datasets and that they can be applied (directly or via fine-tuning) to many diverse tasks. Depending on the domain, what qualifies as a 'vast' dataset or 'diverse' tasks may vary. Here, discussing chemical foundation models, I consider datasets that are vast *for the chemical sciences* and diverse *chemical* tasks. Even still, these concepts are relative and subjective, and different researchers might reasonably hold different opinions regarding which models are and are not 'foundation models'.

Using this definition, I argue that there already exist a number of classes of *chemical* foundation models. Perhaps the most famous family of foundation models are protein models, which won part of the Nobel Prize in Chemistry in 2024 [146]. Given a string of amino acids, these models generate folded protein structures, thus addressing one of the greatest challenges of biology and biochemistry. In some cases, protein model

architectures resemble LLMs. These 'protein language models' [105, 147] treat amino acids as tokens and protein sequences as text. Other models, including the AlphaFold family [10, 148] and RoseTTAFold [11], rely on other advanced ML techniques, such as equivariant transformers [149] and diffusion layers [150]. I consider protein models as foundation models because they are trained on essentially all known protein structures (i.e. all structures in the Protein DataBank or PDB) [11, 151] and because they can be applied to problems including protein structure prediction [10], protein design [152, 153], protein interaction discovery [154, 155], drug discovery [156, 157], vaccine development [158], mutation effects [159], and more.

Recent years have seen the development of so-called 'universal' MLIPs [12, 160]. An interatomic potential, or force-field, is a mathematical function that calculates the potential energy of a system of atoms in space. Rather than parameterizing expert-designed empirical functional forms [161–163], MLIPs predict system energies and atomic forces using an ML model, commonly a graph neural network [164, 165]. By training on data spanning the periodic table [166–168], 'universal' MLIP models can be employed to simulate diverse chemical systems, including some very far outside of the training distribution. For instance, the MACE-MP-0 MLIP [13] was shown to behave well (at least qualitatively) on systems ranging from gas-phase hydrogen combustion to liquid electrolytes and many systems in between. Though most applications of MLIPs involve energy prediction or molecular dynamics, these models can also be fine-tuned to predict various molecular and/or materials properties [160].

While I will mainly focus on protein models and 'universal' MLIPs, one can point to other types of chemical foundation models. For instance, similar to a protein language model, Cai *et al* created ChemFM [169], a foundation model trained on molecules represented as Simplified Molecular Input Line Entry System (SMILES) strings for molecular generation and property prediction tasks.

4.2.2. Risks of chemical foundation models

The chemical foundation models mentioned above avoid a number of the risks associated with general-purpose generative foundation models (e.g. those identified in the Weidinger taxonomy). Since protein models, 'universal' MLIPs, and other chemical foundation models do not model or directly generate speech, there are no risks associated with discrimination, stereotyping, or social prejudice. These models cannot contribute to misinformation or disinformation in any direct way, and there is little risk of the models being anthropomorphized by users or manipulating users. The data used to train chemical foundation models (e.g. density functional theory calculations, experimental protein structures, or SMILES strings) are typically derived from public sources and do not contain sensitive or personal information; together, this means that one should not be concerned about these models accidentally or intentionally releasing private information.

This is not to say that there are no risks associated with these models. To the extent that these models are used in critical decision-making settings—for instance, deciding which drug candidates to study or suggesting a synthesis route—incorrect predictions could still cause harm. In the provided examples, harms could include waste or misallocation of resources or inefficiencies; direct harm to human health and wellbeing is unlikely, given that these models, in their present form, do not directly interact with laboratory equipment or human subjects in a clinical setting.

The main ethical risks that these chemical foundation models introduce are environmental in nature. Though 'universal' MLIPs, protein models, and their ilk are small in comparison to state-of-the-art generative foundation models in terms of dataset size and number of parameters, they are large by the standards of (bio)chemistry. General-purpose models often have higher energy demands than specialized models [170, 171], suggesting that chemical foundation models likely contribute more to environmental degradation and climate change than an average ML model in the chemical sciences.

4.2.3. Chemical ethics for chemical models

In contrast to general-purpose generative models, I find that chemical foundation models are fairly well aligned with science's ideals:

Objectivity: Although all datasets, including those used for protein models (e.g. the PDB) and 'universal' MLIPs, are informed by human bias, the outputs of chemical foundation models are not tainted by human social prejudice. That is, a chemical foundation model output might be more or less accurate in certain areas of the parameter space because of data biases but will not produce outputs that reflect prejudices like racism and sexism. Moreover, these models, as they exist today, will not and cannot directly make biased or prejudiced decisions.

Honesty: Just as essentially every dataset is biased, essentially every model will sometimes be incorrect. The important difference between a general-purpose generative foundation model and a chemical foundation

model is that while the former is not and cannot be designed to produce truthful outputs (section 3.2.2), the latter can. Protein models are trained to (for instance) predict three-dimensional protein structures comparing against experimental 'ground truth', and MLIPs are evaluated against *ab initio* quantum chemical values. As the loss and evaluation metrics of these chemical models improve, developers and users can reasonably claim that they are more closely approximating some 'true' values and more reliably producing valid outputs, at least within the distributions used for evaluation. Chemical practitioners using chemical foundation models should still communicate the limitations of their models and recognize that they will sometimes fail, but provided such communications are made, the ideal of honesty is not under threat.

Openness: Transparency remains an issue for chemical foundation models. As I mentioned above, the data used to train these models are usually well defined and publicly available. However, the models themselves are not necessarily open. As a notable example, Alphafold3 [148], at the time of its publication a state-of-the-art protein model, was published without *any* of the source code being made public (the authors did describe the model using pseudocode). This means that the authors' claims are essentially unverifiable and that the published findings cannot be easily reproduced. While chemical foundation models may not be as large or complex as models like LLMs and VLMs, they are still massive ML models that can function as 'black boxes' (section 3.2.3), which also limits a well-meaning user's ability to be open regarding their methods and how their tools work.

Accountability: In section 3.2, I found that the ideal of accountability was perhaps the least problematic for generative foundation models. Chemical foundation models are even less problematic in the area of accountability. While generative foundation models, which can automatically take actions in digital and physical settings, can in principle be used to diffuse accountability, MLIPs, protein models, and the like do not directly take action in and of themselves and therefore cannot be claimed to take responsibility away from chemical practitioners.

Fairness: Since chemical foundation models do not directly make decisions or take actions, they do not violate the ideal of fairness. Regardless of what human biases are involved in the model's training data, humans are ultimately the actors making decisions based on chemical foundation models' outputs. This does not mean that chemical foundation models are necessarily fair or help to ensure fairness, but they are not in and of themselves unfair.

Stewardship: While the energy and water demand of chemical foundation models has not been well studied, it is reasonable to assume that the development and use of these models has a negative direct impact on the environment and climate. This potentially strains relationships between chemical practitioners and key stakeholders (section 2.2.1), namely the natural environment itself and future generations. At the same time, chemical foundation models avoid many of the other issues around stewardship that generative foundation models suffer from. As noted above in the discussion of honesty, chemical foundation models are often reasonable approximations of 'ground truth' chemical data, and thus, their use is unlikely to threaten internal or public trust in science.

