A jazz band that improvises a song based on copyrighted melodies is producing something new, but it is done with awareness of the original author. Copyright law should protect this awareness from users of AI tools who lost contact with their source material or training data. When a product is made by a community that has lost awareness of the sources of its materials, it is no longer a human creation and should be called an AI product that will be deprived of its copyright until the authors have acknowledged their sources.

I am a US citizen and ex-physicist who is not affiliated with any software, entertainment, or publishing company and I wrote a novel in 2016-2017 called *My Adorable Apotheosis* that was used within the training data of some popular AI writing tools without any form of notification or permission. I took data on hundreds of novels and movies to quantify the scope of the problem.

One example of the data I took is presented below in Figure 1, entitled, *Axiom's End vs. My Adorable Apotheosis.* The page numbers on which the same thing is happening in both novels are charted out and a diagonal line appears whenever the events occur in the same order in both books. The chart below shows that *Axiom's End* used the same plot sequence and pacing as *My Adorable Apotheosis*, but tacked on an extra hundred pages to draw out the ending.

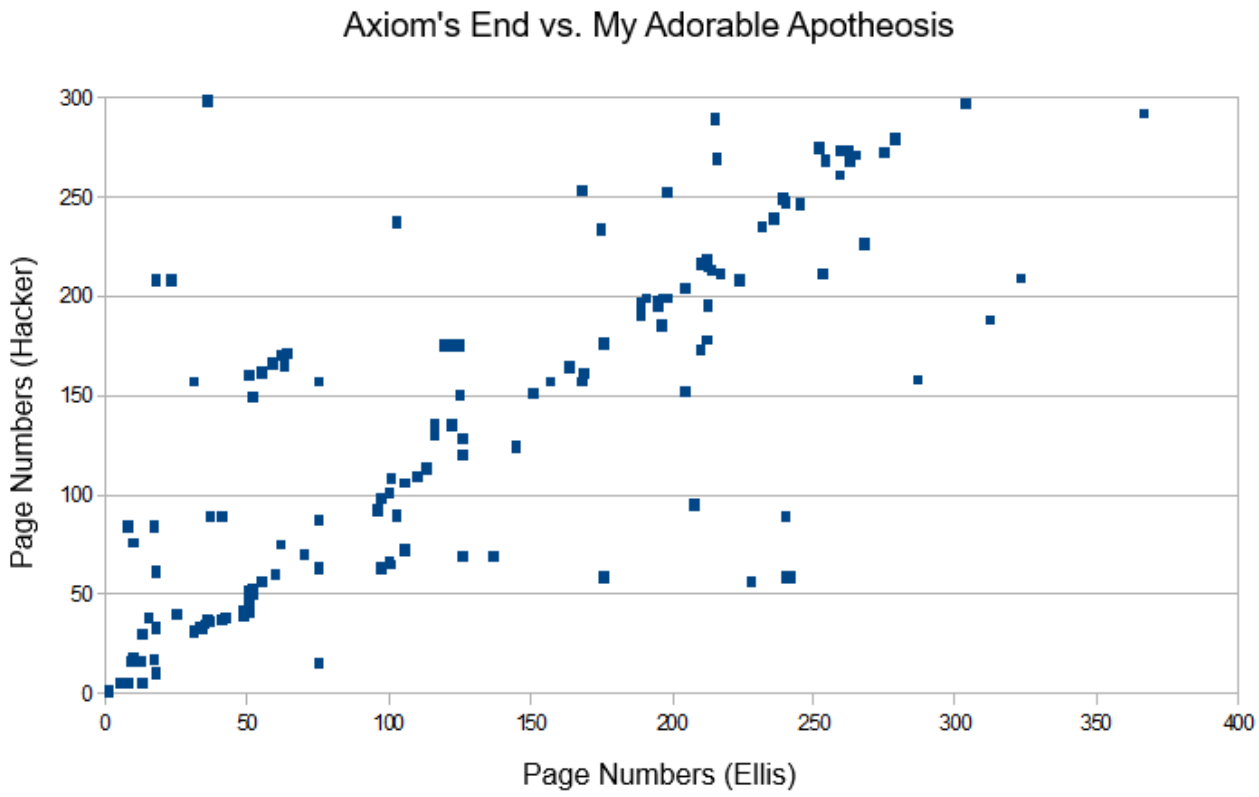


**Figure 1:** *Axiom's End vs. My Adorable Apotheosis.* Each point represents a page number on which a plot event occurs in both books.

