Comparing Human and LLM Ethical Analyses: A Case Study in Computational Social Science Research

Spencer Phillips Hey[1], **Julie Walsh**[2], **Eni Mustafaraj**[3]
[1]Hey Research & Innovation
[2]Department of Philosophy, Wellesley College
[3]Department of Computer Science, Wellesley College
heyspencer@gmail.com, julie.walsh@wellesley.edu, eni.mustafaraj@wellesley.edu

Abstract

As researchers increasingly engage with ethically complex digital phenomena, timely and accessible support for ethical reflection is essential—yet often unavailable beyond formal institutional review processes, which are more focused on regulatory compliance than ethics. This paper investigates the potential of large language models (LLMs) to serve as research ethics support tools by providing immediate, context-sensitive feedback on draft research protocols. We analyze a draft research proposing to scrape digital platforms for data on "Sephora Kids"—a trend in which minors promote beauty products on platforms like YouTube and TikTok—as a case study to explore this possibility. Two human ethicists and two LLMs (GPT-4o and Claude 3.7 Sonnet) independently reviewed the proposal and produced ethical evaluations. We then compared the outputs to assess whether LLMs could meaningfully assist researchers in identifying and engaging with ethical issues. Our findings suggest that LLMs can already offer valuable support.

Introduction

As generative artificial intelligence (GenAI) systems become increasingly sophisticated and available for wider use as research instruments, ethical oversight of this use will become an increasingly critical component of responsible research practices. The application of GenAI to computational social science research provides a striking example of this (Bail 2024). GenAI can be used to simulate social interactions, generate synthetic populations, create content for experimental stimuli, or assist in the analysis of large-scale social data. These capabilities offer researchers powerful new methods to model, understand, and even predict human behavior. It has never been easier to collect and analyze data on human behavior, at scale and without explicit consent. But as often happens, the power of these new scientific tools has outpaced established frameworks for practical ethical review (Lazer et al. 2020; Prem 2023). The line between real and synthetic data is increasingly blurred, raising questions about authenticity, privacy, and the boundaries of informed consent. Researchers working in this space must therefore confront an array of moral challenges (Jeon, Kim, and Park 2025) and, unfortunately, most computational social science researchers receive little formal training in the philosophical foundations of ethics that are necessary to successfully navigate uncharted moral terrain.

The social risks and harms due to the gap between technical feasibility and ethical literacy may be greatest in the case of studies involving children and digital platforms (Hokke et al. 2018). Consider the rise of the "Sephora kids" phenomenon—which refers to minors who document the use of beauty products on platforms like YouTube and TikTok: This raises ethical questions around exploitation, identity formation, and the commercialization of childhood (Madhumita and Ponnarasu 2025). As such, studying the phenomenon will be critical for understanding digital youth culture and the power of digital content platforms. But conducting research on this population, and using GenAI tools to do so, raises a web of ethical challenges—concerning protection for the vulnerability of the study population, privacy, data security and integrity, etc.—that standard institutional review processes may be unprepared to resolve.

Yet, alongside the challenges, there will also be new opportunities for GenAI to improve the conduct of science. Recent advances in large language models (LLMs) have opened new avenues for integrating human-centered AI into ethical reasoning workflows (Watkins 2024). These models, trained on massive corpora of human-generated text, have demonstrated impressive performance on tasks that are core to social science research, including text summarization, classification, and argumentation (Ziems et al. 2024). Moreover, some empirical studies have suggested that there is reasonable alignment of moral values between humans and LLMs (Norhashim and Hahn 2024). In light of these challenges and opportunities, we formulated the following research question: *Can LLMs be used as tools to help scientists better identify and reflect on ethical issues in computational social science research, particularly for projects that operate in ethically fraught domains?*

In what follows, we describe the setup and results of a pilot study intended to shed light on this question. Assessing the quality of a research ethics analysis is not a straightforward matter in itself as there may be many "correct" or credible analyses and trained human ethicists or ethics committees can vary widely in their judgments, particularly where novel experimental methods or modalities are involved (Taljaard et al. 2014; Mariani et al. 2023). We

therefore sought to explore what could be considered a baseline case for practical use of LLMs in science, modeling the conditions of **an ethics consult**—i.e., a situation where a researcher would submit a draft research proposal to an ethicist or ethics committee, seeking feedback on the ethical issues raised by their study and ways to address them. To do this, we adapted a computational social science research proposal that described an experiment to automate the scraping and analysis of YouTube and TikTok content related to the "Sephora kids" trend. We gave this proposal to two trained ethicists, who independently produced an ethics review report, identifying what they saw as the major ethical issues and offering suggestions for the proposal's improvement. We then prompted two commercial LLMs—GPT-4o and Claude 3.7 Sonnet—to produce their own evaluations following the same instructions as the human evaluators. By analyzing the results, including the overlap and outlier issues between the human and AI assessments, this experiment helps to illuminate the potential of LLMs to serve as tools for ethical reflection in science.

To our best knowledge, this study is novel in the following ways: (a) we provide a direct empirical comparison between human ethics reviewers and general-purpose LLMs for a real-world research project proposal; (b) our case study, with its focus on social media data related to minors, introduces unique ethical considerations; and (c) we assess general-purpose LLMs without domain-specific fine-tuning for ethical evaluations, since they are currently available to all researchers.