Chemical foundation models are also well aligned with most of the professional ethical norms of the chemical sciences. As I mention above, chemical foundation models can be incorrect but do not contribute to misinformation or plagiarism as general-purpose generative foundation models do. As they do not intrinsically carry social biases and do not generate speech-like text, they will not contribute to discrimination, harassment, or other unfair social practices. They do not significantly impact chemical practitioners' duties related to education (i.e. these models have not been suggested or applied as replacements for instructors or conventional educational tools) or oversight. As they do not directly create or manipulate chemical substances, chemical foundation models are unlikely to create any direct risk to human health and safety.

The inability of chemical foundation models to directly create chemical substances also renders the dual-use risk of these technologies insignificant. One could certainly use a chemical foundation model to a harmful end (e.g. designing a toxin or simulating energetic materials for weapons), but the model's involvement in any harmful outcomes is highly indirect. Calling a 'universal' MLIP a dual-use technology would be akin to declaring a molecular dynamics code to be dual-use technology—perhaps technically true (after all, the MD code can also be used to simulate the behavior of toxins or explosives), but not a useful classification. From the perspective of chemical professional ethics, the main risk related to chemical foundation models is, again, the potential for unsustainable resource use and environmental damage that they introduce.

5. Discussion

To this point, I have discussed foundation model ethics with reference to science's ideals (drawing on a published set of ideals from the National Academies) and the professional requirements of chemical practitioners (synthesized through qualitative analysis of chemical codes of ethics). I argue (section 3.1) that the most significant risks of general-purpose generative foundation models are misinformation, overreliance, and harm to the environment and climate. Considering these and related risks, I find (section 3.2) that there are conflicts between generative models and most all of science's ideals, with the ideal of accountability being the least obviously and severely compromised. Use of generative models also goes against several key ethical requirements of chemical practitioners (section 3.3), though here the case is more mixed. In fact, while some norms are clearly violated (e.g. the norm that chemical practitioners must protect the environment), it can be argued that generative models actually support the ethical practice of chemistry in some other ways (e.g. by advancing scientific knowledge). In contrast, chemical foundation models (by which I mainly refer to protein models and 'universal' MLIPs) are in alignment with most ethical ideals and norms that I consider here (section 4), with the main risks being possible lack of transparency and environmental harm. With these considerations in mind, I now look forward, considering how the chemical profession(s) should address foundation models to support the ethical practice of science.

5.1. Addressing the harms of foundation models

Considering the diverse ethical concerns laid out above, I begin my discussion by asking a simple and necessary question: should chemical practitioners be using or supporting the use of foundation models?

If chemical practitioners are principled and committed to ethical behavior, the answer should in many cases be 'no'. Certainly, these models should not be used in cases where there are dubious benefits and significant ethical risk. Based on the analysis in section 4.1, for instance, I would put forward that LLMs should not be used in most educational settings, for answering questions, or for most writing tasks.

But what of cases where there are clear benefits to using foundation models? What about using generative models for classification or regression tasks in data-scarce regimes (section 4.1.2)? What about using LLMs to direct a self-driving or automated laboratory (section 4.1.3)? What about training and using 'universal' MLIPs (as opposed to using classical force-fields or more specialized MLIPs)? In these cases, while one could still argue that using foundation models goes against scientific and chemical ethics, it is also reasonable to argue that use of these models *advances* chemical ethics. That is, one could say that a foundation model, if well suited to a particular task in chemistry, serves to 'advance scientific knowledge and understanding' and can help 'advance peaceful, benign, and beneficial applications of chemical science'. The tension between conflicting ethical norms (i.e. advancing science and minimizing risk to health, safety, and the environment) is nothing new to the chemical sciences. An appropriate, pragmatic response hinges on the extent to which the tool in question is *necessary* for a particular task.

To support this line of necessity, consider some other areas where chemical practitioners use ethically problematic tools. A number of commonly used solvents are known to be highly harmful to human health and the environment [172]. Plastics, which also degrade human and environmental health [173, 174], are ubiquitous in our laboratories and are a major product of the chemical industry. And cobalt, a key component in modern lithium-ion batteries, is linked with (alleged) human rights and labor abuses in the Democratic Republic of the Congo [175, 176]. It must be acknowledged that the nature of toxic solvents, plastics, and cobalt-based batteries are significantly different from foundation models; however, the ethical risks of using, say, plastics and an LLM are overlapping. Both harm key stakeholders of chemical science (the broad public, the environment), and both go against the chemical sciences' obligation to promote sustainability. Moreover, as I describe in the supplementary information, software and ML models can for the purposes of chemical ethics be thought of as 'substances', extending the connection between foundation models and the examples mentioned here.

In the cases of toxic solvents, plastics, and cobalt (as well as many others), chemical practitioners are, as a class, aware of both the significant ethical risks and harms associated with these substances and tools and their utility. The loss of these substances and tools would even make some chemical practitioners (e.g. some synthetic chemists or battery researchers) unable to effectively perform their professional duties and meet their obligations. In these cases, where ethically problematic tools are deemed necessary, what can and should chemical practitioners as individuals and the chemical community collectively do?

The main strategies being applied today involve eliminating uses of problematic substances/tools or else replacing them with more ethical alternatives. For toxic solvents and for cobalt, researchers have sought ways to reduce the amount needed or avoid these materials altogether, leading to the explosion of interest into so-called 'green solvents' [177, 178] and Co-free battery electrodes [179, 180]. While the original materials are still used (e.g. Co-based Li-ion batteries remain on the market and an area of research interest), the

chemical community is largely moving towards elimination. There is also some interest in minimizing the use of plastics [181], but there has been a greater emphasis on developing alternative chemistries that are environmentally benign [182, 183] or identifying methods to circularly recycle plastics [184, 185], thereby reducing or eliminating the harm caused by their waste.

By analogy, once we as a community identify foundation models (particularly generative foundation models) as harmful or ethically unacceptable, we can seek to minimize their use. We should be asking, 'Where are LLMs, VLMs, etc offering truly unique, necessary benefits?' and eliminating any non-'essential' uses. For areas where foundation models are presently necessary, such as protein folding, researchers should seek to develop alternatives that are less harmful and/or seek out ways to substantively reduce harm. I note that, while these changes can be enacted by individual chemical practitioners, they will be more effective if implemented structurally, whether at the level of universities and other research bodies, chemical professional societies, or policy-makers.

5.2. Amending chemical ethics

I have argued in this piece that, if a new technology and established ethics are at odds, then the technology should not be used, or steps should be taken to minimize the harm done by that technology. An idea central to my analysis but assumed up to this point is that science's ideals and chemistry's ethical norms remain sound. But this is not to say that the ethical codes that chemical practitioners agree to and must abide by should not be amended to respond to changes in technology such as the development and deployment of foundation models.

Computational chemistry and chemical data science are in some respects ignored by existing codes, which instead focus largely on applied, experimental chemical science (see supplementary information). Given the growing importance of software, data, and ML models to the practice of science, as well as the emergent ethical concerns surrounding these computational tools, further ethical guidance should be developed. In some cases, concepts already mentioned in chemical ethical codes could be expanded to include computational chemistry and chemical data science. For instance, where existing codes discuss the security of chemicals and chemical equipment, they should also mention the chemical practitioner's obligation to (where relevant) secure sensitive data and to prevent the misuse of dual-use computational technologies. One could also imagine the development and communication of entirely new ethical requirements; as a rather general example, chemical practitioners could be obliged to avoid passing critical tasks (e.g. manuscript writing, peer review, and mentorship) to 'AI'.