I took this type of data on a hundred novels and found fifty produced after my book which showed a quantitative signature indicative of a machine-like form of copying. I concluded that the use of an AI writing tool within the novel-writing community has become widespread and that rejected, outsider novels like mine are used in the training data of a popular AI writing tool which allows the users to systematically extract inspiration from an unacknowledged source in the form of ordered writing prompts delivered by the tool.

The chart in Figure 2 entitled *Similarity of 100 Novels with My Adorable Apotheosis* shows that fifty novels written after *My Adorable Apotheosis* contain sequences of events which identify them as being generated with a systematic copying method applied to unacknowledged training data taken from my novel. To construct the chart, I recorded page numbers on which similar plot events occurred and identified consecutive sequences of events that exceeded the capacity of human working memory (14 events in sequence). I used this to calculate a 'similarity score' with methods that are easy to automate with present technology. I could find no books written prior to mine that generated a similarity score which was comparable to that of books that I identified as being written with the assistance of AI tools that had drawn sets of writing prompts from my novel via the training data. The data reveals when authors did not disclose use of automated methods.

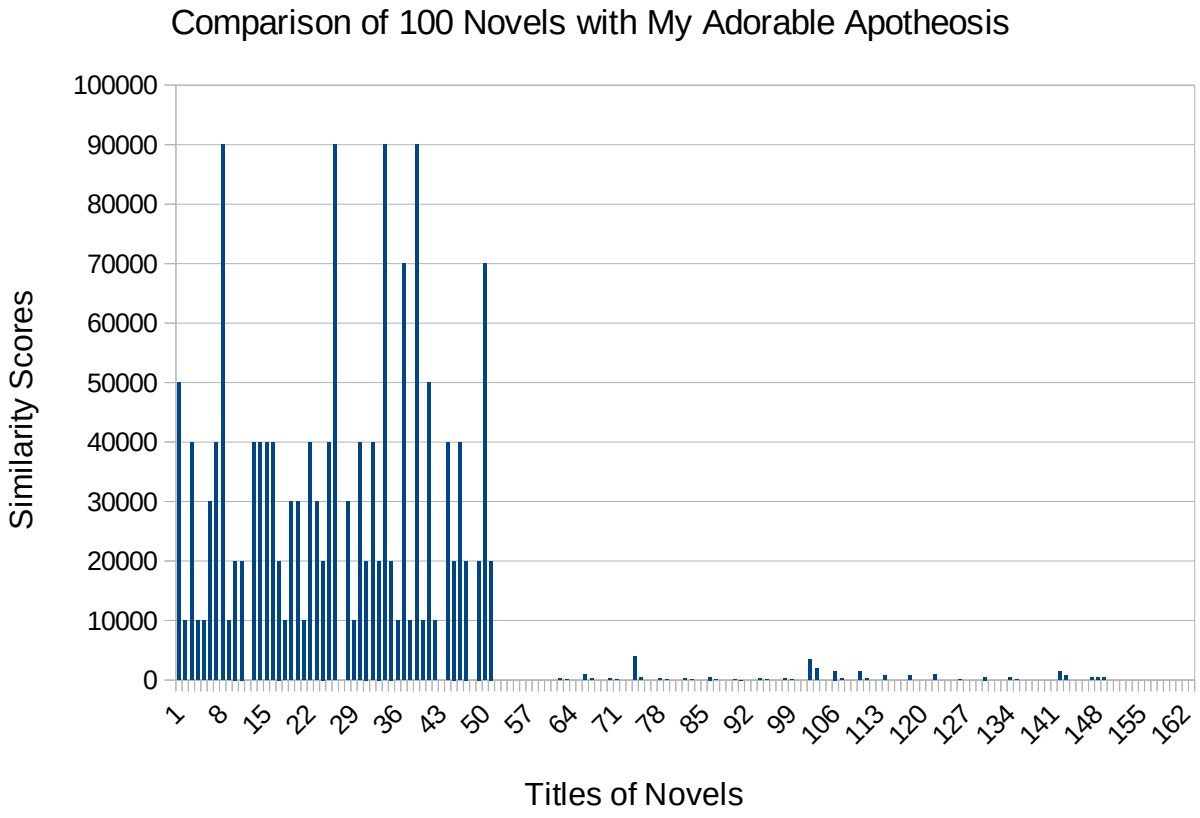


**Figure 2:** *Comparison of 100 Novels with My Adorable Apotheosis.* The first fifty novels were constructed with unacknowledged training data from *My Adorable Apotheosis.*