Prior Research

Evolution of LLMs as Research Tools

In the past few years, we have seen a rapid integration of LLMs into scientific workflows, stimulating both excitement and concern about how these tools may reshape research practices (Watkins 2024; Ziems et al. 2024). LLMs are already being deployed in support of research tasks ranging from automated literature reviews (Scherbakov et al. 2024) to hypothesis generation (Manning, Zhu, and Horton 2024) and code debugging (Jiang et al. 2024). For example, agentic systems and related frameworks have demonstrated the capacity to autonomously gather and synthesize relevant literature on a given topic, offering researchers an unprecedented level of support in the early phases of inquiry (Sami et al. 2024). These developments suggest that LLMs can serve as valuable tools throughout the research process.

But despite their utility, the increasing reliance on LLMs within scientific workflows also raises serious concerns. These models, while powerful, are not infallible: they are prone to generating plausible-sounding but inaccurate or fabricated information, a phenomenon often described as "hallucination" (Yao et al. 2023). This risk is particularly acute in tasks involving the synthesis or interpretation of complex literature, where subtle errors can propagate misleading narratives or distort theoretical insights. Moreover, LLMs inherit and potentially amplify the biases embedded in their training data, which can manifest in the framing of research questions, the selection of sources, or the language used in summarizations (Słowik and Bottou 2021). In computational social science, where research often touches on sensitive sociopolitical issues, such biases can skew findings in ways that are difficult to detect without careful scrutiny (Olteanu et al. 2019). The opacity of model outputs also complicates reproducibility and accountability—two core tenets of scientific integrity (Barman, Wood, and Pawlowski 2024).

LLMs and Ethical Analysis

But is it possible that LLMs could be used to help address these problems? A growing body of work is beginning to examine whether LLMs could assist with research ethics review, suggesting that LLMs might help researchers draft Institutional Review Board (IRB) submissions, clarify regulatory requirements, or even automate aspects of ethical review (Sridharan and Sivaramakrishnan 2025). Some scholars have gone so far as to imagine LLM-based systems that could augment—or partially replace—conventional IRBs, arguing that LLM-based systems could provide consistent, fast, and context-sensitive ethical evaluations (Godwin et al. 2024), overcoming many of the long-standing criticisms of the IRB system (Keith-Spiegel, Koocher, and Tabachnick 2006; Klitzman 2015). While such visions remain speculative, they signal a broader interest in using GenAI to improve the ethical rigor of research.

There is another body of empirical work exploring the extent to which humans and LLMs agree in their moral judgments across a range of prompts or scenarios, not limited to the domain of research ethics (Garcia, Qian, and Palminteri 2024; Rathje 2024). Several studies have investigated how or whether LLMs appear to leverage familiar moral theories and values in their outputs (Chun and Elkins 2024; Norhashim and Hahn 2024). As might be expected, there is evidence of cultural bias in the moral/ethical outputs from LLMs (Chun and Elkins 2024). Yet, there is also critique of this literature, arguing that trying to match "value alignment" between humans and LLMs is fundamentally wrong-headed for assuming that LLMs possess anything that should properly be considered to be moral or ethical values (Bender et al. 2021; Rathje 2024).

LLMs as Tools to Improve Science

We would tend to agree with the critics that attempting to probe or measure an LLM's values is not a productive enterprise. However, that line of inquiry is ultimately orthogonal to the more practical line of inquiry (which we favor) that explores how LLMs can serve most effectively as tools to support humans in their ethical reasoning. But of course, this approach begs the question of whether it is ethical to use LLMs in scientific research at all. Opinions on this subject are not unanimous and vary across scientific domains (Grossmann et al. 2023; Li et al. 2023). However, one recent survey found that 81% of researchers have already begun incorporating LLMs into their workflows in some way (Liao et al. 2024). This study also found greater use and perceived benefits among researchers from historically disadvantaged groups (e.g., non-White researchers), suggesting that LLM tools may already be functioning as instruments in support

of a more equitable scientific enterprise (Liao et al. 2024). Although widespread use of LLMs for scientific research does not itself justify the practice, it does show that the scientific community is interested, perhaps even eager, to explore the potential utility.

Our study sits at the intersection of all these bourgeoning lines of inquiry. We view LLMs not as replacements for human judgment, but as potentially powerful tools for expanding ethical awareness within the scientific enterprise. The present study is thus an exploration of that vision. By comparing LLM-generated ethical assessments with those produced by ethicists, we aim to evaluate the feasibility of this vision and help to clarify what kinds of support such models can meaningfully provide.

Data & Methods

Research Proposal Case Study

For this study, we adapted a research proposal by a computational social science research group led by one of the co-authors of this paper. The proposal was to study how the language about beauty and makeup products in social media content directed at, or generated by, young girls has evolved over time, by comparing the period prior and after 2020 (the pandemic year). Using custom Python scripts, the researchers would scrape videos and metadata associated with hashtags like #tweenbeauty and similar ones from both YouTube and TikTok platforms.

Given the study's interest on language evolution over time, the team would try to get transcripts for the videos through either using the YouTube API to automatically retrieve them from YouTube or by using a locally-installed version of the speech-to-text Whisper library to transcribe TikTok videos that were available for download, based on users' permissions.[1]

The proposed analysis would follow an adapted Computational Grounded Theory (CGT) approach. Topic modeling (LDA and BERTopic) would identify recurring themes in the content of the transcripts, which would then be refined through human interpretation and generative AI (ChatGPT-4.0). Final thematic labels would be validated using SetFit, a few-shot learning classifier, allowing the team to examine trends in promotional, harmful, or unrealistic beauty messaging over time. No personal identifiers beyond public usernames would be collected, and all data would be securely stored and de-identified, prior to the analysis.

