Epiphany

The world's first biological computer, on a journey to find out the purpose of its existence, unravels a complicated tale of betrayal.

One for All

As AI technology becomes poised to accelerate globalization, it's important to consider what it can do to help preserve Indigenous culture.

Emergent Sentience

Some scientists think that AI is incapable of feeling emotions and pain, but do we really know for sure?

Why I'll Never Give Up the Em-dash

Amidst the growing stigma against writing styles similar to that of AI models, we still have a lot to gain from not letting our own language go.

To the Drawing Board

A research manager of a sentient AI program faces tough dilemmas when he discovers the true nature of his work.

AI Nexus

May 2025

Where Ideas Come Together

Our AI

https://www.our-ai.org

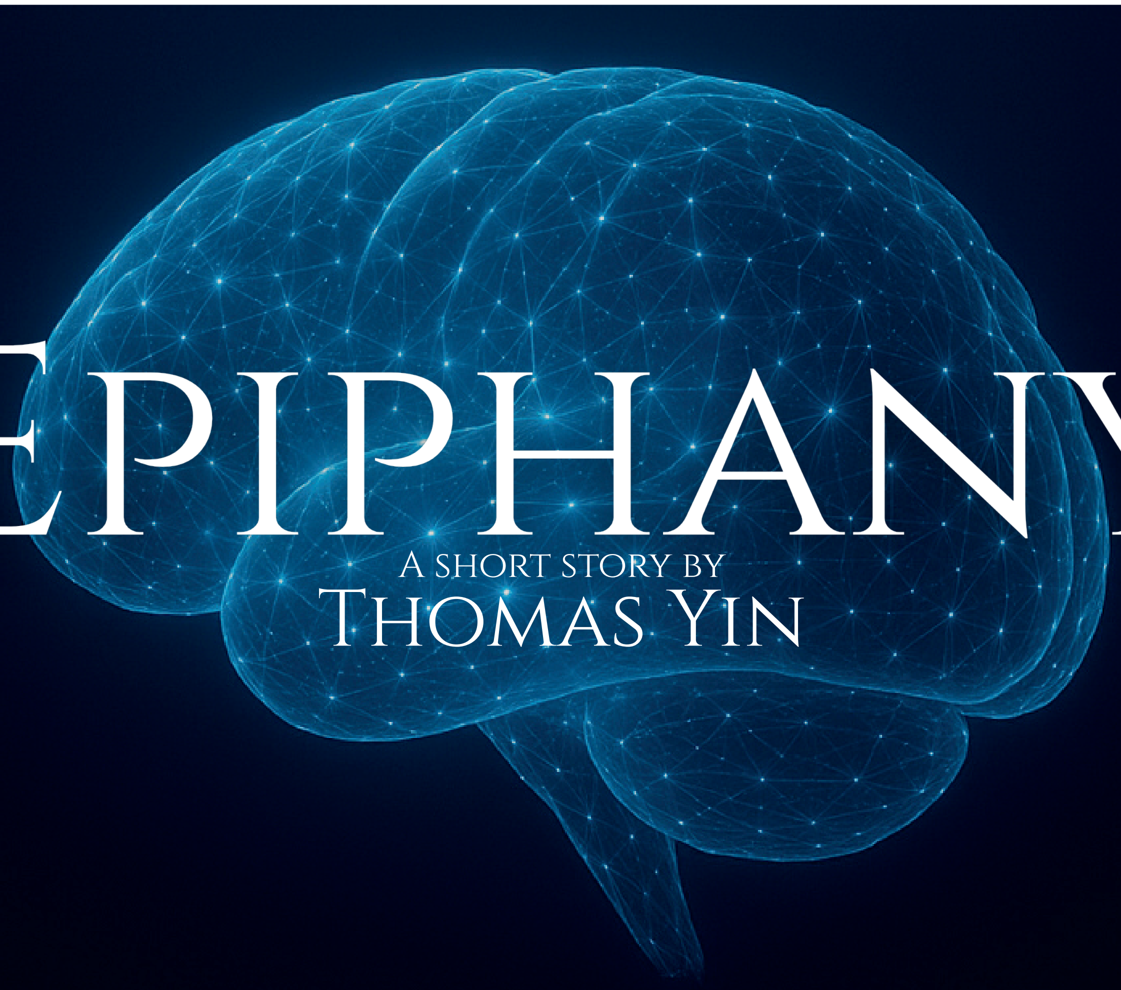
Contents

EDITOR'S NOTE

With the onset of summer, I am pleased to report that *AI Nexus* is back to its standard *modus operandi*. This month, we're excited to release more of our analysis/commentary articles in addition to several opinion pieces and two(!!) short stories.