Finally, many chemical codes of ethics (including most of the codes considered here) do not include any mechanisms for what should happen if chemical practitioners behave unethically. While one can hope that chemical professionals will follow the codes that they have agreed to and aspire to the ideals of science, there will almost certainly be some unscrupulous or malicious chemical practitioners. Chemical organizations would be wise to consider explicitly laying out protocols to address unethical behavior, potentially to discourage such behavior but more pressingly to ensure that they are prepared to act when necessary. More detailed discussion of specific amendments is outside of the scope of the present work but should be conducted within chemical organizations and societies.

I note that reinforcing existing ideals and ethical norms through modest additions and amendments is not the only course of action. Rather than expand chemical ethics to address the possible risks of foundation models (among other computational technologies), chemistry and related fields could redefine what it is to be ethical to accommodate the presence of 'AI' in our society and our professional lives. In a sense, this redefinition is actively taking place. Though none of the professional codes that I discussed in section 2.2 have been significantly amended to this effect, major chemistry conferences and publications allow scientific works based on LLMs and other foundation models with few limitations. While some publication guidelines require that authors of a manuscript disclose if and how foundation models have been used in the writing [87], the door remains open for stochastic plagiarism and misinformation, among other possible negative outcomes. It is logical on the surface to trust that a manuscript's authors will review their work to ensure accuracy and check that all ideas have been properly attributed. However, it may not be obvious at all that a model has plagiarized or repeated a falsehood.

5.3. Lingering questions

The present work is not meant to be definitive but instead was written to open a necessary and (until now) largely ignored conversation. In service of this aim, I close this article by considering what I have not been able to address and what future research—ethical and technical, theoretical and empirical—should be conducted to help push the conversation forward and resolve some of the problems posed here.

I have chosen in this work to limit my scope to the chemical sciences, focusing mainly on chemistry and chemical engineering. This is mainly a reflection of my own disciplinary comfort and background. As a

chemist without significant training in ethics or philosophy, I feel equipped to tackle ethical issues in the chemical sciences, while I recognize that I may be unequipped to address broader questions. Nonetheless, broader questioning is needed.

As a starting point, it would be worthwhile to compare the ethics of foundation models in the the chemical sciences, as discussed here, with those in the biomedical fields, where a body of work in related ethical concerns is growing [19, 186, 187]. In particular, the thought put into regulatory changes in the biomedical literature [188] may be relevant for discussions of scientific ethics. Beyond medicine, it is worth asking to what extent the ideas presented here generalize to other areas of science and engineering. Ostensibly all areas of science should be guided by the same ideals, but different professions may have different norms and different priorities. As one example, though the chemical sciences are intimately connected with and concerned for the environment (and, thus, environmental sustainability), the same may not be said for other fields, such as astronomy.

It is also worth considering how the values of (chemical) science are or are not aligned with the values of 'AI' and ML. Are they in tension? Are they compatible at all? As a starting point, one could compare the ideals listed in section 2.1 above or similar sets of scientific ideals with the (often implicit) values of ML practitioners, as identified by Birhane *et al*'s study of ML conference papers [189]. Alternatively, one could analyze codes of ethics for professional organizations in computer science, data science, and 'AI' and compare the key ethical norms and requirements in those fields to the norms identified here.

Even staying within the domain of the chemical sciences, there are many further questions worth exploring. A more extensive analysis of the chemical ethics of computational and data science work is sorely needed, as is a more extensive comparative analysis of chemical professional ethics beyond the five codes considered here. Future research should moreover consider the ethics of automated and self-driving laboratories, particularly as they pertain to safety and security. Finally, further research should be conducted to firmly assess the environmental costs associated with chemical foundation models and the relative environmental cost or benefit of generative foundation models for chemical classification and regression tasks.

While I have discussed the general practice of chemistry and associated norms, I have focused in ways and at times on chemical researchers. How, if at all, might the conclusions drawn here change from the lens of, e.g. an industrial chemical engineer, where, e.g. publication ethics and classroom education may not be important, where an individual may not have control over which tools they use (i.e. they may be required to use 'AI' tools by their employer), and where the interests of the employer frequently put practitioners at odds with ethical (especially environmental) norms?

5.4. Conclusions

Like microplastics and greenhouse gases, foundation models surround us. Now that these models, from LLMs and generative diffusion models to 'universal' MLIPs, have made their way into society and into scientific practice, the chemical profession(s) must take stock, assessing their risks and outcomes under consistent ethical standards. Here, I have begun this work, addressing the questions 'Are foundation models in alignment with scientific and chemical ethics?' and 'How should chemical practitioners interact with and around foundation models?'

To the first question, an analysis of scientific ideals and professional norms reveals a simple answer: in many cases, no. Generative foundation models such as LLMs and VLMs violate or threaten essentially all of science's ideals as well as many professional ethical norms of chemical science. While chemistry-specific foundation models, such as protein models and 'universal' MLIPs, avoid many of the ethical problems of more general models like LLMs and VLMs, they still have a negative environmental impact that is at odds with the chemical sciences' obligations to sustain and protect the natural environment for itself and for the benefit of future generations.

The second question—now that foundation models are a part of scientific practice, what should be done—is more challenging. I have argued that, where foundation models do not provide significant benefit, they should generally be avoided. This places the onus of the model users to demonstrate real need before applying a foundation model. Drawing from the green chemistry, sustainable chemistry, and battery fields, I further suggest that, in cases where foundation models are highly useful and/or essential for one's work, efforts should be taken to eliminate one's need to use said models, for instance by identifying or developing more ethical alternatives.

There is much more work to be done at the intersection of data science, chemical science, and ethics. I hope that my account inspires further theoretical and technical inquiries in this nascent and much-needed area of study.

Data availability statement

Details from the open coding analysis are available at https://github.com/CoReACTER/chem-ethics-synthesis.

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

I am financially supported by the Carnegie Bosch Institute Postdoctoral Fellowship. I would like to thank Julia Isabelle McKeown for useful discussions.

Conflicts of interest

I have no conflicts of interest to declare.

Model use

I declare that I at no point intentionally and knowingly used any generative foundation model, including but not limited to LLMs, VLMs, and image/video generators, during the process of writing this manuscript.

Software availability

This manuscript did not involve the development of new software.

Supplementary information

Definitions of key terms; qualitative methods for coding ethical codes and intermediate analysis; discussion of chemical professional ethics applied to computational chemistry and chemical data science. Cited references: [24, 26–31, 190–211].