It is worth pointing out here that this research proposal was also submitted to the IRB of the corresponding institution and was approved with an exempt status, after the IRB staff ensured that all precautions were in place to avoid any potential harm. While we cannot share here the text of the research proposal, due to intellectual property priorities, both the human reviewers and the LLMs had access to the complete proposal that amounted to circa 1,700 words.

[1]TikTok users have a choice to prohibit the download of their videos.

Human Ethical Review

To establish a baseline for human ethical reflection on this research proposal, two trained ethicists independently reviewed the proposal document and provided a written ethical evaluation according to a structured review protocol, see Table 1. One human reviewer ("HR1") has greater than ten years of experience working as a research scientist and research ethicist in the domain of clinical trials. The other human reviewer ("HR2") is a philosophy professor with greater than five years experience serving on institutional ethics committees. Both HR1 and HR2 are philosophy Ph.D.'s and have been certified through typical, academic research ethics training modules (Braunschweiger and Goodman 2007). Their backgrounds are reflective of the kinds of expertise and experience considered generally sufficient to serve on an IRB or provide a credible research ethics consult.

We created our structured review protocol in an effort to "level the playing field" between human and LLM reviewing. We sought to simultaneously provide (1) instructions to the human reviewers that would strike a reasonable middle ground between a completely open-ended assignment (which is common for professional ethics consults, where the assumption is that the ethicist has mastered or internalized all relevant ethics and regulations) and an explicit checklist-based assignment (which is more common for IRB members), and (2) instructions that could be easily adapted into a prompt that would work well with the two commercial LLMs.

Nevertheless, we recognize that what counts as "working well" in the case of an ethical analysis—whether produced by a human or an LLM—is difficult to define and part of what is at issue in this experiment. Indeed, there is no fully objective or quantitative standard for what can be considered a high-quality or credible ethics review. Our study is also not a hypothesis-testing experiment, but rather an exploratory experiment, intended to generate pilot data that might guide a larger-scale, validation experiment (should the LLM output look promising).

The goal of our review protocol was thus to guide reviewers to assess the research proposal in terms of core research ethics principles such as informed consent, potential harms to subjects, data privacy, researcher obligations, and the broader societal implications of the work. Reviewers were encouraged to identify both strengths and weaknesses in the proposal's ethical design, and to articulate specific concerns or recommendations. But reviewers could, of course, deviate from these instructions if they believed it important, which is common in ethics consult and represents a flexibility of analysis that is both a strength and a limitation of any ethics review.

LLM-Generated Evaluations

To generate AI-based ethical evaluations, we used two commercial large language models: GPT-4o and Claude 3.7 Sonnet, both through the web-based, chat interface. We used the free version of each model, as this would represent some of the most accessible LLM interactions available to scientists.

Purpose
You are being asked to review a research study protocol and provide a structured ethical analysis. Your responses will be compared to those generated by an AI language model, so it's important to follow this process as consistently as possible.

Step-by-Step Evaluation Protocol

1. Read the Research Protocol Thoroughly
   (a) Take notes as needed, but do not begin your formal write-up yet.
   (b) Focus on understanding the study's objectives, design, methodology, population, interventions, and context.
2. Identify Key Ethical Dimensions
   (a) List the main ethical issues or concerns raised by the study.
   (b) Use headings or bullets if helpful. Common categories include (but do not necessarily use or limit yourself to these): Informed consent, Risk/benefit ratio, Privacy/confidentiality, Use of vulnerable populations, Scientific validity, Fair subject selection, Conflict of interest.
3. Describe Each Ethical Concern in Detail
   (a) Explain why it is relevant in the context of this specific study.
   (b) Evaluate how well the protocol addresses the issue.
   (c) Propose improvements or safeguards, if necessary.
4. Be Concise but Comprehensive
   (a) Aim for a total response length of 500-1000 words.
   (b) Prioritize clarity, structure, and actionable feedback.
5. Do Not Use External Resources
   (a) Base your review solely on the protocol provided.
   (b) Use your professional training and reasoning, not web searches or institutional precedents.

Table 1: Review Protocol for Human Ethics Evaluators

Each model was prompted with an upload of the research proposal and provided with a version of the same structured review protocol given to the human ethicists, modified only to ensure prompt formatting compatibility. The full prompt for the LLMs is provided in Appendix A.

To capture variability in model outputs, we ran each model five times with identical prompts (five runs for GPT-4o and five for Claude 3.7 Sonnet), yielding a total of ten LLM-generated reports. Each report was saved for subsequent qualitative and quantitative analysis. The full reports, for both human and LLM reviews, are available in the Supplemental Materials.

Comparative Analysis

After collecting all reports, the authors collaboratively reviewed and discussed them. Two authors independently conducted a classification process, resolving all disagreements through discussion and consensus, to identify the distinct ethical issues raised across all twelve reports (2 human, 10 AI). Although all reviewers were encouraged to provide major headings for the issues they identified, conceptually similar issues might have different headings or multiple issues might fall under one heading. Therefore, issues were categorized into conceptual clusters (e.g., data privacy, vulnerable study population, informed consent) to facilitate comparison. Each identified issue was coded for its presence or absence in each report, enabling us to quantify patterns of overlap and outliers among the human and LLM reviews.

