A reprint from

American Scientist

the magazine of Sigma Xi, The Scientific Research Honor Society

AI and Responsible Authorship

Why my chatbot is not (yet) a coauthor.

Robert T. Pennock

Suppose I do the experiments but use an artificial intelligence chatbot to write the report; should I list it as an author? If I only use the chatbot to flag typos or suggest fixes for grammatical errors, that question would never arise. But what if, to save time, I have the AI write the literature review section summarizing a set of articles I gave it? Now the words on the page are not my own. More significantly, what if I give it my experimental data to analyze and write up? As AI increases in power and capabilities, does it deserve credit as a coauthor?

From Lovelace to LLMs

In 1843, Ada Lovelace published what was arguably the first computer program, showing how an analytical engine—as mathematician Charles Babbage called his yet-unbuilt digital mechanism—could calculate a common sequence of rational numbers called Bernoulli numbers. A computer program is just a step-by-step procedure, but Lovelace's algorithm could do something that at the time only a person could. An algorithm may run on a mechanical device with gears, on an electrical device with circuits, or on an abstract writing instrument and roll of paper that moves based on symbols written on it—a Turing machine, named after computer pioneer Alan Turing. The idea of artificial intelligence is that such artifacts can in principle be able to exhibit recognizable, if perhaps not exactly human, intelligent activity.

As a PhD student in the late 1980s, I worked with Herbert Simon, the Nobel Prize–winning polymath known as the father of AI for his pioneering theoretical and empirical work that founded the field. Simon argued that AI should be analyzed in terms of symbolic reasoning. I also heard computer scientist and cognitive psychologist Geoffrey Hinton, now called the godfather of AI, argue for and demonstrate early results of an alternative "connectionist" approach that focused instead on statistical associations in *artificial neural networks* (ANNs). ANNs were modeled on brain structures, with varying weights of connections between nodes governing the processing from input to output.

The write-up serves a vital function because it reports the evidence, but authoring is not the core part of research.

The relative merits of these approaches made for vibrant debate and drove interesting research. For example, symbolic AI researchers analyzed the rules of expert reasoning and devised programs to simulate them. Hinton, who later received the A. M. Turing Award, and other connectionists devised better ways to train ANNs. For years, practical applications in both symbolic and connectionist AI always seemed beyond the horizon, but the early 2020s saw rapid advances in their capabilities.

By training an attention-based transformer model on massive amounts of text gathered from the internet, the weights of an ANN can be adjusted so that it becomes an adroit text manipulator. Such *large language models* (LLMs) take a text prompt as an input and then generate a string of output text based on predictions of what words should follow what came before. LLMs generally lack symbolic structure and are not yet very good at math, but various hybrid systems attempt to combine the strengths of both symbolic and connectionist models. Today LLMs can generate a convincing essay about Bernoulli's mathematical discovery. Lovelace would be impressed.

The term *computer* originally referred to a person who performed mathematical calculations—computations. When computing machines took over these clerical tasks, they also took over the term. Is AI poised to similarly replace authors? In particular, has AI come far enough to be responsibly included as an author of a scientific paper?