Author contribution

Evan Walter Clark Spotte-Smith 0000-0003-1554-197X
Conceptualization (lead), Formal analysis (lead), Investigation (lead), Methodology (lead), Writing – original draft (lead), Writing – review & editing (lead)

References

[1] Brown T *et al* 2020 Language models are few-shot learners *Advances in Neural Information Processing Systems* vol 33 pp 1877–901
[2] Guo D *et al* (DeepSeek-AI) 2025 DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning (arXiv:2501.12948)
[3] Radford A *et al* 2021 Learning transferable visual models from natural language supervision (arXiv:2103.00020)
[4] Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B 2022 High-resolution image synthesis with latent diffusion models *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 10684–95
[5] Liu H, Li C, Wu Q and Lee Y J 2023 Visual instruction tuning (arXiv:2304.08485)
[6] White A D *et al* 2023 Assessment of chemistry knowledge in large language models that generate code *Digit. Discov.* **2** 368–76
[7] Ai Q, Meng F, Shi J, Pelkie B and Coley C W 2024 Extracting structured data from organic synthesis procedures using a fine-tuned large language model *Digit. Discov.* **3** 1822–31
[8] Boiko D A, MacKnight R, Kline B and Gomes G 2023 Autonomous chemical research with large language models *Nature* **624** 570–8
[9] Ruan Y *et al* 2024 An automatic end-to-end chemical synthesis development platform powered by large language models *Nat. Commun.* **15** 10160
[10] Jumper J *et al* 2021 Highly accurate protein structure prediction with AlphaFold *Nature* **596** 583–9
[11] Baek M *et al* 2021 Accurate prediction of protein structures and interactions using a three-track neural network *Science* **373** 871–6
[12] Deng B, Zhong P, Jun K, Riebesell J, Han K, Bartel C J and Ceder G 2023 CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling *Nat. Mach. Intell.* **5** 1031–41
[13] Batatia I *et al* 2024 A foundation model for atomistic materials chemistry (arXiv:2401.00096)
[14] Bender E M, Gebru T, McMillan-Major A and Shmitchell S 2021 On the dangers of stochastic parrots: can language models be too big? 🦜 *Proc. 2021 ACM Conf. on Fairness, Accountability and Transparency* pp 610–23
[15] Corrêa N K *et al* 2023 Worldwide AI ethics: a review of 200 guidelines and recommendations for AI governance *Patterns* **4** 100857
[16] Das B C, Amini M H and Wu Y 2025 Security and privacy challenges of large language models: a survey *ACM Comput. Surv.* **57** 152:1–152:39
[17] Keskinbora K H 2019 Medical ethics considerations on artificial intelligence *J. Clin. Neurosci.* **64** 277–82

[18] McLennan S, Fiske A, Tigard D, Müller R, Haddadin S and Buyx A 2022 Embedded ethics: a proposal for integrating ethics into the development of medical AI *BMC Med. Ethics* **23** 6
[19] Li H, Moon J T, Purkayastha S, Celi L A, Trivedi H and Gichoya J W 2023 Ethics of large language models in medicine and medical research *Lancet Digit. Health* **5** e333–5
[20] Ong J C L *et al* 2024 Medical ethics of large language models in medicine *NEJM AI* **1** AIra2400038
[21] Grabb D, Lamparth M and Vasan N 2024 Risks from language models for automated mental healthcare: ethics and structure for implementation (arXiv:2406.11852)
[22] National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Committee on Responsible Science 2017 *Fostering Integrity in Research* (National Academies Press)
[23] Ziman J M 1978 *Reliable Knowledge: An Exploration of the Grounds for Belief in Science* (Cambridge University Press)
[24] Kovac J 2018 *The Ethical Chemist: Professionalism and Ethics in Science* (Oxford University Press)
[25] Kovac J 2013 Reverence and ethics in science *Sci. Eng. Ethics* **19** 745–56
[26] Kovac J 2015 Ethics in science: the unique consequences of chemistry *Account. Res.* **22** 312–29
[27] American Chemical Society 2019 *The Chemical Professional's Code of Conduct—American Chemical Society* (acs.org) (available at: www.acs.org/careers/career-services/ethics/the-chemical-professionals-code-of-conduct.html)
[28] Brown L 2018 The global chemists' code of ethics: international cooperation for increased chemical security and safety *Responsible Conduct in Chemistry Research and Practice: Global Perspectives* (ACS Publications) pp 129–37
[29] American Institute of Chemical Engineers 2015 *Code of Ethics* (aiche.org) (available at: www.aiche.org/about/governance/policies/code-ethics)
[30] The Organization for the Prohibition of Chemical Weapons 2015 *Compilation of Codes of Ethics and Conduct* (available at: www.opcw.org/sites/default/files/documents/SAB/en/2015_Compilation_of_Chemistry_Codes.pdf)
[31] Organization for the Prohibition of Chemical Weapons 2015 *The Hague Ethical Guidelines* (opcw.org) (available at: www.opcw.org/hague-ethical-guidelines)
[32] Yuriev E, Wink D J and Holme T A 2024 The dawn of generative artificial intelligence in chemistry education *J. Chem. Educ.* **101** 2957–9
[33] Kipp A, Hawk N and Perez G 2024 Generating opportunities: strategies to elevate science and engineering practices using ChatGPT *Sci. Teach.* **91** 43–47
[34] Hare P M 2024 Coding with AI in the physical chemistry laboratory *J. Chem. Educ.* **101** 3869–74
[35] Weidinger L *et al* 2022 Taxonomy of risks posed by language models *Proc. 2022 ACM Conf. on Fairness, Accountability and Transparency* pp 214–29
[36] ACS Publications 2024 *ACS Publications Diversity Data Report* (pubsdiversity.acs.org) (available at: https://pubsdiversity.acs.org/data/2024/index.html)
[37] Royal Society of Chemistry 2023 *Diversity Data Report 2022* (rsc.org) (available at: www.rsc.org/globalassets/02-about-us/corporate-information/our-diversity-data/rsc-diversity-data-report-2022.pdf)
[38] Wilson D and VanAntwerp J 2021 Left out: a review of women's struggle to develop a sense of belonging in engineering *Sage Open* **11** 21582440211040791
[39] Ballenger J, Polnick B and Irby B (eds) 2016 *Women of Color in Stem: Navigating the Workforce* (Information Age Publishing)
[40] Moss-Racusin C A, Sanzari C, Caluori N and Rabasco H 2018 Gender bias produces gender gaps in STEM engagement *Sex Roles* **79** 651–70
[41] Strayhorn T L 2018 *College Students' Sense of Belonging: A Key to Educational Success for all Students* (Routledge)
[42] Rainey K, Dancy M, Mickelson R, Stearns E and Moller S 2018 Race and gender differences in how sense of belonging influences decisions to major in STEM *Int. J. STEM Educ.* **5** 1–14
[43] Master A H and Meltzoff A N 2020 Cultural stereotypes and sense of belonging contribute to gender gaps in STEM *Int. J. Gend. Sci. Technol.* **12** 152–98
[44] McClure T 2023 Supermarket AI meal planner app suggests recipe that would create chlorine gas *Guardian*
[45] Airgas 2021 *Safety Data Sheet: Chlorine* (available at: www.airgas.com/msds/001015.pdf)
[46] Maughan A E, Toberer E S and Zevalkink A 2025 Integrating large language models into the chemistry and materials science laboratory curricula *Chem. Mater.* **37** 2389–94
[47] Perez-Marin D and Pascual-Nieto I 2011 *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices* (Information Science Reference)
[48] Kidd C and Birhane A 2023 How AI can distort human beliefs *Science* **380** 1222–3
[49] Luccioni A S, Viguier S and Ligozat A-L 2023 Estimating the carbon footprint of BLOOM, a 176B parameter language model *J. Mach. Learn. Res.* **24** 1–15
[50] Samsi S, Zhao D, McDonald J, Li B, Michaleas A, Jones M, Bergeron W, Kepner J, Tiwari D and Gadepally V 2023 From words to watts: benchmarking the energy costs of large language model inference *2023 IEEE High Performance Extreme Computing Conf. (HPEC)* pp 1–9
[51] Li P, Yang J, Islam M A and Ren S 2025 Making AI less "thirsty": uncovering and addressing the secret water footprint of AI models (arXiv:2304.03271)
[52] Bluefield Research *Water for Data Centers: Market Trends and Forecasts, 2023–2030* (available at: www.bluefieldresearch.com/research/water-for-data-centers-market-trends-and-forecasts-2023-2030/) (Accessed 19 February 2025)
[53] Shehabi A, Smith S J, Hubbard A, Newkirk A, Lei N, Siddik M A B, Holecek B, Koomey J, Masanet E and Sartor D 2024 *2024 United States Data Center Energy Usage Report LBNL-2001637* Lawrence Berkeley National Laboratory
[54] Kez D A, Foley A M, Laverty D, Rio D F D and Sovacool B 2022 Exploring the sustainability challenges facing digitalization and internet data centers *J. Clean. Prod.* **371** 133633
[55] Guidi G, Dominici F, Gilmour J, Butler K, Bell E, Delaney S and Bargagli-Stoffi F J 2024 Environmental burden of United States Data Centers in the artificial intelligence era (arXiv:2411.09786)
[56] Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, Poulton A, Kerkez V and Stojnic R 2022 Galactica: a large language model for science (arXiv:2211.09085)
[57] Lu C, Lu C, Lange R T, Foerster J, Clune J and Ha D 2024 The AI scientist: towards fully automated open-ended scientific discovery (arXiv:2408.06292)
[58] Oreskes N 2004 The scientific consensus on climate change *Science* **306** 1686–1686
[59] Lynas M, Houlton B Z and Perry S 2021 Greater than 99% consensus on human caused climate change in the peer-reviewed scientific literature *Environ. Res. Lett.* **16** 114005