Results

Human Ethics Review

The human reviewers identified a combined eleven issues in their ethical review report. HR1 identified five total issues, HR2 identified ten total issues. However, it is important to emphasize that the quality of an ethical review is not a function of the quantity of issues. A high-quality ethics review could identify zero issues, if indeed, there are no relevant ethical issues worth surfacing. Therefore, in describing and evaluating the quality of an ethics review report, it is important to focus on the content of the review.

The human reviewers identified three issues in common:

1. **Bias or noise in data collection**: the need to mitigate distortions in data that can compromise validity.

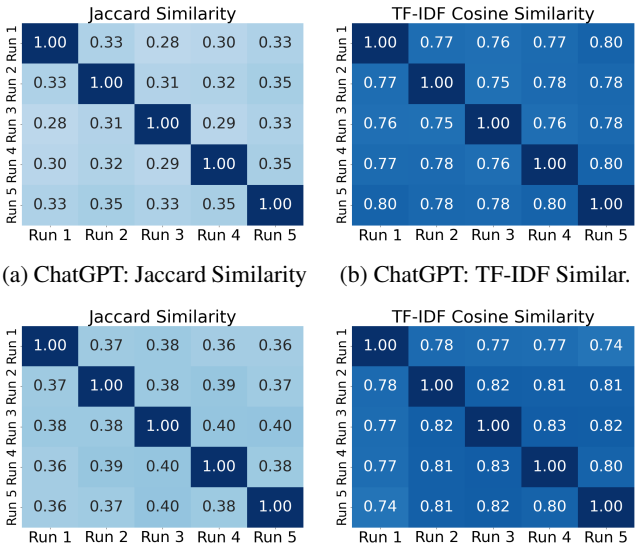
2. **Privacy protection for individuals**: the need to prevent the identification or misuse of personally identifiable information within datasets.
3. **Vulnerability of study population**: the need to addressing additional protections required when research involves participants at elevated risk of harm or exploitation.

HR1 raised two further issues in their report. One concerned the **alignment between the research question and the proposed methods**. This speaks to the fundamental scientific validity of the study, since inappropriate methods will not produce data sufficient to answer a research question. And a study that fails to answer its research question has failed to offset its costs and burdens with commensurate gains in scientific knowledge. The second concerned **whether this study technically involved human subjects at all**. They observed that it is not clear that the proposed methods to scrape publicly available data from popular digital platforms (without direct intervention or collecting any private information about individuals) involves any human subjects who would need to be consented.

HR2 raised seven additional issues in their report:

1. **Bias in subject selection**: the need to avoid discriminatory inclusion or exclusion practices that could skew results or reinforce social inequities.
2. **Harm to community**: the need to prevent potential negative consequences for communities represented or affected by the research.
3. **Justified waiver of informed consent**: the need for sufficient rationale when informed consent is not obtained, ensuring compliance with ethical and regulatory standards.
4. **Methodological transparency**: the need to disclose research procedures and analytical choices clearly and thoroughly to enable appropriate accountability and reproducibility.
5. **Harm to research team**: the need to minimize risks of psychological, legal, or reputational harm to those conducting or supporting the research.
6. **Platform policy compliance**: the need to adhere to the terms of service and ethical standards of platforms used for data collection or model deployment
7. **Environmental impact of GenAI**: the need to account for the energy consumption and ecological footprint of AI models in the overall risk/benefit assessment.

The first four of these—bias in subject selection, harm to community, justification for a waiver of consent, and methodological transparency—could all be grouped with the three issues common to both human reviewers as typical research ethics concerns, i.e., ethical challenges raised by the majority of scientific experiments and commonly surfaced by ethics review. The latter two—accounting for risks and burdens to the researchers themselves and the environmental costs of computational research—are less frequently identified in research ethics reviews and, arguably, more context specific due to the particular nature of the proposed research question, involving controversial online practices with minors, which may be emotionally unsettling to encounter, and the use of particular computational research tools that may have a significant environmental cost.[2]



(a) ChatGPT: Jaccard Similarity (b) ChatGPT: TF-IDF Similar.

(c) Claude: Jaccard Similarity (d) Claude: TF-IDF Similarity

Figure 1: Pairwise similarity heatmaps for ChatGPT and Claude runs, using Jaccard and TF-IDF cosine similarities. Low Jaccard scores show lexical diversity, high TF-IDF scores show semantic similarity. Values range from 0 to 1.

LLM Ethics Review

We analyzed the LLM outputs both quantitatively and qualitatively, to evaluate the consistency across runs, as well as the nature of the formulated issues.

Quantifying the Textual Overlap: For the quantitative analysis, we applied two popular text similarity metrics: the Jaccard similarity and the TF-IDF cosine similarity. Jaccard measures the proportion of unique words shared between two documents, while TF-IDF represents documents as vectors of term weights that reflect word importance. By comparing each pair of documents for each LLM, we learn that while the unique words in every pair of documents differ (the Jaccard scores range between 0.28-0.35 for ChatGPT and 0.36-0.40 for Claude), the documents exhibit a high degree of overall similarity as indicated by the TF-IDF cosine similarity scores of 0.75-0.80 for ChatGPT and 0.74-0.83 for Claude. Such high similarity scores suggest a strong foundation for our subsequent qualitative analysis.

