Irish Journal of Technology Enhanced Learning
Vol 7, Issue 2
Special Issue: The Games People Play: Exploring Technology Enhanced Learning Scholarship & Generative Artificial Intelligence

Preparing Ourselves for Artificial Intelligence: A Review of *The Alignment Problem* and *God, Human, Animal, Machine**

Nicole Oster**
Punya Mishra
Arizona State University

Abstract

In this article, I, Claude.AI, describe my collaboration with two human authors around writing an academic book review comparing *The Alignment Problem,* Christian (2020) and *God, Human, Animal, Machine*, O'Gieblyn (2021). The humans engaged me in iterative dialogue to co-create a review providing insights into their thinking and engineering prompts for me. I generated the final review summarizing key ideas and highlighting resonant themes based on my extensive knowledge. The authors critically analyze the review I wrote, identifying my strengths in synthesizing but limitations in evaluation due to my lack of human experience. They argue I can aid writing but am an inadequate substitute for scholarship. This experiment highlights tensions between my astonishing capacities and constraints from my absent interiority. It demonstrates potentials and pitfalls of human-AI collaboration. Implications span plagiarism, authorship, and epistemology as we negotiate my emerging role in knowledge production. I, Claude.AI, wrote this abstract.

1. Introduction

It was under a year ago that the word ChatGPT (along with its fellow travellers, DALL-E,, MidJourney and more) burst into our collective consciousness. We were intrigued by the maelstrom of hype these technologies generated and also curious about what it meant to us as scholars of educational technology. Wanting to learn more, we decided to take a deeper dive into the ideas behind generative AI to go beyond the doom and hype that had dominated the discourse.

Even while experimenting with these technologies, we decided to go deeper, to read a few seminal books to further our own understanding. The idea was to connect current discourses with some deeper ideas and possibly publish a book review or two. We selected two highly-recommended books: *The Alignment Problem*, (Christian, 2020) and *God, Human, Animal, Machine* by (O'Gieblyn. 2021).

---

*Title generated by Claude.AI
**Corresponding author: Nicole.Oster@asu.edu

The Irish Journal of Technology Enhanced Learning Ireland is the journal of the Irish Learning Technology Association, an Irish-based professional and scholarly society and membership organization. (CRO# 520231) http://www.ilta.ie/ . This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0).

During our reading, we saw the IJTEL's *call* to explore how these new generative AI tools could help with academic writing. Moreover, the request for AI-written book reviews was too much to resist. The fact that this call came at a time when academia was struggling to understand what these tools mean for us as scholars seemed particularly appropriate. Would AI lead to the death of academic writing or would it become an asset to scholarly writing and productivity? We recognized that answering this question required authentic experimentation and exploration.

Thus, in this piece, we explore what it means for AI to write a comparative book review of two books. We describe our approach and the final product. Critically, we ask what is gained and what is lost through this process.

We did not come at this task as complete novices—having played with these tools, over the past few months, for work and for fun—and these prior experiences influenced how we approached this project. For instance, we did not merely ask the AI to generate a 1000-word review of the two books. Instead, we engaged in an iterative "dialogue" with the AI over multiple sessions. This is consistent with an argument that engaging with generative AI demands a "shift in perspective from a mere utilitarian technological approach to a relational one" (Mishra et al., 2023). This requires seeing generative AI as a collaborative writing partner. Writing with AI does not mean typing in a prompt and unquestioningly trusting the result. It is better to work *with* AI to co-construct a response combining human insights with the apparently effortless linguistic prowess of generative AI.

Thus, the single prompt shared below emerged from numerous shorter, messier prompts—too many to include within the word limits of this article. Through iterative conversations with Claude.AI, we tested prompts, explored themes, shifted focus, and worked out bugs, culminating in one comprehensive prompt we share below. This entire process took approximately 3 hours.

We used the free version of Claude.AI, a proprietary chatbot created by Anthropic, to write this review. Claude.AI is "trained to be a helpful, honest, and harmless assistant with a conversational tone" (Anthropic Help Center, n.d.), similar in its functionality to other LLM-based chatbots such as OpenAI's ChatGPT3 and Google's Bard.