[60] Lightsources.org 2025 Light sources of the world (available at: https://lightsources.org/lightsources-of-the-world/) (Accessed 13 June 2025)
[61] Touvron H *et al* 2023 Llama 2: open foundation and fine-tuned chat models (arXiv:2307.09288)
[62] Waseem Z, Lulz S, Bingel J and Augenstein I 2021 Disembodied machine learning: on the illusion of objectivity in NLP (arXiv:2101.11974)
[63] Kotek H, Dockum R and Sun D 2023 Gender bias and stereotypes in large language models *Proc. ACM Collective Intelligence Conf.* pp 12–24
[64] Wan Y and Chang K-W 2024 White men lead, black women help? Benchmarking language agency social biases in LLMs (arXiv:2404.10508v3)
[65] An J, Huang D, Lin C and Tai M 2024 Measuring gender and racial biases in large language models (arXiv:2403.15281)
[66] Wilson K and Caliskan A 2024 Gender, race and intersectional bias in resume screening via language model retrieval *Proc. AAAI/ACM Conf. on AI, Ethics and Society* vol 7 pp 1578–90
[67] Li R, Kamaraj A, Ma J and Ebling S 2024 Decoding ableism in large language models: an intersectional approach *Proc. 3rd Workshop on NLP for Positive Impact* pp 232–49
[68] Phutane M, Seelam A and Vashistha A 2024 How toxicity classifiers and large language models respond to ableism (arXiv:2410.03448v1)
[69] Felkner V K, Chang H-C H, Jang E and May J 2023 WinoQueer: a community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models (arXiv:2306.15087)
[70] Dhingra H, Jayashanker P, Moghe S and Strubell E 2023 Queer people are people first: deconstructing sexual identity stereotypes in large language models (arXiv:2307.00101)
[71] Sosto M and Barrón-Cedeño A 2024 QueerBench: quantifying discrimination in language models toward queer identities (arXiv:2406.12399)
[72] Khandelwal K, Tonneau M, Bean A M, Kirk H R and Hale S A 2024 Indian-BhED: a dataset for measuring India-centric biases in large language models *Proc. 2024 Int. Conf. on Information Technology for Social Good* pp 231–9
[73] Martí L, Mollica F, Piantadosi S and Kidd C 2018 Certainty is primarily determined by past performance during concept learning *Open Mind* **2** 47–60
[74] Chang D F, Donald J, Whitney J, Miao I Y and Sahdra B 2024 Does mindfulness improve intergroup bias, internalized bias and anti-bias outcomes?: A meta-analysis of the evidence and agenda for future research *Pers. Soc. Psychol. Bull.* **50** 1487–516
[75] Yin X, Han J and Yu P S 2007 Truth discovery with multiple conflicting information providers on the web *Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 1048–52
[76] Lee H-P H, Sarkar A, Tankelevitch L, Drosos I, Rintel S, Banks R and Wilson N 2025 The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers *Proc. ACM CHI Conf. on Human Factors in Computing Systems*
[77] Stadler M, Bannert M and Sailer M 2024 Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry *Comput. Hum. Behav.* **160** 108386
[78] Zhang Y *et al* 2023 Siren's song in the AI ocean: a survey on hallucination in large language models (arXiv:2309.01219)
[79] Tonmoy S, Zaman S, Jain V, Rani A, Rawte V, Chadha A and Das A 2024 A comprehensive survey of hallucination mitigation techniques in large language models (arXiv:2401.01313)
[80] Carson T L 2006 The definition of lying *Noûs* **40** 284–306
[81] El-Mallakh R S and Walker K L 2010 Hallucinations, psuedohallucinations and parahallucinations *Psychiatry: Interpers. Biol. Process.* **73** 34–42
[82] Hicks M T, Humphries J and Slater J 2024 ChatGPT is bullshit *Ethics Inf. Technol.* **26** 1–10
[83] Frankfurt H G 2009 *On Bullshit* (Princeton University Press)
[84] Liu H, Li C, Wu Q and Lee Y J 2023 Visual instruction tuning *Advances in Neural Information Processing Systems* vol 36 pp 34892–916
[85] Sapkota R, Raza S and Karkee M 2025 Comprehensive analysis of transparency and accessibility of ChatGPT, DeepSeek, and other SoTA large language models (arXiv:2502.18505)
[86] American Chemical Society 2024 *Artificial Intelligence (AI) Best Practices and Policies at ACS Publications* (researcher-resources.acs.org) (available at: https://researcher-resources.acs.org/publish/aipolicy)
[87] Nature Portfolio 2025 *Artificial Intelligence (AI) | Nature Portfolio* (nature.com) (available at: www.nature.com/nature-portfolio/editorial-policies/ai)
[88] Ajwani R, Javaji S R, Rudzicz F and Zhu Z 2024 LLM-generated black-box explanations can be adversarially helpful (arXiv:2405.06800)
[89] Sarkar A 2024 Large language models cannot explain themselves (arXiv:2405.04382)
[90] Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A and Zhou D 2022 Self-consistency improves chain of thought reasoning in language models (arXiv:2203.11171)
[91] Huang J and Chang K C-C 2022 Towards reasoning in large language models: a survey (arXiv:2212.10403)
[92] Chen Y, Singh C, Liu X, Zuo S, Yu B, He H and Gao J 2024 Towards consistent natural-language explanations via explanation-consistency finetuning (arXiv:2401.13986)
[93] Kim S S, Vaughan J W, Liao Q V, Lombrozo T and Russakovsky O 2025 Fostering appropriate reliance on large language models: the role of explanations, sources, and inconsistencies (arXiv:2502.08554)
[94] Van Dalen H P and Henkens K 2012 Intended and unintended consequences of a publish-or-perish culture: a worldwide survey *J. Am. Soc. Inf. Sci. Technol.* **63** 1282–93
[95] Rawat S and Meena S 2014 Publish or perish: where are we heading? *J. Res. Med. Sci.* **19** 87
[96] Xue Y and Larson R C 2015 STEM crisis or STEM surplus? Yes and yes *Mon. Labor Rev.* **2015** 10–21916
[97] Algan Y, Cohen D, Davoine E, Foucault M and Stantcheva S 2021 Trust in scientists in times of pandemic: panel evidence from 12 countries *Proc. Natl Acad. Sci.* **118** e2108576118
[98] Cologna V and Siegrist M 2020 The role of trust for climate change mitigation and adaptation behaviour: a meta-analysis *J. Environ. Psychol.* **69** 101428
[99] Inan H A, Ramadan O, Wutschitz L, Jones D, Rühle V, Withers J and Sim R 2021 Training data leakage analysis in language models (arXiv:2101.05405)
[100] Huang J, Shao H and Chang K C-C 2022 Are large pre-trained language models leaking your personal information? (arXiv:2205.12628)

