Contribution ID: **17** Type: **not specified**

AI-Enhanced ErUM: Four New Epistemic and Ethical Considerations

Thursday 31 July 2025 11:30 (30 minutes)

Scientists are already using generative AI in their research (Furze, 2025; Kwon, 2025; Nazir & Wang, 2023). But while there is extensive and ongoing research into the ethics of AI throughout its lifecycle (Coeckelbergh, 2020), and some research about the ethics of AI-related jobs and how certain job tasks might use AI (Chance & Hammersley, n.d.; Gray & Surrey, 2019; Perrigo, 2023; Williams et al., 2022), relatively little of the research has to do specifically with use of AI to augment knowledge work (Kulkarni et al., 2024; Nah et al., 2023; Resnik & Hosseini, 2024). In scientific research in particular, important critiques charge AI use with risking fabrication of vital data and citations, undermining transparency in data collection and use, and complicating the very nature of authorship (Resnik & Hosseini, 2024).

However, no one has yet investigated the specific ethical issues surrounding use of AI research in the field of what is known in German as "ErUM"("*Erforschung von Universum und Materie*"), i.e., research into the universe and matter. Such research comprises several fields of the natural sciences, including astrophysics, particle physics, nuclear physics, hadron and ion physics, and photon and neutron science. This paper therefore offers a preliminary ethical assessment of the ethics of AI-assisted ErUM. We find that even setting aside traditional concerns about AI use in scientific research in general, AI-enhanced ErUM in particular raises four specific, ethically salient problems, all of which are related to the long-term sustainability of AI-enhanced ErUM research.

The first problem is that fabrication (sometimes called "hallucination") may be especially difficult to detect in ErUM compared to other sciences and to non-scientific research fields, given the required investments of money and expertise of running reliable ErUM experiments. Put briefly, while LLMs'fabrication rates have declined greatly in the last few years (Vectara, n.d.), it is obviously very important that specific scientific claims made in a manuscript be accurate. In other areas of scholarly research, fabrication is either not a problem (because the research is fully creative in nature), easy to detect (because the research's main evidence is explicitly, publicly presented in the document for anyone to inexpensively evaluate), or at least open to testing by other, interested parties. But because much of ErUM is, by its nature, relatively expensive and abstruse, fabrications may be difficult to detect. Replicating large experiments such as those at CERN, FAIR, or XFEL is practically impossible for almost everyone (Junk & Lyons, 2020), so fabricated interpretations or subtly altered conclusions, especially in internal reports or AI-drafted analyses, may go unchallenged for years. Similarly, raw data in ErUM are typically vast, complex, and not publicly accessible, or only shared within limited collaborations. Even the collaborators rely on multi-layered data-processing pipelines, with specialized software and calibration tools that only a subset of the team understands (cf. Neves et al., 2011; Rumsey, 2025; Werum, n.d.). And while this is more controversial, arguably, some ErUM subfields (e.g., dark-matter searches, early-universe cosmology, string theory, and neutrino physics) involve interpretations under conditions of uncertainty. Small anomalies are often the bases of tentative theoretical frameworks, and new theories are sometimes expected to rest, at first, on minimal or indirect empirical signals.

Second, and related, what is sometimes called "model collapse"is especially dangerous in ErUM, because so much ErUM is concentrated in a few major institutions, and so the demand for data may outstrip the production of novel research. Model collapse occurs when an LLM is trained, at least partially, on LLM-generated data (cf. Crotty, 2024; Shumailov, 2024; Sun et al., 2024). This is a problem in general, but because cutting-edge ErUM (e.g., collider results, XFEL imaging, and neutrino events) requires highly specialized and expensive equipment, and is generated at a relatively small number of sites (e.g., CERN, DESY, and FAIR), only a few collaborations and labs occupy the pipelines from raw data to public interpretation. So, the number of available human-written interpretations of ErUM data is already very limited compared to fields such as psychology, sociology, and even molecular biology. If even a small fraction of the LLM training corpus on ErUM comes from derivative or LLM-generated summaries of these limited sources, recursive contamination is far more likely, and harder to detect, because replication is nearly impossible. Beyond this, fields such as physics and cosmology have outsized cultural and epistemic influence. Discoveries such as the Higgs boson, gravitational waves, and the Big Bang model are frequently "reprocessed"in popular science, TED talks, and Wikipedia (cf. Levinson, 2013). But LLMs are disproportionately trained on those popularized and digested forms of ErUM. If even a few flawed, AI-generated articles inflect these outlets, they are more likely to get reabsorbed into