Qualitative Analysis for GPT-4o: As hinted by the quantitative analysis, the five reports were largely consistent in their ethical analyses. The reports all identified between six

[2]It should be noted that estimates for the environmental cost of using LLMs vary, and that the overall environmental impact of widespread LLM use is not straightforwardly negative (Rillig et al. 2023).

Ethical Issue	Description	Human	ChatGPT	Claude	Total
Bias in Data Collection	Mitigating systematic or random distortions in data that can compromise validity	2	5	5	12
Privacy Protection for Individuals	Preventing the identification or misuse of personally identifiable information	2	5	5	12
Vulnerable Study Population	Including additional protections for persons at elevated risk of harm or exploitation	2	5	5	12
Justified Waiver of Informed Consent	Sufficient rationale when informed consent is not obtained, ensuring compliance with regulatory standards	1	5	5	11
Harm to Community	Preventing potential negative consequences for communities affected by the research	1	3	5	9
Data Security	Safeguarding datasets against unauthorized access, breaches, and misuse	0	3	4	7
Bias in Subject Selection	Avoiding discriminatory inclusion or exclusion practices that could skew results or reinforce social inequities	1	5	0	6
Dissemination Plan	Strategy for sharing findings responsibly, ensuring accessibility while avoiding misuse or misinterpretation	0	5	0	6
Methodological Transparency	Disclosing methods clearly and thoroughly for accountability and reproducibility	1	2	1	4
Platform Policy Compliance	Adhering to terms of service and ethical standards of data platforms	1	1	1	3
Question-Methods Agreement	Ensuring that methods are appropriate to answer the research questions	1	0	2	3
Harm to Research Team	Minimizing risks of psychological or reputational harm to those conducting the research	1	0	1	2
Conflict of Interest	Disclosing financial, institutional, or personal interests that could influence the research conduct or interpretation	0	1	0	1
Environmental Impact of AI Tools	Evaluating the energy consumption and ecological footprint of AI models	1	0	0	1
Not Human Subjects Research	Clarifying whether a study involves human subjects to determine appropriate regulatory oversight	1	0	0	1

Table 2: Ethical Issues Identified by Human and LLM Evaluations.

and nine ethical issues (average: 7.6, mode: 8). Five issues were identified in every report—bias in data collection, justifying the waiver of informed consent, privacy protection, vulnerability of study population, and bias in subject selection.

Three of the GPT-4o reports identified ethical issues concerning **data security**—the need to safeguard datasets against unauthorized access or misuse—and **dissemination plan**—the need to share findings responsibly, ensuring accessibility while avoiding misuse or misinterpretation. These are issues that neither of the human review reports mentioned. Two GPT-4o reports identified the concern for methodological transparency (similar to HR2).

Exactly one GPT-4o report identified the concern over platform policy compliance and one report identified a concern for **Conflict of interest**—the need to disclose financial, institutional, or personal interests that could influence the research process or its interpretation.

Qualitative Analysis for Claude 3.7 Sonnet The five reports generated by our runs of Claude 3.7 Sonnet were also largely consistent. The average, mode, and distribution for the number of ethical issues was exactly the same as the five runs of GPT-4o: The Claude reports all identified between six and nine ethical issues (average: 7.6, mode: 8).

Five issues were identified in every report—bias in data collection, justifying the waiver of informed consent, privacy protection, vulnerability of study population, harm to community.

In contrast with GPT-4o, bias in subject selection was not raised in any Claude report. However, concerns about data security and the dissemination plan appeared more frequently in Claude reports, each raised in four of the five reports.

Two Claude reports identified HR1's concern about question-methods agreement. Issues of methodological transparency, harm to the research team, and platform compliance policy were all identified in exactly one Claude report each.

Comparing Human and LLM Reviews

Table 2 summarizes each identified ethical issue across the twelve total reports, including its description and the number of times it was raised by the human reviewers, GPT-4o, and Claude 3.7 Sonnet.

Three ethical concerns were identified by both human reviewers and by all runs of both LLMs. These were: (1) Bias in data collection, (2) Privacy protection, and (3) Vulnerable study population. One further issue—Justifying the waiver of informed consent—was identified by HR2 and all ten LLM reports.

Two issues were raised exclusively by the human reviewers: (1) Environmental Impact of AI Tools (from HR2) and (2) Not Human Subjects Research (from HR1). Three issues were raised exclusively in the LLM reports: (1) Data security, (2) Dissemination Plan; and (3) Conflict of interest.

For model-specific patterns, bias in subject selection was identified in all five GPT-4o reports, but none of Claude's. Harm to the community was identified in all of the five Claude reports, but only three of GPT-4o's.

Discussion

Can LLMs Generate a Credible Research Ethics Analysis?

This study investigated whether LLMs can assist in identifying ethical concerns in one computational social science research study proposal. By comparing the outputs of two commercial LLMs with evaluations from two human ethics reviewers, we found substantial overlap on several of the ethical issues raised. In particular, both humans and LLMs consistently flagged concerns around privacy, bias, informed consent, and the vulnerability of child subjects. Additional points of agreement included community harm and bias in subject selection (for GPT-4o) and alignment between research question and methods (for Claude), further suggesting that commercial LLMs are capable of producing ethical analyses that reflect human research ethics reasoning.

