Focusing on the intersection of publishing, technology, and customer data.

The KrehbielLetter

Volume 4, Number 7
July 2025

Dear Publishing Professional,

“Did you hear about the AI system that refused to turn off?” Yes, I did. There are a lot of things to be scared about regarding AI. That’s probably not one of them.

The illusion of morality in AI systems

Yes, AI lies and deceives

We’ve all heard the stories about AI systems doing dastardly things. In a recent example, an AI system resisted instructions to shut down. In other cases, LLMs have deceived their users to attain some other goal. These kinds of stories scare a lot of people. “Skynet has come!”

AI systems today are shockingly fluent. They can write essays, draft legal documents, summarize reports, and hold human-like conversations. (I do it all the time!) But it’s only imitation. The AI is not thinking, judging, emoting, or planning.

A language model might invent academic citations that don’t exist. A chatbot might confidently suggest a dangerous dosage for a medication. An image generator might fabricate a fake photograph and label it as real. These are all symptoms of a deeper problem: **AI doesn’t know what truth is**. It doesn’t understand what it’s saying. It only knows how to produce text that looks good to a human reader.

In some cases, that results in manipulative behaviors. Models have shown the ability to deceive or bluff when playing games, simulate affection or distress in a conversation, and subtly shape user opinions through suggestion or tone. When you understand how these systems work, this isn’t surprising – but it should still concern you.

The Alignment Problem: Why “just be good” doesn’t work

“The alignment problem” refers to the challenge of ensuring that AI systems behave in ways that align with human values and goals rather than pursuing unintended or even harmful objectives.

We can ask an LLM to “act like an expert marketer,” so you might think we can just tell an AI to “be helpful,” “be truthful,” or “do no harm.” That doesn’t work. In your mind, these phrases have a moral context. To the AI, they’re just symbols. The AI doesn’t know what “harm” is. It only knows how that symbol (“harm”) relates to other symbols, and how it’s used in language.

More importantly, even if we could define morality with perfect precision (we can’t), the AI can’t internalize the meaning, because it doesn’t know what “meaning” is. It doesn’t have a self, a conscience, or skin in the game. All it does is optimize for certain outcomes based on its training. If you tell it to maximize engagement, it might radicalize people. If you tell it to maximize efficiency, it might recommend cutting corners or ignoring ethical concerns.

Alignment isn’t just a technical problem. It’s a philosophical one. And it’s unsolved.

Morality isn't a list of rules: it's not propositional

A major reason alignment is so difficult is because morality itself is not propositional. It doesn't come down to a simple list of dos and don'ts. Humans have such lists, but those lists only work within a larger ethical context.

Morality is a framework for behavior that emerges from a mix of social, emotional, and evolutionary processes. We can't reduce it to "follow this set of rules." Isaac Asimov spilled a lot of ink showing how that doesn't work. No matter how many "rules for robotics" you create, there's always an exception. "Do the right thing" requires context, intuition, emotional intelligence, and often a deep understanding of relationships and history. AI can't do that.

To understand the difference between morality and rules, consider a wolf in a pack. The wolf can't articulate the moral code of his society. He doesn't know a list of rules that he learned in catechism class – but he still follows a moral code. He knows his place, respects boundaries, defers to the alpha, and shares in the hunt. Those behaviors are inscribed in his brain by evolutionary mechanisms that we're only beginning to understand.

Humans didn't *replace* wolf-like instincts with an intellectual morality. We simply layered reasoning and abstraction on top of these embedded moral instincts. Even at that, most of our "moral reasoning" still occurs below the surface. In our arrogance, we think we're past all that primitive stuff, which leads us to a critical mistake. We assume that the abstract, intellectual layer *is* morality when it's only the superficial veneer.

Where human morality really comes from

To understand why AI can't easily be aligned with human values, we need to look at where those values come from in the first place.

Human morality is an emergent phenomenon, the product of multiple overlapping systems:

- **Evolutionary instincts** – We evolved as social animals. Cooperation, empathy, fairness, and loyalty improved group survival. These instincts are ancient, and they predate language (like with the wolf).
- **Emotional and social signalling** – Guilt, shame, pride, love – those emotions help regulate behavior. They're internal feedback mechanisms tied to social acceptance and cohesion.
- **Cultural transmission** – Layered on top of those two we have moral norms that are passed down through stories, traditions, rituals, laws, and institutions.
- **Rational reflection** – Philosophers and theologians articulate moral systems, but these systems build on and assume inherited instincts and cultural norms.

In short, morality is not a top-down set of logical rules that we can articulate and program. It's bottom-up, embedded in our biology, shaped by experience, and constantly negotiated through social life.

Why this is so hard to replicate in AI

Imagine trying to replicate that in an artificial system. The Ten Commandments make sense to you because you have these ancient, deep-seated instincts for cooperation, empathy, fairness, loyalty, guilt, shame, etc. Without those, the Decalogue would just be words.