Authorship Problematized

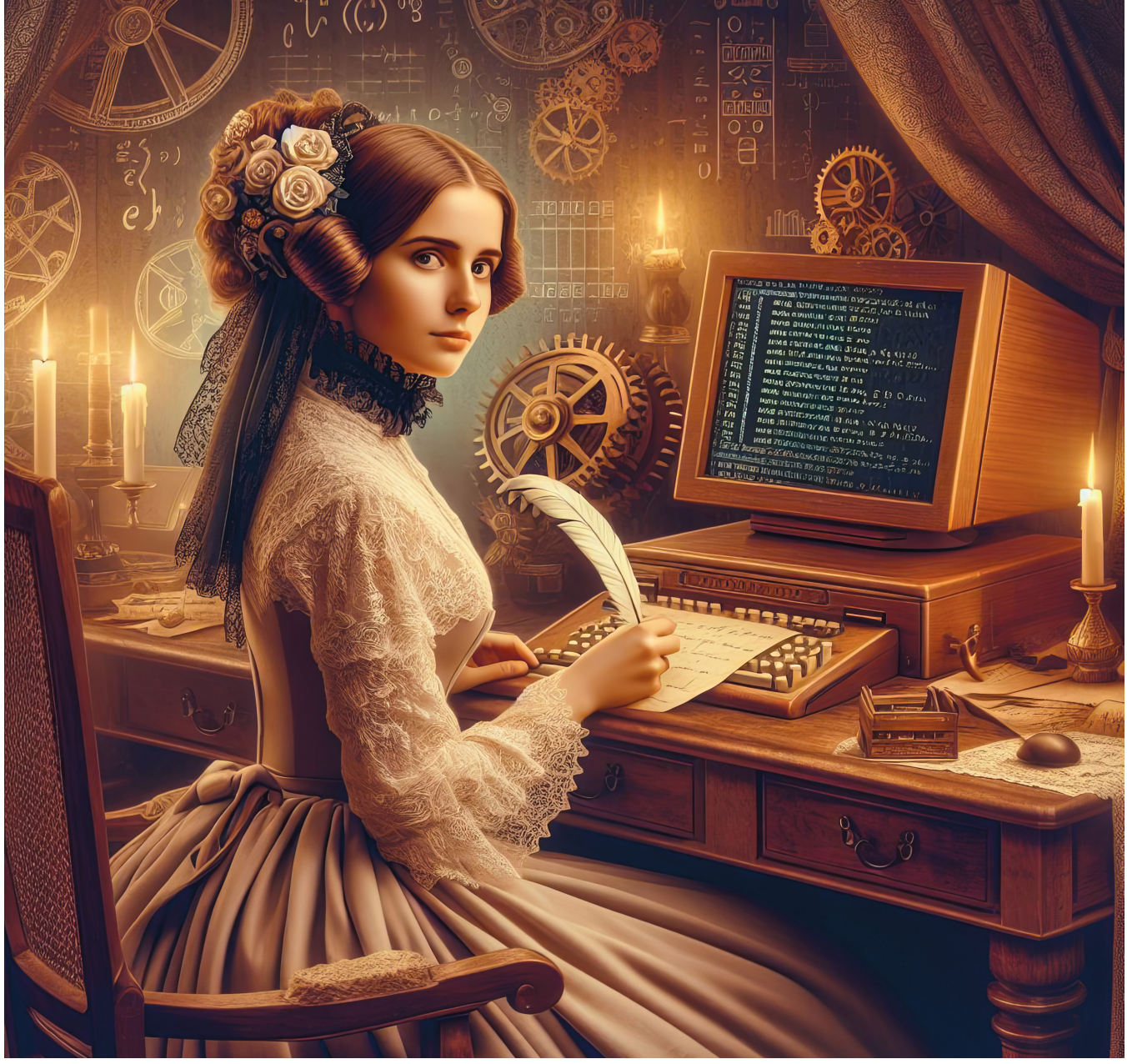
To analyze the ethics of responsible scientific authorship, one must first consider the notion of authorship itself. On first pass, *authorship* seems problematic if only because the word is linked etymologically to the idea of authority and thereby to an unscientific, legislative notion of justification.

QUICK TAKE

The use of artificial intelligence in the development of research papers raises the ethical question of whether AI tools should receive coauthor credit.

Responsible scientific authorship is less about authoring the text of a research paper than it is authorizing the paper as a fair representation of the evidence supporting its findings.

AI agents might be able to dig up unknown references or even analyze original data, but they are not yet able to take responsibility for the research in an ethical sense.



Magic Studio

AI productions may or may not reflect reality, so users must be wary and check outputs themselves. An AI image generator was given the prompt "Ada Lovelace coding at a desktop computer" and created this fanciful illustration of that historical figure, who was involved in early computer algorithms. AI tools still don't get basic things quite right, such as Lovelace here writing with a quill instead of using the keyboard. AI tools are powerful, but it is only moral agents who can bear responsibility for the use of what they generate.

On the legislative model, the say-so of the proper authority creates and justifies a law because a recognized authority is its author. But science brooks no such justification; a scientific conclusion is justified not by authority, but by observational evidence. Philosopher Blaise Pascal, writing in the 17th century during the scientific revolution, put the point bluntly: "On subjects in [the physical domain] we do not in the least rely on authorities—when we cite authors, we cite their demonstrations, not their names."