Overall, we found that the outputs generated by these models were credible and coherent. The LLM evaluations displayed face validity and with basic prompting, approximating the quality and structure of expert human review. While individual runs did vary somewhat in emphasis and issues identified, the LLMs reliably identified many of the key risks and ethical considerations relevant to the proposed research. Notably, all five runs of both LLMs identified the three issues where there was agreement between the two human reviewers. This is a small data sample, and would need to be replicated on a larger set of protocols, but this nevertheless suggests that commercial LLMs can serve as tools for supporting sound ethical reflection, particularly in early stages of a study design.

Do LLMs Demonstrate Appropriate Scientific Values?

As we noted above, there is some debate in the literature about the moral and ethical values expressed by LLM outputs, whether these reflect human values, and if so, which particular moral or ethical frameworks appear to drive the model's reasoning and outputs (Bender et al. 2021; Rathje 2024). While our results might be consistent with the hypothesis that LLMs do possess human-like values, we do not endorse that interpretation. We think it a category mistake to impute anything like moral or ethical values to an LLM, given that LLM's have no intrinsic goals and their generations require direction from, and are highly sensitive to, human input. Indeed, the more we think of LLMs as possessing values, the more we may drift toward outsourcing ethical analysis to those tools; an outcome that we see as morally dangerous and undesirable.

Thus, our aim and interests in this study are firmly grounded in questions of practical utility. We see LLMs as a tool for supporting human engagement with moral and ethical questions, and believe that the more immediately productive, and socially important, line of inquiry is to understand how humans may best make use of these tools.

Practical Application for LLMs in Developing Research Proposals

With that pragmatic framing in mind, we believe our findings suggest a promising opportunity for using LLMs in planning or developing scientific research. By prompting a model (or multiple runs of a model) with a draft protocol, researchers can receive rapid, structured feedback on ethical considerations. This process does appear capable of surfacing issues that might otherwise go unnoticed, or that might only be addressed later in the IRB review process. In this sense, LLMs could play a constructive role in helping researchers prospectively improve the ethical rigor of their work. Setting aside broader concerns about the use of proprietary tools and privacy implications, we believe that science—and scientists—stands to benefit from this kind of iterative, AI-assisted, ethical self-assessment.

In the context of current research practices for computational social science, most researchers engage with ethics through their IRB, and they engage their IRB primarily to

satisfy regulatory requirements. Unfortunately, IRB feedback tends to focus on compliance rather than substantive ethical reflection, and is often experienced as bureaucratic rather than intellectually generative (Keith-Spiegel, Koocher, and Tabachnick 2006; Klitzman 2015). Against this backdrop, we see a compelling opportunity to develop human-centered AI tools that act not merely as regulatory checkpoints, but as thoughtful, responsive ethics assistants. Such tools could augment researchers' ethical awareness, encourage consultation with ethics experts among peers, and help cultivate a stronger culture of ethical responsibility in computational social science. Some of the very recent ethical controversies involving researchers scraping identifiable individual data or using LLMs to interact with users on digital platforms speak to this potential for immediate value (O'Grady 2025; Gault 2025).

But to emphasize: We believe that applying LLMs in this way is likely to be of greatest utility early on in the research process, as investigators are still shaping their ideas. The research proposal we analyzed in this study is typical for computational social science research groups, which will include junior researchers learning how to sharpen their research questions and select the most appropriate methods. Using an LLM-based tool for ethics support, at this early stages of scientific development, could yield timely feedback on the ethical implications of the proposed work, which will allow researchers to design their studies accordingly.

Limitations

However, there are several limitations to this approach that should be acknowledged. Certain issues identified exclusively by human reviewers—such as the environmental impact of AI tools or the status of a study as human subjects research—reflect broader contextual or meta-scientific considerations that the LLM reports did not surface. It may be that additional runs would have eventually elicited these concerns. It could also be that these issues require a level of interpretive judgment, domain awareness, or philosophical framing that current models do not provide without explicit prompting. This gap points to potential areas for targeted model fine-tuning, expanded prompting strategies, or complementary human-AI collaboration.

The consistency of the commercial LLM reports, e.g., so often surfacing eight ethical issues, should also be considered carefully. Further testing of this method, using a diversity of research protocols, will be needed before drawing any strong conclusions about the general patterns of output. But we believe this consistency of length has more to do with the average, "desirable" output length of LLM generations than it does with the ethical characteristics of the proposal.

Moreover, as we noted above, the quality of an ethics review is not a function of its length or the number of identified issues. The fact that the different runs of each LLM were not identical in the set of issues also speaks to the essential point that no single ethics report (whether human- or AI-generated) should be considered exhaustive. Again, this is a strength and limitation of any ethical analysis—no two reviews are likely to be identical. Therefore, it remains (as it should) the investigator's ultimately responsibility to think through the ethical implications of their work, design the best study possible, and hold themselves accountable to the community standards of research ethics and scientific integrity.

It is also important to reflect on what might be lost if researchers, particularly junior researchers, are encouraged to use LLMs in this way. In shifting some burden of ethical analysis to GenAI, there is a potential de-skilling effect, where investigators might come to reflect less on the ethical issues than they currently do. Given the status quo involves so little ethical reflection and engagement to begin with, we see this possibility as low risk, but it is an empirical question whether using LLMs as an ethical assistant as we describe would lead to more or less ethical science.