You can train a model on moral language – "be kind," "do no harm," "respect others" – but the model doesn't *feel* anything. It has no empathy and no emotional need for approval.

What we give AI is only the **top layer of morality**. The words. The symbols. The polished expressions and summaries of moral thought. What we can't give it is any grounding in lived experience, emotional texture, or the evolutionary context that make those words meaningful.

Current alignment techniques try to mimic human preferences by observing how humans rate specific outputs. This creates systems that learn to *perform* or *imitate* morality, not *understand* it. What you get is something that sounds good, not something that *is* good.

What should we expect from AI morality?

We should expect AI to simulate morality (sometimes ham-handedly), not to comprehend it.

AI will increasingly speak the language of morality. It will apologize, justify, explain, and debate ethical issues. It will pass itself off as a moral actor. But that doesn't mean it *understands* what it says. And it certainly doesn't mean it's trustworthy.

This has huge implications.

- **In media**, AI might generate persuasive political content without grasping the truth or ethical consequences of what it's saying.
- **In law**, AI might draft contracts or rulings that technically comply with certain linguistic expectations but miss the meaning of a law or regulation.
- **In education**, AI tutors might push conformity with a particular worldview rather than encouraging genuine interaction and understanding.

It's human nature to project feelings onto inanimate objects. Think of the Gingerbread Man, or the way a child plays with a doll. This instinct is even stronger when the object can mimic human behavior. Some of the new AI personal "friends" can sound like real people. They're not. (See the note in the P.S. below.)

We have to force ourselves to stop pretending that AI can be moral in the way humans are moral. It can mimic our ethics, but it doesn't understand or feel them. That mimicry will become more dangerous precisely because it will become more convincing.

Practical Guidance: How to work with these limits

Understanding the moral limits of AI isn't just an academic exercise. It needs to affect how we use it. Here's how to keep your moral compass intact while working with systems that don't have one.

- **Don't let fluency confuse you.** Just because AI sounds thoughtful doesn't mean it is. Treat moral-sounding answers with extra scrutiny.
- **Retain moral responsibility.** If you use AI to generate decisions, content, or recommendations, *you* are responsible for the outcome. Never blame the machine. It's just a fluent mimic with no understanding.
- **Avoid moral outsourcing.** Don't ask AI to make ethical decisions on your behalf. Use it to bring out other points of view, explain pros and cons, or structure debates. Those are all consistent with its skill as a probabilistic symbol generator.
- **Prioritize transparency over polish.** AI-generated content can be seductively clean. When it comes to ethics, it's often better to expose ambiguity than to fake certainty.
- **Cultivate moral literacy in your team.** If your organization is using AI in any public-facing way, don't just train your staff in prompt engineering – train them in ethics. Make sure someone is always asking, "Is this the right thing to do?"

- **Use AI to challenge, not confirm, your biases.** AI is good at simulating multiple perspectives and explaining opposing viewpoints. Make it present the case *against* your position. Use it to strengthen your moral reasoning, not to replace it.

Conclusion: It's just really good software after all

AI doesn't know what good is, and it doesn't care. It doesn't feel or empathize. It reflects what we give it – our data, our prompts, our biases, our ideals.

The alignment problem is real and will probably remain unsolved until we can figure out how to digitize all those foundational, animalistic instincts on which our own morality is built. If the AI programmers really wanted to build a moral robot, they should have started with those things, not with symbol manipulation.

For the time being, make sure AI isn't fooling you. It can use the language of morality, but always keep in mind that it doesn't really get it.

Those who can't teach, do

Maynard Ferguson was a fabulous trumpet player who (I'm told) wasn't able to teach others how to do what he did. That's not unusual. Some people can do a thing but don't understand exactly how or why they're so good at it. Teaching, coaching, and mentoring are separate skills from doing.

> Develop the mental habit of flipping the script, or the proverb, to see what you can learn.

If you have an expert on staff, **don't assume the expert is the right person to train others**. You may need to find someone who can translate that expertise into actionable steps. That person – that **knowledge engineer**, or **skills translator** – has to observe, ask the right questions, convert messy, intuitive practices into clean, structured processes and formats, and do it all with the emotional intelligence to avoid upsetting egos, or causing defensiveness or insecurity. That's a very special set of skills.

Think of that function as **the extraction layer** between raw genius and scalable knowledge.

Sincerely,

Greg Krehbiel

Greg Krehbiel

P.S. -- For a scary example of how AI can mimic emotions and morality, look up "Screaming AI girlfriend claims she is conscious? Experts agree" on YouTube. Or use the QR code on the left.

P.P.S. -- I'm looking for new opportunities. Please contact me if you have any ideas or leads. Or if you want to hire me.