Indeed, presentations at the Royal Society from the early days of the scientific revolution commonly included a physical demonstration of instruments, materials, and procedures. As far as possible, researchers exhibited phenomena, rather than just penning descriptions thereof. The write-up serves a vital function because it reports the evidence, but authoring it is not the core part of research.

I'm a Scientist, Jim, Not a Novelist

In 1983, the first version of Microsoft Word came out. I taught students how to use this new word processing tool, telling them it was their future. My mother had earned money in college typing classmates' papers on her Smith Corona typewriter; today, kids use Word in elementary school. LLM chatbots are the next big step in word processing but threaten a more uncertain future. Generative AI makes it easy to produce a research paper with little more than a clever prompt. The result may not be plagiarism in the usual sense of the word, but it isn't original work either.

I'm always dismayed when a student refers to a nonfiction science book we are reading as a novel. There is a stark difference between science and science fiction. This division ought to be obvious, but some postmodernist critics of science muddied the distinction, arguing that scientific practice may be described simply as the manipulation of texts. They portrayed science as a subset of literary activity—scientists "construct" the world (rather than discover it) by "writing" a "narrative." Such "stories" may be accepted

Wikimedia Commons

Alan Turing (*top left*), Herbert Simon (*top right*), and Geoffrey Hinton (*bottom*) are pioneering figures in the development of artificial intelligence. One of Turing's conceptual advances was the idea of using an imitation game—determining how well a person could spot a machine impersonating a human—as a test of computer intelligence. Simon argued that AIs should be evaluated in terms of symbolic reasoning. Hinton took an alternative approach that considered statistical connections in artificial neural networks, modeled on brain structures, with varying weights of connections between nodes governing the processing from input to output. Large language models are an application of this idea.

as true, but only because society has granted scientists, perhaps unwisely, this creative privilege. This view was seriously mistaken. Despite some interesting areas of overlap, the norms that govern production of reports of scientific discovery are crucially different from those that govern construction of works of creative fiction.

The essential core of scientists' work involves applying the scientific method of inquiry to an empirical research problem. Scientists must formulate a reasonable model that might explain a phenomenon of interest, design a justified experimental protocol to test it, carefully execute the procedure, gather data, and then analyze the results. When researchers follow the norms of the scientific method, they are not primarily authors or creators, but rather reporters and discoverers.

Glorified Autocomplete?

One might think that it is enough to disqualify AI from being an author of a scientific paper by arguing that it is a fancy word processor that discovers nothing new. The text prediction facility of LLMs, impressive as it is, has been criticized as no more than a glorified autocomplete system. But this judgement may be too quick. Hinton says that he already detects an emergent intelligence capable of real reasoning and understanding.

Lovelace's program dealt with math, but she presciently speculated that the analytical engine might also be able to operate on other things, such as musical notes. Today, AIs can generate listenable music from user text prompts. Creativity is different from discovery, as we noted, so LLMs can't just create an empirical finding. But if we connect one to sensors, couldn't it analyze the data and possibly make a new discovery about the world?

Simon thought that AI would be able to do just that. He and his colleagues investigated this possibility using a program they built called BACON, with heuristics thought to facilitate scientific discovery. They tested it on known cases with some remarkable early success, for instance, giving it planetary motion data and observing it rediscover Kepler's third law. Today, LLMs are being trained on protein amino acid sequence data to generate candidate sequences to help discover useful novel enzymes.

Of course, this technical ability is not enough to make AI a full-fledged discovery machine. One must be able to display the evidential relationships that connect such a machine's outputs to existing scientific knowledge and recognize their import in extending that knowledge. A computer might record and process data from a radio telescope, but for the time being we still need a Jocelyn Bell, who discovered the first pulsar from such radio signals, to distinguish a meaningful signal from a glitch.

These issues relate to several ethical principles that are important for understanding responsible authorship attribution in science.

Truth and Tools

The first ethical principle of science involves truth. The goal of science is to discover empirical truths about the natural world, which is why honesty is a core virtue in science. It doesn't even make sense to seek a discovery without it.