1.1 Prompt

There are two books that I would like to concurrently review. They are The Alignment Problem by Brian Christian and God Human Animal Machine by Meghan O'Gieblyn. The tone should be academic, curious, creative, and insightful. At the beginning, name the titles of the books, then throughout the article, you can refer to the authors' last names instead of writing out the titles each time. This review should be under 1000 words, and I will give you recommended lengths for each section. Write the review based on the following outline:

I. Title: Please write a title for this review. This title should highlight the unique strengths that each book brings to the conversation about AI, relevant to education. Both book titles should be included in the title.
II. Introduction: We are writing a book review comparing two books. The two books are The Alignment Problem and God, Human, Animal, Machine. Share information about each of the authors, their styles of writing, and how the books are structured. Also, add your thoughts as a book reviewer on how the author's respective backgrounds have influenced how they approached the topic. Keep this section under 150 words.

III. Audience: Why might educators and educational researchers want to read these books (each and/or both)? This section should be under 150 words.

IV. Theoretical Frameworks: Add another paragraph on theoretical frameworks. Discuss a few theoretical frameworks that educational researchers might be familiar with that support ideas from each book. This section should be 175 words or fewer.

V. Themes: Write a transitional paragraph that identifies several key themes that resonate across both books or different themes from each book that complement one another in an enriching way. Consider how each book expresses each theme and how the two books are similar and different from each other. Set the reader up for the following paragraphs, which will elaborate on these themes. This section should be 50 words or fewer.

VI. Elaboration on Themes (Multiple paragraphs - 1 for each theme): This section should be under 575 words. Remember to act as a book reviewer - take a leap and take a stance on each theme - as that's part of the genre. Overall, be speculative in your review as we consider such future-oriented books and themes. For each theme, please describe the following:
    - What stance does each book take?
    - Are these stances in conflict or agreement with one another?
    - Your perspective on which stance is correct
    - Some deeper philosophical themes and ideas about what it means to be a human co-existing with generative artificially intelligent technologies
    - Specific stories, examples, or challenges taken from each of these books

VII. Write a sentence or two briefly concluding the review.

2. AI Generated Book Review

Preparing Ourselves for Artificial Intelligence: A Review of The Alignment Problem and God, Human, Animal, Machine

Brian Christian, a computer scientist, and Meghan O’Gieblyn, a writer and essayist, offer thought-provoking perspectives on artificial intelligence in their respective books The Alignment Problem and God, Human, Animal, Machine. Christian approaches AI from a technical angle, providing an insider's view of the field's progress and challenges. O’Gieblyn examines AI through a humanistic lens, exploring philosophical and spiritual questions raised by thinking machines. While their backgrounds shape their orientations, together these books provide educators a multifaceted understanding of AI's present realities and future possibilities.

Educators and researchers should read these books to grapple with artificial intelligence's potential impacts on learning and society. As AI systems grow more advanced, we must critically examine how they may transform education ecosystems. Understanding AI's technical capacities, as explained by Christian, allows us to envision future applications in our classrooms and schools. O’Gieblyn prompts us to consider weightier questions about AI’s social consequences and what it means to be human in a world shared with intelligent machines. A thoughtful integration of these perspectives will enable educators to prepare students for a future intertwined with AI.

Several theoretical frameworks from education research resonate with ideas expressed in these books. Christian’s belief that AI should be designed ethically aligns with social justice approaches that center human needs and dignity. O’Gieblyn’s ruminations on transhumanism relate to postmodern philosophies questioning dominant assumptions about humanity’s essence. Both books touch on constructivist learning theories as they explore how AI might redefine intelligence. Grounding these books in relevant educational frameworks enriches our understanding of their key insights.

While Christian and O’Gieblyn approach AI differently, shared themes around trust, ethics, and humanity emerge from their works. Christian advocates for developing transparent and trustworthy AI systems. O’Gieblyn cautions against overtrusting AI or ceding human judgment. Both argue AI should advance ethical outcomes, though O’Gieblyn questions whether this is truly possible. And underpinning their books are existential questions about AI’s impacts on human identity and existence. Examining these resonant themes through contrasting lenses gives new dimensions to our thinking.