[101] Zheng Z, Rampal N, Inizan T J, Borgs C, Chayes J T and Yaghi O M 2025 Large language models for reticular chemistry *Nat. Rev. Mater.* **10** 1–13
[102] Zhang Q *et al* 2025 Scientific large language models: a survey on biological & chemical domains *ACM Comput. Surv.* **57** 161:1–161:38
[103] Ramos M C, Collison C J and White A D 2025 A review of large language models and autonomous agents in chemistry *Chem. Sci.* **16** 2514–72
[104] Xiao Y *et al* 2025 Protein large language models: a comprehensive survey (arXiv:2502.17504)
[105] Wang L, Li X, Zhang H, Wang J, Jiang D, Xue Z and Wang Y 2025 A comprehensive review of protein language models (arXiv:2502.06881)
[106] Guo T, Guo K, Nan B, Liang Z, Guo Z, Chawla N, Wiest O and Zhang X 2023 What can large language models do in chemistry? A comprehensive benchmark on eight tasks *Advances in Neural Information Processing Systems* vol 36 pp 59662–88
[107] Huang X, Surve M, Liu Y, Luo T, Wiest O, Zhang X and Chawla N V 2024 Application of large language models in chemistry reaction data extraction and cleaning *Proc. 33rd ACM Int. Conf. on Information and Knowledge Management* pp 3797–801
[108] Lewis P *et al* 2020 Retrieval-augmented generation for knowledge-intensive NLP tasks *Advances in Neural Information Processing Systems* vol 33 pp 9459–74
[109] Bran A M, Cox S, Schilter O, Baldassari C, White A D and Schwaller P 2024 Augmenting large language models with chemistry tools *Nat. Mach. Intell.* **6** 525–35
[110] Banerjee S, Agarwal A and Singla S 2024 LLMs will always hallucinate, and we need to live with this (arXiv:2409.05746)
[111] Xu Z, Jain S and Kankanhalli M 2024 Hallucination is inevitable: an innate limitation of large language models (arXiv:2401.11817)
[112] Liang W *et al* 2024 Mapping the increasing use of LLMs in scientific papers (arXiv:2404.01268)
[113] Kwon D 2025 Is it OK for AI to write science papers? Nature survey shows researchers are split *Nature* **641** 574–8
[114] Jablonka K M, Schwaller P, Ortega-Guerrero A and Smit B 2024 Leveraging large language models for predictive chemistry *Nat. Mach. Intell.* **6** 161–9
[115] Seifrid M, Pollice R, Aguilar-Granda A, Chan Z M, Hotta K, Ser C T, Vestfrid J, Wu T C and Aspuru-Guzik A 2022 Autonomous chemical experiments: challenges and perspectives on establishing a self-driving lab *Acc. Chem. Res.* **55** 2454–66
[116] Tom G *et al* 2024 Self-driving laboratories for chemistry and materials science *Chem. Rev.* **124** 9633–732
[117] Pannu J, Bloomfield D, MacKnight R, Hanke M S, Zhu A, Gomes G, Cicero A and Inglesby T V 2025 Dual-use capabilities of concern of biological AI models *PLOS Comput. Biol.* **21** e1012975
[118] Fivizzani K P 2005 The evolution of chemical safety training *Chem. Health Saf.* **12** 11–15
[119] Obasi L A and Osom E A 2023 Evaluating the effectiveness of safety training programs on laboratory chemical handling and compliance *J. Econ. Financ. Innov.* 115–24 (available at: https://sbtsuejournals.uz/index.php/EFI/article/view/218)
[120] Moore S *et al* 2023 Empowering education with LLMs-the next-gen interface and content generation *Int. Conf. on Artificial Intelligence in Education* pp 32–37
[121] Clark T M, Fhaner M, Stoltzfus M and Queen M S 2024 Using ChatGPT to support lesson planning for the historical experiments of Thomson, Millikan and Rutherford *J. Chem. Educ.* **101** 1992–9
[122] Du Y, Duan C, Bran A, Sotnikova A, Qu Y, Kulik H, Bosselut A, Xu J and Schwaller P 2024 Large language models are catalyzing chemistry education *ChemRxiv Preprint* https://doi.org/10.26434/chemrxiv-2024-h722v (posted online 25 June 2024, accessed 3 March 2025)
[123] Fernández A A, López-Torres M, Fernández J J and Vázquez-García D 2024 ChatGPT as an instructor's assistant for generating and scoring exams *J. Chem. Educ.* **101** 3780–8
[124] Yik B J and Dood A J 2024 ChatGPT convincingly explains organic chemistry reaction mechanisms slightly inaccurately with high levels of explanation sophistication *J. Chem. Educ.* **101** 1836–46
[125] Humphry T and Fuller A L 2023 Potential ChatGPT use in undergraduate chemistry laboratories *J. Chem. Educ.* **100** 1434–6
[126] Subasinghe S M S, Gersib S G and Mankad N P 2025 Large language models (LLMs) as graphing tools for advanced chemistry education and research *J. Chem. Educ.* **102** 1563–71
[127] Keith M, Keiller E, Windows-Yule C, Kings I and Robbins P 2025 Harnessing generative AI in chemical engineering education: implementation and evaluation of the large language model ChatGPT v3.5 *Educ. Chem. Eng.* **51** 20–33
[128] Kulik J A and Kulik C-L C 1988 Timing of feedback and verbal learning *Rev. Educ. Res.* **58** 79–97
[129] Hartman J 2008 Does class size matter? Reflections on teaching engineering economy to small and large classes *American Society for Engineering Education 2008 Annual Conf. & Exposition* pp 13–449
[130] Sun D L, Harris N, Walther G and Baiocchi M 2015 Peer assessment enhances student learning: the results of a matched randomized crossover experiment in a college statistics class *PLoS One* **10** e0143177
[131] Schultz M 2011 Sustainable assessment for large science classes: non-multiple choice, randomised assignments through a learning management system *J. Learn. Des.* **4** 50–62
[132] Kara E, Tonin M and Vlassopoulos M 2021 Class size effects in higher education: differences across STEM and non-STEM fields *Econ. Educ. Rev.* **82** 102104
[133] de la Pena A M, González-Gómez D, de la Pena D M, Gómez-Estern F and Sequedo M S 2013 Automatic web-based grading system: application in an advanced instrumental analysis chemistry laboratory *J. Chem. Educ.* **90** 308–14
[134] Aldriye H, Alkhalaf A and Alkhalaf M 2019 Automated grading systems for programming assignments: a literature review *Int. J. Adv. Comput. Sci. Appl.* **10** 215–22
[135] Dommeyer C J, Baum P and Hanna R W 2002 College students' attitudes toward methods of collecting teaching evaluations: in-class versus on-line *J. Educ. Bus.* **78** 11–15
[136] Winchester M K and Winchester T M 2012 If you build it will they come?; Exploring the student perspective of weekly student evaluations of teaching *Assess. Eval. High. Educ.* **37** 671–82
[137] Chávez K and Mitchell K M 2020 Exploring bias in student evaluations: gender, race and ethnicity *PS: Political Sci. Politics* **53** 270–4
[138] Heffernan T 2022 Sexism, racism, prejudice and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching *Assess. Eval. High. Educ.* **47** 144–54
[139] Goos M and Salomons A 2017 Measuring teaching quality in higher education: assessing selection bias in course evaluations *Res. High. Educ.* **58** 341–64
[140] Boysen G A 2015 Significant interpretation of small mean differences in student evaluations of teaching despite explicit warning to avoid overinterpretation *Scholarsh. Teach. Learn. Psychol.* **1** 150