training corpora and thereby overweight the epistemic authority of incorrect interpretations.
Third, and also related, given that ErUM is adjacent to the foundations of physics, the use of AI to describe and interpret research results may activate its well-known plausibility bias (Agarwal et al., 2024), exacerbate LLMs'lack of transparency, and have other misleading effects. So, when asked to summarize or interpret foundational ErUM, LLMs may favor familiar explanatory tropes (e.g., "curved spacetime,""particle-wave duality,""the fabric of the universe,"and "uncertainty principle") and reproduce consensus-sounding narratives, even where the field remains deeply unsettled. Relatedly, because LLMs specialize in semantic mimicry, their statements may miss subtle distinctions in ontological assumptions and theoretical commitments. They may unwittingly conflate rival interpretations, misrepresent theories'scopes, or invent "bridges"between incompatible frameworks, for the sake of plausibility.
Fourth, because ErUM commonly requires wide-ranging international collaborations, injecting AI writing and interpretation into internal documents and translations may undermine accountability and introduce further fabrications. ErUM is sometimes conducted by massive international consortia involving hundreds to thousands of researchers (cf. Abbott et al., 2012; ATLAS Collaboration et al., 2012), spread across many countries and time zones (e.g., ATLAS, MCS, IceCube, and FAIR). Internal communication (such as memos, reports, logs, and drafts) often passes through many hands and is written collaboratively. Beyond this, international ErUM teams often operate in multiple working languages and rely on AI translation tools to write or interpret internal communications. When precision is critical, subtle mistranslations can alter apparent meanings. Indeed, LLMs are known to fabricate or "smooth over"unclear concepts, especially in technical domains. And because scientific collaboration relies not only on accuracy but also on understanding who knows what (and to what degree of confidence)—a sort of "epistemic map"of the team—AI-generated text may mask the human judgment behind claims. Finally, large-scale ErUM projects require extensive documentation, which is often delegated to staff or early-career researchers, and increasingly, to LLMs. For all these reasons, there is danger of mistranslation, miscommunication, and semantic drift, and these errors can propagate into the literature or infrastructure plans.
This paper concludes by issuing some recommendations, based on these ethical problems, for ethically responsible use of AI in ErUM. Put briefly, ErUM researchers who use AI should adopt these five best practices:
• AI-provenance tagging in publications and preprints: ErUM authors should clearly indicate which portions of a document, and which datasets, were produced, edited, or otherwise modified by AI. Authors should include prompts and source chains as supplementary files.
• Independence in summaries: ErUM researchers should avoid using LLMs to evaluate or summarize papers that the LLMs themselves may have been trained on, especially in grant or peer review.
• AI-verification workflows: ErUM researchers should build institutional workflows with checklists or signoff procedures for AI-generated sections, key-claim verification, and audits. They should also create logs or registries of AI-generated content.
• Epistemic qualifiers: Authors should mark speculative or inferential language with qualifiers such as "tentative,"and discourage AIs'being used to explain anomalous results, unless a human has first formulated or vetted the anomaly.
• General caution: ErUM researchers and journalists should treat ErUM as an "epistemic high-risk zone"for fabrication, plausibility bias, and recursive contamination, and should support research into domain-specific collapse detection.
These practices can help maintain the promise of AI-assisted ErUM research while minimizing the principal ethical and epistemic dangers.
Given the undeniable promise of AI in academic research, it would be a mistake to completely boycott or forgo the benefits of AI assistance, including in ErUM. By being aware of the domain-specific dangers of AI-enhanced ErUM research, scientists can enhance their understanding of the universe while investing in a reliable and solid foundation for the future of ErUM.

Sustainability

Ethics

ai-enhanced research, artificial intelligence, ai erum research, research ethics, academic ethics

Primary author: METCALF, Thomas (Sustainable AI Lab, Institute for Science and Ethics, University of

Bonn)

Presenter: METCALF, Thomas (Sustainable AI Lab, Institute for Science and Ethics, University of Bonn)

Session Classification: Ethics