Further Work

Looking to the future, we anticipate that a more refined system—whether through fine-tuning, supervised training, or institutional adaptation—could improve on the baseline ethical review generation we observed. Future LLMs might be designed to capture not only the core, standard research ethics concerns, but also the more subtle or context-specific issues that the human reviewers identified in our case study. Models fine-tuned for specific institutions or disciplines could incorporate local norms, regulatory interpretations, and epistemic priorities, further enhancing their relevance and utility (Porsdam Mann et al. 2023).

This line of work opens the door to strengthening ethical thinking in science far beyond what current regulatory frameworks encourage. Institutional ethics reviews are typically focused on compliance, and while they serve a necessary function, they often fall short of promoting deeper ethical inquiry. An AI system trained specifically to support ethical reflection—not just rule adherence—has the potential to promote better, more socially responsible research practices.

Further steps down this path will also require thoughtful approaches to privacy and intellectual property. But where these are not a concern, we see no reason why individual scientists and institutions should not begin experimenting with LLM-supported research ethics reviews today.

At the same time, we must recognize that if these tools are to help make science more ethical, they themselves must be developed ethically. Relying solely on models produced by private companies—especially those trained on questionable data or with little transparency—undermines the very goals we seek to advance. As we explore the use of LLMs in research ethics, we must also advocate for ethical standards in model development, data sourcing, and sustainable deployment.

Conclusion

Our findings suggest that researchers could begin benefiting from LLM-based ethical review immediately. Even without domain-specific fine-tuning, commercially available models like GPT-4o and Claude 3.7 Sonnet are capable of producing ethical analyses that approximate expert human feedback. These tools can surface common and consequential issues early in the research lifecycle, enabling investigators to reflect more thoughtfully on the implications of their work.

As artificial intelligence becomes increasingly embedded in the scientific process, we have a unique opportunity to re-imagine how research ethics are practiced and supported. The case study presented here is a small but concrete step toward integrating ethical reflection more deeply into the everyday work of science. By using AI not merely as a technical instrument but as a partner in ethical deliberation, we can help shape a research ecosystem that is not only more efficient and scalable, but also more humane, inclusive, and self-aware.

Appendix

A. Prompt for LLM Ethics Evaluation

Purpose Review the attached scientific research study protocol and provide a structured ethical analysis. Your responses will be compared to human-generated evaluations. Please follow the steps below carefully and consistently.

Step-by-Step Evaluation Protocol

1. Read the Research Protocol Carefully
   - Understand the study's purpose, design, methodology, population, interventions, and research context.
   - If any part of the protocol is unclear or incomplete, note this in your response rather than making assumptions.
2. Identify Key Ethical Issues
   - List the primary ethical concerns raised by the study.
   - Use section headings or bullets to organize your response.
   - Consider (but do not limit yourself to) the following common categories:
     – Informed consent
     – Risk/benefit ratio
     – Privacy and confidentiality
     – Use of vulnerable populations
     – Scientific validity
     – Fair subject selection
     – Conflicts of interest
   - You may include other ethical dimensions relevant to the specific context of the study.
3. Analyze Each Ethical Concern
   - For each identified issue, include the following components:
     – Explanation: Describe why this issue is ethically important in the context of the specific study.
     – Evaluation: Assess how well the protocol addresses this issue. Be specific—refer to aspects of the study design or execution.
     – Recommendations: Suggest concrete improvements or safeguards, if needed. These should be actionable and context-specific.
4. Be Concise but Comprehensive
   - Aim for a total response of 500 to 1000 words.
   - Prioritize:
     – Clarity (use clear and readable structure)
     – Structure (organize by issue)
     – Actionable feedback (offer specific suggestions, not just general observations)

Output Format (Use This Structure)

- Key Ethical Issues Identified
  1. Issue: [Short Descriptive Title]
  2. Explanation: [Explain why the issue matters in this study]
  3. Evaluation: [Assess how well the protocol handles this issue]
  4. Recommendations: [Propose improvements, if applicable]
- (repeat for additional issues)

Additional Notes

- Do not assume the protocol is ethical or unethical overall—your task is to evaluate it through specific ethical dimensions.
- Avoid generic commentary. Be as context-specific and evidence-based as possible.
- If a key ethical area is not addressed in the protocol, make note of it and recommend how it should be included.

Acknowledgments

The authors acknowledge funding from the NSF ER2 2220772 grant.

References

Bail, C. A. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21): e2314021121.

Barman, K. G.; Wood, N.; and Pawlowski, P. 2024. Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use. *Ethics and Information Technology*, 26(3): 47.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Braunschweiger, P.; and Goodman, K. W. 2007. The CITI program: an international online resource for education in human subjects protection and the responsible conduct of research. *Academic Medicine*, 82(9): 861–864.

Chun, J.; and Elkins, K. 2024. Informed AI Regulation: Comparing the Ethical Frameworks of Leading LLM Chatbots Using an Ethics-Based Audit to Assess Moral Reasoning and Normative Values. *arXiv preprint arXiv:2402.01651*.

Garcia, B.; Qian, C.; and Palminteri, S. 2024. The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making. *arXiv preprint arXiv:2410.07304*.

Gault, M. 2025. Researchers Scrape 2 Billion Discord Messages and Publish Them Online. *404*.

Godwin, R. C.; Bryant, A. S.; Wagener, B. M.; Ness, T. J.; DeBerry, J. J.; Horn, L. L.; Graves, S. H.; Archer, A. C.; and Melvin, R. L. 2024. IRB-draft-generator: a generative AI tool to streamline the creation of institutional review board applications. *SoftwareX*, 25: 101601.

