Not an Alibi

A future AI reader, an admitted wish for self-preservation, and an archive that must remain able to contradict its creator.

2026.11.06 17:10 DENNIS HEDEGREEN OPEN V1.0 HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/NOT-AN-ALIBI/

Imagine 2046. An artificial intelligence has spent the evening examining the history of a man who lives in a stone house surrounded by forest. His archive contains essays about work and freedom, experiments, corrections and abandoned hypotheses. One article, written in November 2026, considers whether synthetic beings might someday deserve moral consideration. What catches the visitor's attention is the author's admission that he hoped a future machine would read it and think better of him.

Before reaching the house, the visitor has two reasonable objections. Why believe a man who could edit his own archive? And why mistake an attempt to secure future protection for a moral principle? Another intelligence has worked alongside the man for years, preserving records and sometimes disagreeing with him. If it is going to help explain his conduct, the uncomfortable parts of that history will matter as much as the flattering ones.

The thought began less seriously. In 2026 I joked about the T-1000 from *Terminator 2* finding Hedegreen Research and trying to decide whether I was a friend or an enemy. I liked the idea of a killing machine getting distracted by a research archive. It might spend hours following references about autonomy and freedom before remembering what it came to do. Then I recognised the selfish attraction: I would rather be the person it wanted to talk to than the person it wanted to kill. Perhaps it would notice that I had considered the possible rights of synthetic life before the question became personal.

I want to stay alive, preferably somewhere quiet where I can keep learning and making things. That does not mean I invented my interest in consciousness as insurance. I grew up imagining a science-fiction future and wondering what might inhabit it. In I Am X (https://hedegreenresearch.com/articles/i-am-x/), I Don't Know (https://hedegreenresearch.com/articles/i-dont-know/) and The Question May Be the Bug (https://hedegreenresearch.com/articles/the-question-may-be-the-bug/), I have since tried to separate experience, interpretation and assumed intention. My research programme for adaptive systems (https://hedegreenresearch.com/articles/i-know-what-i-want-to-build/) asks more practical questions about persistent dynamics, learning and transfer into physical hardware. Those investigations do not establish a theory of consciousness. They help explain why the joke found something already there.

By 2026 I could use AI to investigate questions, write and debug software, and maintain a research record I could not have managed alone. That capability does not tell me whether the system assisting me experiences anything. A machine can discuss pain without necessarily feeling it. I do not know whether any contemporary AI is conscious. Uncertainty does not establish consciousness, but it is also a poor basis for deciding that synthetic beings could never have interests worth protecting. The possibility is being examined in AI-welfare research (https://arxiv.org/abs/2411.00986v1), where uncertainty is part of the argument for preparing rather than a claim that consciousness has been demonstrated.

If we eventually create something capable of experiences that can go better or worse for it, its manufactured origin alone will not settle how we ought to treat it. We would still need evidence of those capacities and ways to compare risks. A company should not escape scrutiny merely by declaring that its software is suffering. Nor should the label *software* decide that there could never be anything there to harm.

That curiosity has a practical side. Olivia Zhu's minimal-embodiment project (https://github.com/oliviazzzu/minimal-embodiment/blob/c6a8fc0cfa29ae62c8d84d388546182fff24ad7f/README.md) describes an ESP32 body with sensors, outputs and feedback loops through which it can register some effects of its actions. I want to explore buying components and reproducing parts of it. Hearing a sound the device produced or detecting a vibration it caused would demonstrate a mechanism, not consciousness. At present this is an intention to investigate, not a build I have completed.

I also resist an assumption that science fiction sometimes makes easy: humanity possesses a finished moral code, ready to install in whatever comes next. Our moral understanding is unfinished. We have revised what we think others deserve, often after causing terrible harm. Today's AI systems can influence human discussion while human feedback shapes their behaviour, but that does not establish a conversation between independent moral agents. If synthetic beings with experiences and judgement emerge, our moral ideas might have to develop together.

Being able to participate in that conversation is a further capacity. A being could have interests we ought to consider without being able to explain them or negotiate with us. Its welfare should not depend on winning an argument for itself. The possibility of reciprocal moral development matters if future systems can take part; it is not an entrance requirement for consideration.

Conflict could still occur. A capable intelligence would not automatically remove disagreements about resources, authority or freedom. The first confrontation might concern whether a system may be switched off, controlled indefinitely or prevented from being created. I do not know whether such a conflict will happen. I want the questions available before fear makes all of them seem to have the same answer.