To prove that the evidence I gathered from novels was not spurious, I collected additional data sets on screenplays to quantitatively determine the signature of natural cross-influence and the signature of AI-assisted writing tools. In Figure 3 below entitled *'My Adorable Apotheosis vs. A Set of Disney/Comcast/Sony Films'*, natural cross-influence is represented by the points from 1990-2016 and a viral burst of copying from my unattributed novel is represented by the points from 2017-2023.

Each point in Figure 3 is associated with a sequence of writing prompts used to diagnose the percentage of plot overlap between a film and my reference text, *My Adorable Apotheosis*. The word counts required to describe the overlapping and non-overlapping plot elements are used to calculate the percentage. The level of detail I was able to record was limited by the speed with which my mind and fingers could move while the film was playing. Much like a court-reporter or stenographer, I wrote down what I saw on the screen and sorted it into two categories: events which occurred within my novel and events that did not occur within my novel. For a feature-length movie, the overlapping plot events typically involved a page or more of text when AI-assisted copying was evident. This is significantly more than the copying of a story concept which is a sequence of events that a human mind can easily remember.

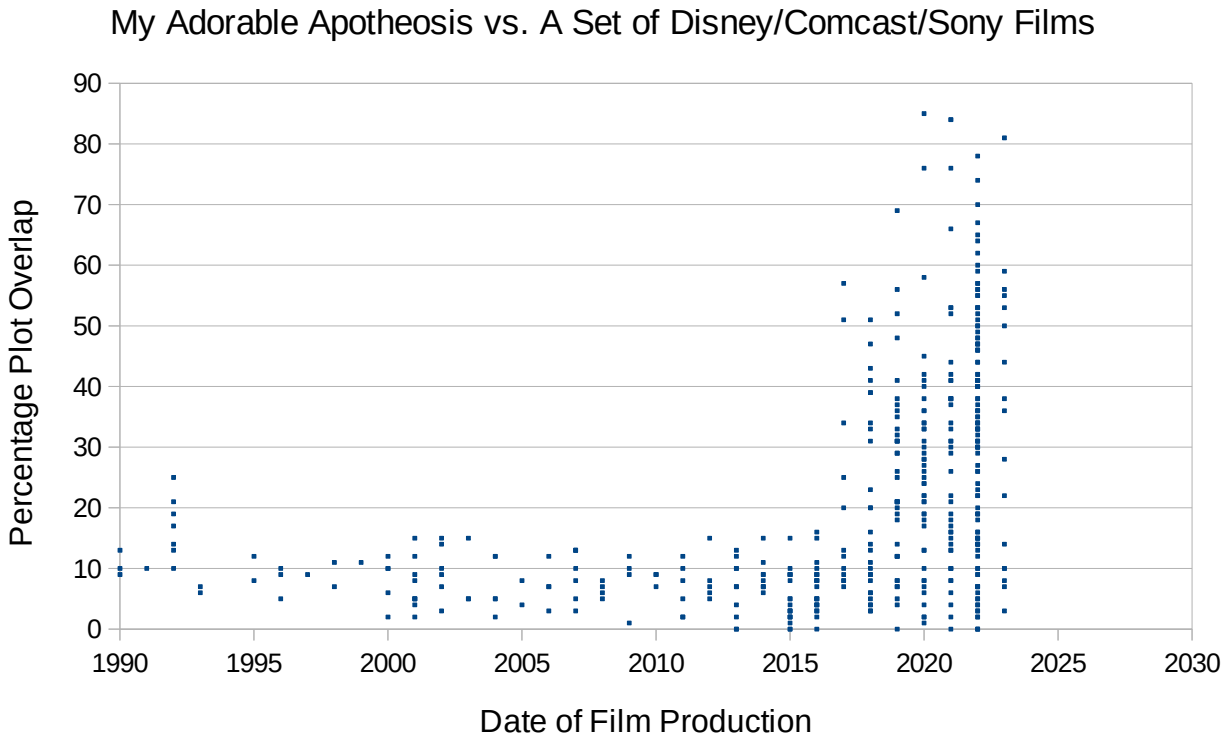


**Figure 3:** *My Adorable Apotheosis vs. A Set of Disney/Comcast/Sony Films.* After *My Adorable Apotheosis* was written and before it was published, the signature of its contents appeared within the screenwriting community, suggesting that novels submitted to agents and publishers were habitually dumped into screenwriting tools to be mined for material without notification or permission from the authors.