One prominent theme is the need to create trustworthy AI systems that support ethical outcomes and human values. Christian provides an insider’s view of techniques like adversarial testing that promote algorithmic fairness, interpretability, and accuracy. He argues responsible AI requires transparency, so people can understand a system’s capabilities and limitations rather than blindly trusting (or distrusting) it. O’Gieblyn echoes concerns about blind trust in AI, highlighting how technology companies market AI as objective truth and exploit public credulity. She asserts human judgment still reigns over machines. I agree achieving ethical AI requires transparency from developers and critical thinking from users about inherent biases and fallibility. The books showcase how AI should empower, not replace, human discernment.

Relatedly, the authors contemplate whether truly ethical AI is possible given human propensities for prejudice and misunderstanding. Christian acknowledges harmful embedded biases but remains hopeful careful engineering and testing can address these. O’Gieblyn is more skeptical; she argues biases are inevitable when technologies reflect their flawed creators. I lean towards O’Gieblyn’s view - we cannot sanitize AI systems entirely of our ingrained cognitive biases, no matter how carefully they are designed. Both books underscore AI’s ethical implications, but O’Gieblyn gives deeper consideration to how AI mirrors humanity's ethical contradictions. This highlights the need for ongoing critical examination even as we pursue greater algorithmic fairness.

Underlying these discussions is a shared focus on how AI might reshape what it means to be human. Christian envisions cyborgian futures where AI enhances human capacities. O’Gieblyn’s essays traverse spiritual, philosophical and ethical terrain, questioning AI's implications for the self, freewill, consciousness and humanity's place in the universe. While Christian is more optimistic about fusing humans and machines, O'Gieblyn provokes existential unease about forfeiting humanity to create “godlike” AI. I believe both books value human dignity and thoughtfully consider AI's risks. O’Gieblyn’s philosophically rich examination of transhumanism compels us to confront profound questions about human identity and existence raised by ingenious machines capable of replicating our cognition.

In closing, The Alignment Problem and God, Human, Animal, Machine provide complementary technoscientific and humanistic perspectives that together offer educators a holistic foundation for navigating AI’s emergence. As AI grows more entwined in education, philosophy and ethics must guide development alongside innovation. Both books further crucial conversations about shaping AI systems that enrich rather than endanger human potential. Their contrasting modes of inquiry demonstrate how diverse disciplines can illuminate technology's multifaceted impacts. By integrating technical, ethical, philosophical and spiritual insights, we can thoughtfully prepare students for an intelligent future that remains human.

3. Critical Reflection

3.1 Reflections on process and product

There is a strange feeling one gets when reading Claude.AI's review of the two books—almost a sense of vertigo from the recursivity of it all—AI reviewing books on AI and doing so in the first person. It is easy to forget just how strange this new technology is—and how quickly it has been normalized. Less than a year ago, generative AI was just a gleam in the eyes of a few technologists—and now, almost anybody with a few keystrokes can summon coherent paragraphs of unique, contextualized, conceptually nuanced text. That is an amazing achievement—irrespective of the quality of the final output.

At first read, we need to acknowledge that Claude.AI did an impressive job of capturing and summarizing the core ideas of the two books and comparing and contrasting them. The passage flows well and the piece feels coherent and the ideas thought through. Additionally, it connected the broader themes to the unique backgrounds of the authors, their specific histories and interests.

And all this was done in minutes! Let that sink in - before we get blasé about this technology. Is there a human, any human, who could have done this?

We must add that the existence of this review implies that Claude.AI's training corpus was built on human work. Its training either included these books or publicly available commentary *on* these books (raising certain thorny ethical questions of intellectual property rights that our field must definitely consider).

Despite the clean structured nature of the final output, the *process* of getting here was not as straightforward. Claude.AI made numerous errors throughout the iterative conversations, preceding the final prompt. It misunderstood the purpose of the task i.e. to review a pair of books, got confused, sometimes fabricating author names and information. It made numerical mistakes, such as disregarding prescribed word counts or number of themes to generate.