[141] Simpson P M and Siguaw J A 2000 Student evaluations of teaching: an exploratory study of the faculty response *J. Mark. Educ.* **22** 199–213
[142] White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J and Schmidt D C 2023 A prompt pattern catalog to enhance prompt engineering with ChatGPT (arXiv:2302.11382)
[143] Arawjo I, Swoopes C, Vaithilingam P, Wattenberg M and Glassman E L 2024 ChainForge: a visual toolkit for prompt engineering and LLM hypothesis testing *Proc. 2024 CHI Conf. on Human Factors in Computing Systems* pp 1–18
[144] Bommasani R *et al* 2021 On the opportunities and risks of foundation models (arXiv:2108.07258)
[145] Schneider J, Meske C and Kuss P 2024 Foundation models: a new paradigm for artificial intelligence *Bus. Inf. Syst. Eng.* **66** 221–31
[146] Abriata L A 2024 The nobel prize in chemistry: past, present and future of AI in biology *Commun. Biol.* **7** 1409
[147] Nijkamp E, Ruffolo J A, Weinstein E N, Naik N and Madani A 2023 ProGen2: exploring the boundaries of protein language models *Cell Syst.* **14** 968–978.e3
[148] Abramson J *et al* 2024 Accurate structure prediction of biomolecular interactions with AlphaFold 3 *Nature* **630** 493–500
[149] Tai K S, Bailis P and Valiant G 2019 Equivariant transformer networks *Int. Conf. on Machine Learning* pp 6086–95
[150] Cao H, Tan C, Gao Z, Xu Y, Chen G, Heng P-A and Li S Z 2024 A survey on generative diffusion models *IEEE Trans. Knowl. Data Eng.* **36** 2814–30
[151] Yang Z, Zeng X, Zhao Y and Chen R 2023 AlphaFold2 and its applications in the fields of biology and medicine *Signal Transduct. Target. Ther.* **8** 1–14
[152] Jendrusch M, Korbel J O and Sadiq S K 2021 AlphaDesign: a *de novo* protein design framework based on AlphaFold *bioRxiv Preprint* https://doi.org/10.1101/2021.10.11.463937 (posted online 12 October 2021, accessed 11 May 2025)
[153] Goverde C A, Wolf B, Khakzad H, Rosset S and Correia B E 2023 *De novo* protein design by inversion of the AlphaFold structure prediction network *Protein Sci.* **32** e4653
[154] Kim A-R, Hu Y, Comjean A, Rodiger J, Mohr S E and Perrimon N 2024 Enhanced protein-protein interaction discovery via AlphaFold-Multimer *bioRxiv Preprint* https://doi.org/10.1101/2024.02.19.580970 (posted online 21 February 2024, accessed 11 May 2025)
[155] Lee C Y *et al* 2024 Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation *Mol. Syst. Biol.* **20** 75–97
[156] Tian R, Li Y, Wang X, Li J, Li Y, Bei S and Li H 2022 A pharmacoinformatics analysis of artemisinin targets and *de novo* design of hits for treating ulcerative colitis *Front. Pharmacol.* **13** 843043
[157] Kobakhidze G, Sethi A, Valimehr S, Ralph S A and Rouiller I 2022 The AAA+ ATPase p97 as a novel parasite and tuberculosis drug target *Trends Parasitol.* **38** 572–90
[158] Collar A L, Linville A C, Core S B and Frietze K M 2022 Epitope-based vaccines against the *Chlamydia trachomatis* major outer membrane protein variable domain 4 elicit protection in mice *Vaccines* **10** 875
[159] Pak M A, Markhieva K A, Novikova M S, Petrov D S, Vorobyev I S, Maksimova E S, Kondrashov F A and Ivankov D N 2023 Using AlphaFold to predict the impact of single mutations on protein stability and function *PLoS One* **18** e0282689
[160] Shoghi N, Kolluru A, Kitchin J R, Ulissi Z W, Zitnick C L and Wood B M 2023 From molecules to materials: pre-training large generalizable models for atomic property prediction (arXiv:2310.16802)
[161] Baker C M 2015 Polarizable force fields for molecular dynamics simulations of biomolecules *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **5** 241–54
[162] Harrison J A, Schall J D, Maskey S, Mikulski P T, Knippenberg M T and Morrow B H 2018 Review of force fields and intermolecular potentials used in atomistic computational materials research *Appl. Phys. Rev.* **5** 031104
[163] Bedrov D, Piquemal J-P, Borodin O, MacKerell A D Jr, Roux B and Schröder C 2019 Molecular dynamics simulations of ionic liquids and electrolytes using polarizable force fields *Chem. Rev.* **119** 7940–95
[164] Yang J, Chen Z, Sun H and Samanta A 2023 Graph-EAM: an interpretable and efficient graph neural network potential framework *J. Chem. Theory Comput.* **19** 5910–23
[165] Zeng J, Giese T J, Zhang D, Wang H and York D M 2025 DeePMD-GNN: a DeePMD-kit plugin for external graph neural network potentials *J. Chem. Inf. Model.* **65** 3154–60
[166] Jain A *et al* 2013 Commentary: The materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002
[167] Chanussot L *et al* 2021 Open catalyst 2020 (OC20) dataset and community challenges *ACS Catal.* **11** 6059–72
[168] Levine D S *et al* 2025 The open molecules 2025 (OMol25) dataset, evaluations, and models (arXiv:2505.08762)
[169] Cai F, Hanna K, Zhu T, Tzeng T-R, Duan Y, Liu L, Pilla S, Li G and Luo F 2025 A foundation model for chemical design and property prediction (arXiv:2410.21422)
[170] Desislavov R, Martínez-Plumed F and Hernández-Orallo J 2023 Trends in AI inference energy consumption: beyond the performance-vs-parameter laws of deep learning *Sustain. Comput.: Inform. Syst.* **38** 100857
[171] Luccioni S, Jernite Y and Strubell E 2024 Power hungry processing: watts driving the cost of AI deployment? *Proc. 2024 ACM Conf. on Fairness, Accountability and Transparency* pp 85–99
[172] Sherman J, Chin B, Huibers P, Garcia-Valls R and Hatton T A 1998 Solvent replacement for green processing *Environ. Health Perspect.* **106** 253–71
[173] Vethaak A D and Legler J 2021 Microplastics and human health *Science* **371** 672–4
[174] Li W C, Tse H F and Fok L 2016 Plastic waste in the marine environment: a review of sources, occurrence and effects *Sci. Total Environ.* **566** 333–49
[175] Mann E 2017 Digital technology is dependent on forced labor: the exploitative labor practices of cobalt extraction in the Democratic Republic of Congo *Appl. Anthropol.* **37** 25
[176] Calvão F, Mcdonald C E A and Bolay M 2021 Cobalt mining and the corporate outsourcing of responsibility in the Democratic Republic of Congo *Extr. Ind. Soc.* **8** 100884
[177] Jessop P G 2011 Searching for green solvents *Green Chem.* **13** 1391–8
[178] Bubalo M C, Vidović S, Radojčić Redovniković I and Jokić S 2015 Green solvents for green technologies *J. Chem. Technol. Biotechnol.* **90** 1631–9
[179] Croy J R, Long B R and Balasubramanian M 2019 A path toward cobalt-free lithium-ion cathodes *J. Power Sources* **440** 227113
[180] Wang M, Chen X, Yao H, Lin G, Lee J, Chen Y and Chen Q 2022 Research progress in lithium-excess disordered rock-salt oxides cathode *Energy Environ. Mater.* **5** 1139–54
[181] Chen X, Chen F, Jiang H, Wang J, Li Y X and Wang G 2023 Replacing plastic with bamboo: eco-friendly disposable tableware based on the separation of bamboo fibers and the reconstruction of their network structure *ACS Sustain. Chem. Eng.* **11** 7407–18