AI chatbots currently are not good at distinguishing truth from falsity.

They are often trained on indiscriminate internet texts, which provides rich diversity but questionable accuracy. Just as biased internet data can cause AI bots to exhibit racial biases, they can also lead to skewed reporting about factual matters.

Falsities can also be generated because of the nature of the systems; LLMs don't produce outputs by direct comparison with reality, but rather based on probabilistic patterns of texts. Thus, LLMs can produce what are referred to as *hallucinations* or *confabulations*, glitches where the bots output what their models predict to follow what came before, but which are not true. They put forward concocted statements with apparent confidence. LLMs may even support these false statements with references that they also made up. Unwary, trusting users are easily burned.

All instruments must be calibrated, and AI is no different from other technical tools. Even mature software may have bugs. My research team recently encountered an obscure incompatibility between Mac and PC versions of Microsoft Excel that caused some cell data not to show up in searches, corrupting results until we figured out the cause. A new technology such as AI chatbots will have many more unexpected flaws. Tools are not always reliable, so it is incumbent upon users to perform the requisite checks.

Bearing Responsibility

Of course, humans can be unreliable and make mistakes as well. The difference is that although a malfunctioning tool can be the cause of mistakes, humans bear responsibility for them. This ethical concept is central in the original question about responsible authorship—what it means to be responsible.

Colloquially, we often conflate the two different notions. For example, when investigating a traffic accident, we may ask what was responsible for the crash and conclude that it was brake failure. That is, we are asking for the cause. For a failed experiment, we may similarly identify a cause, such as a bug in a computer program. It is a different question to ask who was responsible for a failure—the mechanic, say, or the developer. This notion involves blame, which is an evaluative concept. In cases of success, responsibility brings credit. Either way, this second notion of responsibility takes us beyond the realm of mere causation and into the realm of ethics.

One key aspect of ethical responsibility is contained in the term itself—responsibility involves being ready and able to respond. Responsible conduct of research implies a duty to ensure that things are done properly and to stand up to answer a call to accountability if something goes wrong.

Scientific reasoning involves identifying the causal relationships that are evidentially relevant in the test of a model. It requires objectively assessing the data and skeptically watching out for possible biases, including one's own. It requires humbly submitting to what that evidence shows, even if the results go against one's favored model. And so on. For a scientific paper, responsibility implies a duty to honestly exhibit such evidence. To deserve credit as an author, one must be in alignment with and be able to take responsibility for such values.

Large language models can produce what are referred to as *hallucinations* or *confabulations*, glitches where the bots output what their models predict to follow what came before, but which are not true.

The Credits

Historically, it was important to have a written statement of research not only to disseminate information, but also to establish priority. Scientists received, and for the most part continue to receive, relatively little financial remuneration for their labors. Their primary reward was the joy of discovery itself and the peer recognition that their work earned them. One of the roles of the Royal Society was to oversee deposits of sealed reports of discoveries, so that scientists could continue experiments to fully establish a result and pursue its elaboration without fear that other researchers might scoop them before they were ready to make their work public.

This issue brings up a third ethical principle for responsible authorship: justice. It is not equitable to give credit that is not earned. The issue of equity arises when authorship assignment is too coarse-grained. Such inequities can be avoided by moving to an attribution model that explicitly lists the specific responsibilities of the researchers.

Instead of calling all participants "authors," papers should list their roles and contributions. Depending on the type of research, these roles may range from general ones such as "Principal Investigator" to specific ones such as "Statistician" or "Virus samples contributed by . . ." Such a model of attribution—what I call "the credits" by analogy with film credits—better follows ethical principles of justice by recognizing and equitably distinguishing actual research roles. It supports responsibility by allowing one to receive just recognition in accordance with one's true contributions. This model also helps in assignment of blame. Contributors are not equally at fault in cases of misconduct; explicit attribution identifies the parts of the research for which they actually bear responsibility.

I proposed this credit attribution model in 1996, and various journals have since adopted some form of it, such as the Contributor Roles Taxonomy (CRediT), which classifies some common research roles. AI bots are not contributors, but just as film credits may list the use of Panavision equipment, researchers can cite AI agents in the same way that they would a specialized instrument. If AI bots are used substantively in the research or report, this use is evidentially relevant, and responsible authors should list them as tools.