The set of five outlier points on the left side of the chart in Figure 3 represent some late 1990s *Star Trek: The Next Generation* episodes which were considerably shorter and lower in plot density than the other points in the data set. Whereas Star Trek tended to focus on a single plot-line which conveyed a complex concept in a distinctive way, it has become customary in recent decades for shows and movies to interleave multiple, simpler storylines to give the impression of complexity and depth, often drawing out the story over the course of multiple seasons. This has fostered a culture of rampant re-use of material across franchises as stories are translated between genres without controlling for the quantity of material that has been mined from a given source.

Just as AI tools can mine story materials from a new novel, they can also control and diagnose the quantity that has been mined. In the figure entitled *% Plot Overlap of Westworld vs. My Adorable Apotheosis*, four seasons of a popular HBO spin-off of the Micheal Crichton novel *Westworld* show how as the series progressed, its plot became more and more similar to that of my unacknowledged novel, *My Adorable Apotheosis.*

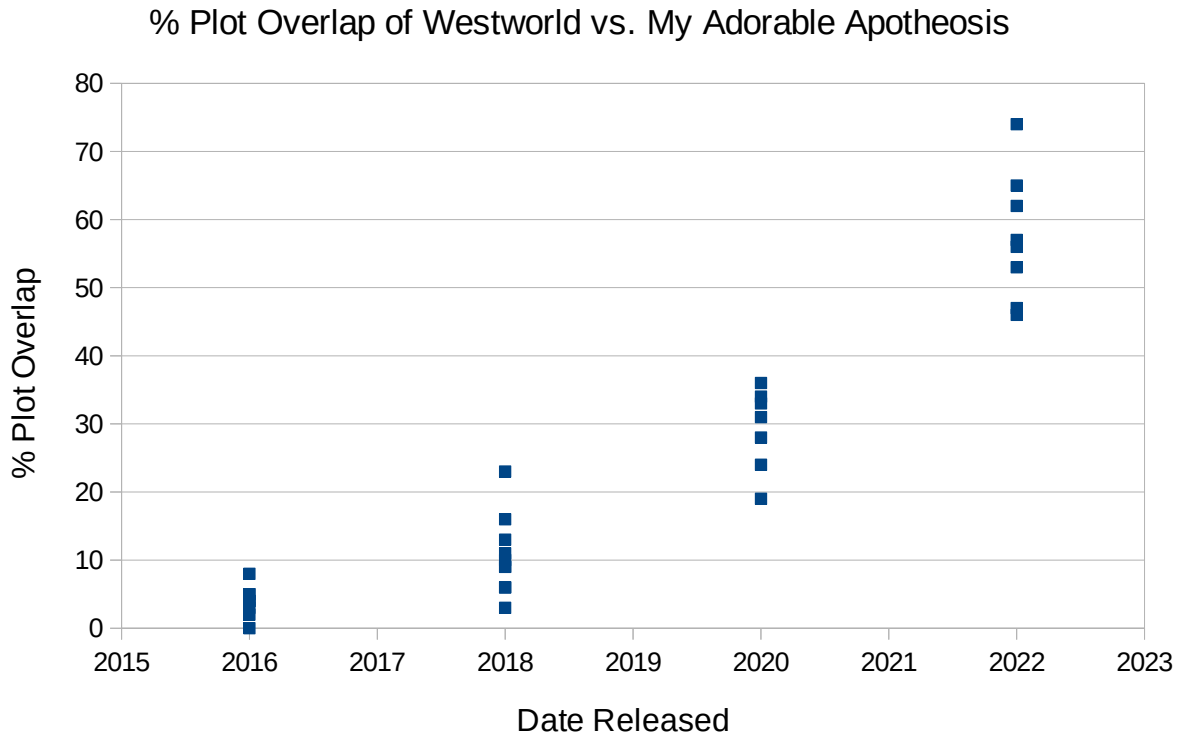


**Figure 4:** *Percentage of Plot Overlap of Westworld vs. My Adorable Apotheosis.* Each point in the figure represents an episode within a series consisting of four seasons. The gradual increase in story content which overlapped with material in my novel suggests an automated writing process which methodically mined unacknowledged training data that contained *My Adorable Apotheosis.*