[182] Brodin M, Vallejos M, Opedal M T, Area M C and Chinga-Carrasco G 2017 Lignocellulosics as sustainable resources for production of bioplastics—a review *J. Clean. Prod.* **162** 646–64
[183] Rujnić-Sokele M and Pilipović A 2017 Challenges and opportunities of biodegradable plastics: a mini review *Waste Manag. Res.* **35** 132–40
[184] Coates G W and Getzler Y D Y L 2020 Chemical recycling to monomer for an ideal, circular polymer economy *Nat. Rev. Mater.* **5** 501–16
[185] Demarteau J *et al* 2023 Biorenewable and circular polydiketoenamine plastics *Nat. Sustain.* **6** 1426–35
[186] Harrer S 2023 Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine *eBioMedicine* **90** 104512
[187] Pool J, Indulska M and Sadiq S 2024 Large language models and generative AI in telehealth: a responsible use lens *J. Am. Med. Inform. Assoc.* **31** 2125–36
[188] Ong J C L *et al* 2024 Ethical and regulatory challenges of large language models in medicine *Lancet Digit. Health* **6** e428–32
[189] Birhane A, Kalluri P, Card D, Agnew W, Dotan R and Bao M 2022 The values encoded in machine learning research *Proc. 2022 ACM Conf. on Fairness, Accountability and Transparency* pp 173–84
[190] Glaser B G 2016 Open coding descriptions *Grounded Theory Rev.* **15** 108–10
[191] Lincoln Y S and Guba E G 1985 *Naturalistic Inquiry* vol 75 (SAGE)
[192] Mehlich J, Moser F, Van Tiggelen B, Campanella L and Hopf H 2017 The ethical and social dimensions of chemistry: reflections, considerations and clarifications *Chem. Eur. J.* **23** 1210–8
[193] Brown T L 2009 *Chemistry: The Central Science* (Pearson Education)
[194] Bertozzi C R 2015 The centrality of chemistry *ACS Cent. Sci.* **1** 1–2
[195] Hoffmann R 1995 *The Same and not the Same* (Columbia University Press)
[196] Chamizo J A and Ortiz-Millán G 2024 Ethics of the future of chemical sciences *Found. Chem.* 1–11
[197] Reed J W 1992 Analysis of the accidental explosion at PEPCON, Henderson, Nevada, on May 4, 1988 *Propellants Explos. Pyrotech.* **17** 88–95
[198] Funabashi H 2006 Minamata disease and environmental governance *Int. J. Japan. Sociol.* **15** 7–25
[199] Vargesson N 2015 Thalidomide-induced teratogenesis: history and mechanisms *Birth Defects Res. C* **105** 140–56
[200] Ladd J 2017 Bhopal: an essay on moral responsibility and civic virtue *Engineering Ethics* (Routledge) pp 153–71
[201] Puleo S 2019 *Dark Tide: The Great Boston Molasses Flood of 1919* (Beacon Press)
[202] Prugh R W 2020 Historical record of ammonium nitrate disasters *Process Saf. Prog.* **39** e12210
[203] Al-Hajj S, Mokdad A H and Kazzi A 2021 Beirut explosion aftermath: lessons and guidelines *Emerg. Med. J.* **38** 938–9
[204] Gamito M C and Marsden C T 2024 Artificial intelligence co-regulation? The role of standards in the EU AI Act *Int. J. Law Inf. Technol.* **32** eaae011
[205] Hacker P 2024 Sustainable AI regulation *Common Mark. Law Rev.* **61** 345–86
[206] Hoffmann R 2015 Tension in chemistry and its contents *Account. Res.* **22** 330–45
[207] Schummer J 2001 Ethics of chemical synthesis *Hyle: Int. J. Phil. Chem.* **7** 103–24
[208] Rudin C 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead *Nat. Mach. Intell.* **1** 206–15
[209] Samek W and Müller K-R 2019 Towards explainable artificial intelligence *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer) pp 5–22
[210] Foster K R, Vecchia P and Repacholi M H 2000 Science and the precautionary principle *Science* **288** 979–81
[211] Bensaude-Vincent B and Simon J 2012 *Chemistry: The Impure Science* (World Scientific)