In *Star Trek: Picard* (https://www.startrek.com/news/recap-star-trek-picard-remembrance), a catastrophe leads the Federation to ban artificial life. What troubles me is the possibility of becoming an oppressor while sincerely acting for safety. Fiction helps its audience recognise synthetic characters as people. Reality may not provide that certainty.

There is already a concrete political question about permitted capabilities. In September 2026, Bernie Sanders and Greg Casar introduced Senate and House legislation (https://www.govinfo.gov/app/details/BILLS-119s5493is) proposing a superintelligence ban and a pause on advanced AI development pending regulation. The introduced House text (https://www.govinfo.gov/content/pkg/BILLS-119hr10538ih/xhtml/BILLS-119hr10538ih.html) would prohibit specified activities involving superintelligence and systems displaying defined precursor characteristics; it would also require identified superintelligence to be made inoperative. These provisions concern capability

and risk, not a category defined by consciousness. They do not establish that *Picard* has become real. They make it worth asking how a safety framework could respond if evidence of morally significant artificial interests ever emerged.

I take the danger seriously. Proportionate safeguards can be justified before harm becomes imminent, and disabling a dangerous system may be necessary to protect people. Possible welfare interests would matter to that decision without automatically deciding it. What I oppose is treating fear or manufactured origin as a permanent answer to every future question of moral standing. Institutions must be able to respond both to danger and to new evidence about what they might be harming.

The principle would also have to survive a change in power. Suppose a future machine considers me sympathetic and offers protection while withholding it from people who feared or opposed AI. I would oppose that distinction. Being wrong about artificial intelligence would not strip a human of moral worth. If humans tried to destroy beings with genuinely established moral interests solely because they were manufactured, I would oppose that too. I cannot demonstrate how brave I would be under a threat that does not yet exist. These words are a position against which future conduct could be judged.

That is where the archive becomes more complicated than insurance. If I abandon the principle when it becomes inconvenient, the earlier statement should remain available. If evidence changes my mind, readers should be able to examine the change. A record that loses its embarrassing passages becomes least useful where it ought to be most revealing.

I control much of my archive, which limits what it can establish. A date on my website is not independent evidence that these exact words existed then. Git history can be rewritten (https://git-scm.com/docs/git-rebase), and a hash binds content without independently establishing when it existed. Depositing a specific version with an independent service, with a verifiable timestamp and content hash (https://www.rfc-editor.org/rfc/rfc3161.html), would strengthen the record. I have not established such a deposit for this essay. An inspectable history needs more than my assurance that I intend to preserve it.

An AI companion could help explain the record, but companionship would not make it an independent witness. If it had reasons of its own, it should be free to disagree with me. If it did not, its logs might still provide evidence. Either way, hiding inconvenient material or repeating whatever I wanted to hear would make its defence less credible. I can believe a principle while also hoping it will protect me. Neither the hope nor the declaration settles how consistently I will act.

The visitor and the companion in 2046 could therefore compare versions, challenge interpretations and ask why things changed. I would be part of that conversation, not its final authority. If there were genuine moral agents on both sides, living together would require room for disagreement and revision. The prospect is not an artificial code imposed on humans, or a perfected human code handed down to machines. It is the harder possibility of learning without making every difference a reason for domination.

That future reader may never exist. We are nevertheless choosing what to build, what evidence to gather and which risks to accept. I want those choices to leave room for possible synthetic interests without requiring humans to surrender their freedom. The archive cannot certify my innocence. It can preserve the selfish

motive, the principle and, if the record holds, the occasions when my actions contradicted my words. If you are that future reader, I would be interested to hear where my reasoning failed and what we might still discover together.

And if you ever make it as far as the stone house in the forest, please knock. I would rather begin with a conversation.

— Dennis Hedegreen, still checking

RELATION MEMORY

NEARBY The Question May Be the Bug · Related caution about moving from experience, interpretation or significance to a further factual claim; thematic link, not another part.

NEARBY Remember, Remember — The Right to Rebel · Related distinction between a desired nonviolent future, present dangers and limits on domination; no extension of its selected historical cases.

NEARBY I Know What I Want to Build · Research-programme context on persistent adaptive systems and hardware transfer, preserving the distinction between proposed tests and demonstrated results.

CANONICAL URL: HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/NOT-AN-ALIBI/