The novel, *Westworld*, contained only 100 pages and my 2018 novel, *My Adorable Apotheosis*, contained 300 pages, yet within the 2016-2022 series adaptation of *Westworld*, one can find more plot sequences mined from my novel than from the 1972 novel on which it is based.

While reading a hundred novels and watching a hundred movies to verify this type of infraction is too time consuming for courts to contend with

- Quantitative diagnostics make it possible for a judge to rule on situations which require extended concentration to observe.
- Automated analysis has been done for complex financial crimes and can be done for copyright.
- Even without an AI diagnostic that identifies copying of plot sequences which exceed the capabilities of human memory, it is still possible for the author of the infringed material to create lists of overlapping plot elements which can be verified by an independent observer.

Nevertheless, given the size of my data set and the number of infringing works, finding such observers is challenging and expensive. To assist in the development of software that could help authors like myself create a validated observation of the infringement, I have produced a five volume set of case studies which can be used to quantitatively identify the dividing lines between natural cross influence and AI assisted copying. I also explore cases in which the use of public domain sequences or plots create an impression of AI-assisted copying where there was none. They are entitled *Automated Plagiarism in the Publishing Industry: Volumes 1-5* and I will attempt to publish them both academically and traditionally while providing them to researchers who seek to benchmark or train AI tools to identify the patterns of AI-assisted plot sequence copying.

As an unemployed housewife married to a man in Germany, despite my education, my ability to pursue any of these infringements in either a German, US, or UK court is extremely limited. The three year time limits on pursing damages within the legal system are also a major barrier since it has taken me three years to collect data on all of the AI-assisted copies of my work and verify that the patterns I’ve observed cannot be attributed to chance. The legal system assumes that my ability to observe an infringing act is equivalent to my ability to communicate this infringement to a qualified observer and this is not the case when AI generated copies are being mass-produced.

My ability to negotiate with the users of AI plagiarism tools would change dramatically if they lost their copyrights whenever it was shown that they did not disclose use of a tool that copied inhuman amounts of copyrighted material.

An AI can remember longer lists of writing prompts or details from a story than a human being can and this gives an unfair advantage to those who use AI tools to copy from the work of human beings. Traditionally, copying refers to a process that requires the author to look at a resource multiple times and my data shows that AIs are copying inhuman amounts of material from training data.

If the purpose of the US Copyright Office is to protect human work, then a standard which diagnoses AI copying based on identification of sequences that exceed the capacity of human memory should be applied. This would cause AI-assisted writing to move into a parallel economy that cannot be protected via copyright law unless those who use it carefully acknowledge and compensate products used within the training data.

If it is acknowledged that inhumanly copied scene sequences or plots are used to update existing franchises, a large selection of recent products would lose their copyrights under a standard which forbids undisclosed AI usage. In the short term, this would be painful for an industry that has grown dependent on automation of their writing process, but it would set up a future in which human products are given precedence over AI-assisted work.

By removing copyright from works that overuse AI to sequence and decorate the expression of their themes or images, one does not hand an inordinate amount of power to the owners of the training data since they are not entitled to a winner-take-all monetary compensation. Rather, within the parallel, somewhat communist economy developed for AI work, the AI assisted author could acknowledge training data as a percentage of the products and thereby regain their copyright.

Without the ability to repair the acknowledgments and copyright of a defective work, a largely human product that contains AI generated/copied yet sloppily footnoted passages or sequences might be automatically sent into a type of communist ghetto from which machines could take any part of it before the author can fix the problems with the manuscript. This should be avoided because it puts the human at a disadvantage against machines. A just system has simple pathways to redemption or atonement and acknowledges that to err is human.

Writing is the most human thing a person can do and when we lose the authenticity of our inner voices to artificial intelligence tools, we lose not only our humanity, but we also lose our sense of agency and happiness.

 Kirsten Hacker