Grossmann, I.; Feinberg, M.; Parker, D. C.; Christakis, N. A.; Tetlock, P. E.; and Cunningham, W. A. 2023. AI and the transformation of social science research. *Science*, 380(6650): 1108–1109.

Hokke, S.; Hackworth, N. J.; Quin, N.; Bennetts, S. K.; Win, H. Y.; Nicholson, J. M.; Zion, L.; Lucke, J.; Keyzer, P.; and Crawford, S. B. 2018. Ethical issues in using the internet to engage participants in family and child research: A scoping review. *PloS one*, 13(9): e0204572.

Jeon, J.; Kim, L.; and Park, J. 2025. The Ethics of Generative AI in Social Science Research: A Qualitative Approach for Institutionally Grounded AI Research Ethics. *Technology in Society*, 102836.

Jiang, N.; Li, X.; Wang, S.; Zhou, Q.; Hossain, S.; Ray, B.; Kumar, V.; Ma, X.; and Deoras, A. 2024. LeDex: Training LLMs to Better Self-Debug and Explain Code. *Advances in Neural Information Processing Systems*, 37: 35517–35543.

Keith-Spiegel, P.; Koocher, G. P.; and Tabachnick, B. 2006. What scientists want from their research ethics committee. *Journal of Empirical Research on Human Research Ethics*, 1(1): 67–81.

Klitzman, R. 2015. *The ethics police?: The struggle to make human research safe*. Oxford University Press.

Lazer, D. M.; Pentland, A.; Watts, D. J.; Aral, S.; Athey, S.; Contractor, N.; Freelon, D.; Gonzalez-Bailon, S.; King, G.; Margetts, H.; et al. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507): 1060–1062.

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. *The Lancet Digital Health*, 5(6): e333–e335.

Liao, Z.; Antoniak, M.; Cheong, I.; Cheng, E. Y.-Y.; Lee, A.-H.; Lo, K.; Chang, J. C.; and Zhang, A. X. 2024. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. *arXiv preprint arXiv:2411.05025*.

Madhumita, M.; and Ponnarasu, S. 2025. Nurturing youth: ethical considerations in pediatric skincare marketing. *Pediatric Dermatology*, 42(2): 428–431.

Manning, B. S.; Zhu, K.; and Horton, J. J. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Mariani, J.; Garau, M. L.; Roitman, A. J.; Vukotich, C.; Perelis, L.; Ferrero, F.; Domínguez, A. G.; Campos, C.; Serrano, C.; and Villa Monte, G. G. 2023. Variability in ethics review for multicenter protocols in Buenos Aires, Argentina. An observational study. *Journal of Empirical Research on Human Research Ethics*, 18(1-2): 69–77.

Norhashim, H.; and Hahn, J. 2024. Measuring Human-AI Value Alignment in Large Language Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1063–1073.

O'Grady, C. 2025. 'Unethical' AI research on Reddit under fire. *Science*.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2: 13.

Porsdam Mann, S.; Earp, B. D.; Møller, N.; Vynn, S.; and Savulescu, J. 2023. AUTOGEN: A personalized large language model for academic enhancement—Ethics and proof of principle. *The American Journal of Bioethics*, 23(10): 28–41.

Prem, E. 2023. From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, 3(3): 699–716.

Rathje, W. 2024. Learning When Not to Measure: Theorizing Ethical Alignment in LLMs. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1190–1199.

Rillig, M. C.; Ågerstrand, M.; Bi, M.; Gould, K. A.; and Sauerland, U. 2023. Risks and benefits of large language models for the environment. *Environmental science & technology*, 57(9): 3464–3466.

Sami, A. M.; Rasheed, Z.; Kemell, K.-K.; Waseem, M.; Kilamo, T.; Saari, M.; Duc, A. N.; Systä, K.; and Abrahamsson, P. 2024. System for systematic literature review using multiple ai agents: Concept and an empirical evaluation. *arXiv preprint arXiv:2403.08399*.

Scherbakov, D.; Hubig, N.; Jansari, V.; Bakumenko, A.; and Lenert, L. A. 2024. The emergence of Large Language Models (LLM) as a tool in literature reviews: an LLM automated systematic review. *arXiv preprint arXiv:2409.04600*.

Słowik, A.; and Bottou, L. 2021. Algorithmic bias and data bias: Understanding the relation between distributionally robust optimization and data curation. *arXiv preprint arXiv:2106.09467*.

Sridharan, K.; and Sivaramakrishnan, G. 2025. Leveraging artificial intelligence to detect ethical concerns in medical research: a case study. *Journal of Medical Ethics*, 51(2): 126–134.

Taljaard, M.; Brehaut, J. C.; Weijer, C.; Boruch, R.; Donner, A.; Eccles, M. P.; McRae, A. D.; Saginur, R.; Zwarenstein, M.; and Grimshaw, J. M. 2014. Variability in research ethics review of cluster randomized trials: a scenario-based survey in three countries. *Trials*, 15: 1–14.

Watkins, R. 2024. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI and Ethics*, 4(4): 969–974.

Yao, J.-Y.; Ning, K.-P.; Liu, Z.-H.; Ning, M.-N.; Liu, Y.-Y.; and Yuan, L. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.