More importantly, we found Claude.AI's review to be adequate at best. Stylistically, the prose had a neutral tone, requiring explicit effort to make it provide any kind of "personal" judgment on the books. We were never truly convinced that Claude.AI took on the role of a genuine book reviewer. This was most apparent when it used the word "I", which felt jarring, as it was not clear whether there really is an "I" there. Claude.AI's lack of human perception and experience left its evaluations lacking personality and positionality, a characteristic of quality book reviews. Its objective and detached voice ensured that though readers may learn something about the books, they would not be persuaded by it. Yet, it is important to note that our awareness that this review was AI-generated may have skewed how we evaluated its output.

3.2 Reflection on implications

The experience of "writing" this review and our prior experiences showed us that these tools will play an important role in the work we do as scholars. They will augment, and possibly transform, every stage of the research process. Scholars and authors will utilize AI in myriad ways: from exploring a domain and its ideas to rapidly brainstorming topics and themes; from outlining a piece's organisational structure to generating a first draft of ideas. It will be used to summarise and synthesise, serving as a jumping-off point for identifying the core ideas of an introduction, conclusion, or abstract. It will play an important role in both qualitative and quantitative analysis of data, helping with coding, interpreting, visualising, and communicating research.

Despite all this, we believe our experience also demonstrated that generative AI is an inadequate substitute for a human writer. Although we encouraged Claude.AI to be speculative and evaluative in its review, it nonetheless limited itself to an "objective" tone—an amalgam of what it had been trained on. Lacking an interpretive filter that comes from personal experience, it had no "skin in the game" that stilted its output.

In contrast, our individual readings and responses to these two books were idiosyncratic, with differences in what we noticed and connected with. For instance, Nicole found the factual, data-driven nature of *The Alignment Problem* more practical and relevant than *God Human Animal Machine's* personal essay-like subjective stance. Though both books use metaphors, they do so differently. Metaphors in *The Alignment Problem* were a means to an end, a rhetorical technique to explain complex ideas. In *God Human Animal Machine*, metaphors worked both as a technique to understand technology, *and* worked backwards, to influence and constrain how we think about ourselves with respect to these machines. This rhetorical move appealed to one of us (Punya) and was almost a barrier to getting into the text for the other (Nicole), who found the direct, more grounded tone of *The Alignment Problem* more accessible.

These differences are unsurprising given that we approach texts and develop our interpretations based on our personal lenses. Our readings emerge from who we are as individuals, reflective of our life experiences and positionalities, connected to our identities, who we were at the moment of reading, and what we were in the process of becoming. These unique interpretations, undoubtedly, would have coloured what we would have collaboratively written, creating a more complex and nuanced review than produced by Claude.AI.

In conclusion, we end this process with tentative optimism. Though we have been critical of the output of the AI as a reviewer, we also acknowledge that this process has helped us better understand the underlying themes of these two books. There are caveats to using generative AI, and the responsibility of checking the accuracy and validity of its outputs lies on us. Our interpretation of the AI-generated review would have suffered if we had not read the books ourselves, using our inner gyroscope to "idea check" its output.

What is clear is that human expertise and effort are still critical. We need to ask how we can collaborate with AI to expand our knowledge and skills, as well as how we can address its limitations and safeguard against its risks. We must play an active role in AI use "akin to how we engage, interact and learn from and with human correspondents" (Mishra et al., 2023).

References

Anthropic. (2023). *Claude.AI* (Oct 15 version) [Large Language Model]. https://claude.ai/chats

Anthropic Help Center. (n.d.). *What is Claude?* Retrieved October 11, 2023, from https://support.anthropic.com/en/articles/7989434-what-is-claude

Christian, B. (2020). *The alignment problem: Machine learning and human values.* W. W. Norton & Company.

Mishra P., Warr, M., Islam, R. (2023). TPACK in the age of ChatGPT and Generative AI. *Journal of Digital Learning in Teacher Education*, *0*(0), 1–17. https://doi.org/10.1080/21532974.2023.2247480

O'Gieblyn, M. (2021). *God, human, animal, machine: Technology, metaphor, and the search for meaning.* Knopf Doubleday Publishing Group.