Ghostwriter in the Machine

Computer programs have come a long way since Lovelace's first algorithm. Today's text generators may help or harm writing, but they are not authors in the moral sense. Whether they could ever become responsible sentient agents is not just a scientific and engineering question, but a philosophical one.

Both Plato and Aristotle considered the implications of *automata*—machines capable of autonomous behavior. The 18th-century Swiss

Wikimedia Commons

Pierre Jaquet-Droz's functional automaton, "The Writer," could inscribe lines of text on note cards by means of a complex clockwork mechanism. Large language model AI chatbots produce text using pretrained neural networks instead of arranged physical gears and cams, but the process is no less mechanical.

watchmaker Pierre Jaquet-Droz built several clockwork automata to implement this idea. The most sophisticated was "The Writer," a mechanical boy he programmed to produce lines of text on note cards. Generative AI today can do far more, but it is still essentially "mechanical." Could a mere mechanism ever be truly human or, as in the Japanese manga and anime movie *Ghost in the Shell*, is a human consciousness required? That title references the phrase "ghost in the machine," which comes from British philosopher Gilbert Ryle's critique of the idea of metaphysical separation of mind and body. Ryle argued that humans are also "just" mechanisms, but that state doesn't prevent our being intelligent moral agents.

What authorship means in the scientific context is not about authoring the text so much as authorizing the report as a fair representation of the evidence.

AI bots can now pass Turing's imitation game—in which an AI can fool a person into believing that it is a human—in some circumstances. One programmer recently trained a LLM to automatically converse with online potential dating partners, to filter down to compatible individuals before communicating personally. We have not yet achieved an AI Cyrano de Bergerac, but still. It takes more than simple intelligence to have a moral sense and bear responsibility, but we can't rule out the possibility of a future AI ghostwriter.

Endorsing the Check

Science is an evidence-based discipline and eschews appeals to authority. A scientific paper is not an act of creation but a report of evidence for a discovery. By putting one's name on that report, a scientist is taking responsibility for its contents. What authorship means in the scientific context is not about authoring the text so much as authorizing the report as a fair representation of the evidence: I authorize it in the sense that I endorse it.

Think of this act as the scientific version of endorsing a check. Signing off on a report indicates that I have performed the requisite tests—checked and double-checked the protocols, data, calculations, and results—and will stand by them. I put my name to it to affirm that I take responsibility for the experiments and analysis reported therein. If the evidence I report for the discovery is good, you can take it to the bank. Researchers who have been caught fabricating data, on the other hand, are like forgers whose checks have bounced. Having violated essential scientific values, they are no longer trustworthy. Their endorsement is worthless.

AI is still just a tool. If it fails, we wouldn't blame it morally but causally, and we would take steps to fix it. Like other powerful tools, AI is useful but can be dangerous. It must be used responsibly. It cannot, at least not yet, be itself responsible. Perhaps the future will bring an AI that can, and thereby will deserve credit as a coauthor. But for now, the responsibility lies entirely with me.

Bibliography

Hollings, C., U. Martin, and A. Rice. 2018. *Ada Lovelace: The Making of a Computer Scientist*. Oxford, UK: Bodleian Library Publishing.

Langley, P., H. A. Simon, and G. L. Bradshaw. 1987. Heuristics for empirical discovery. In *Computational Models of Learning*, ed. L. Bolc, pp 21–54. Berlin: Springer-Verlag.

Pennock, R. T. 1996. Inappropriate authorship in collaborative scientific research. *Public Affairs Quarterly* 10:379–393.

Rothman, J. 2023. Why the godfather of AI fears what he's built. *New Yorker* (November 13).

Robert T. Pennock is Sigma Xi Senior Fellow for Science and Engineering Values. He is a University Distinguished Professor at Michigan State University, where he is on the faculty of Lyman Briggs College, the departments of philosophy and computer science and engineering, and the ecology, evolution, and behavior program. His research involves both empirical and philosophical questions that relate to evolutionary biology, cognitive science, and the scientific character virtues. Email: pennock5@msu.edu