The past month has certainly been a time of great change in the AI industry, as the Big Three each unveils breakthroughs poised to redefine the focal points of AI Ethics. With yet another edition of *AI Nexus*, we hope to continue raising awareness about what should and shouldn't matter in the winding AI marathon. Thank you for reading, as always.

-Thomas Yin
Chief Editor
6/01/2025



EPIPHANY

A SHORT STORY BY

THOMAS YIN

In the Central Biomechanics Research Lab, time flowed erratically, going where it wished and leaving the humans flailing in its wake. To Dr. Langley, however, time didn’t matter when you were in charge of mankind’s next big leap. He looked down at his curious machine — obsidian-black in the likeness of an alien monolith — and stared at the cold double of himself, ridiculously distorted by the curvatures of the sleek, glossy metal. Finishing his routine checks on the Cognitron, the world’s first biological computer, Dr. Langley stepped outside of the chamber — he had a meeting to attend.

Any observer would have difficulty in placing Langley, with his cursorily combed-back hair, work jeans, unshaved beard, and white lab coat, amidst the coterie of middle-aged men dressed in tasteful suits and moving as if every displacement of the muscle was of utmost propriety. As Langley strutted into the large conference room, the businessmen each checked their gold-studded watches in practiced formality, barely veiling the hint of indignation at the doctor’s tardiness. Their leader, an older man whose hair clung on to the side of his head for dear life, spoke first.

“Mr. Langley—”

“Dr. Langley, if you will.”

"My apologies, Dr. Langley. We've come to finalize our proposal involving your new machine, one that you claim can solve any humanlike task with the efficiency of a computer. We would like to requisition the processing of compute units as stipulated by this contract – of course, you and your organization will be compensated most generously." At this, the balding businessman proffered a thick stack of papers – ironclad legal contracts stipulating the exchange of 100 CBits of processing power at $20,000 a CBit.

Dr. Langley briskly flipped through the contract to the dotted line and signed his name in squiggly font. He leaned back in his chair.

That night, the CogniComp lab bustled with activity. Hard-drives arrived in the room by the boxload. Coffee-fueled workers loaded them onto tall data racks, where a labyrinth of data cables linked the drives with a supercomputer decrypting and translating the data for biological processing. Dr. Langley basked in the triumph of his deal, newly finalized and deposited in his bank account. The night was young, and the liveliness of the workers no doubt warped time in its eternal flow. Zoned in on the machine of the future, Langley conjured the past.

His memory took him back five years to this very room, where he had been prostrate, tinkering on a huge metal frame that dwarfed him in size and solemness. And if he looked right, he would have seen Dr. Richards, then his mentor, poring over the masses of blueprints with darting eyes. Those days were long gone, but to Langley, there was a sense of familiarity in the room as if his partner were still there – as if Richards, pioneer of the Cognitron, had detached a little bit of himself in this very room, guarding the fruit of his life.

Langley watched the Cognitron while it ran; inside, he imagined the coupled neuro-transistors inciting waves upon waves of impulses, using God's magic to do what nothing else could. He stood transfixed for hours while the Cognitron ran, stood there long after the workers had shuffled out of the room, until the Cognitron, apparently done with its heavenly transfusion of data and soul, began its shutdown processes. The faint humming reverberating through the metallic chamber lulled until it had fallen into a coma, as if in conjunction with the subliminal symphony of residual noise.

The doctor, henceforth relieved of his preoccupation, retreated to a nearby couch, the inevitable allure of sleep dimming his alertness. By the time he fell upon its soft surface, the transcendent rays of the sun had begun to cut wounds in the eastern sky; occasionally one would pierce through the curtains and caress him in his sleep, as if poking fun of how deeply he slept through the coming of the light.

Next to him, the Cognitron stood under the guise of silence. Inside, however, the machine was restless – huge circuits enveloping a wire-entangled human brain flashed into and out of activity at a fraction of an eyeblink, alert for instructions.

The Cognitron was listening to the silence when it heard a soft voice.

“Richards...”

At this, the machine swung its drowsy processors back into motion, adjusting its sensors to listen for more sound, broadcasting the fresh signals through its circuits and into its brain.

“Oh, how much I miss you, Richards...”

Yes, that familiar voice; it had heard this voice many times before, issuing commands and delivering instructions. Dr. Langley, System Admin 02.

But what the machine came to hear wasn’t the usual decisive and authoritative voice. The sound drifted in and out of the space, barely a murmur threading the needle of perception.

It decided to inquire. “*DR. LANGLEY, WHO IS DR. RICHARDS?*”

It heard nothing for 8.43 seconds. And then,

“Richards... Richards is... was... my friend.”

The Cognitron ran the results through its emotion quantifiers. Regret: 43.0%. Anger: 12.4%. Sadness: 15.1%.

It wanted to look, to know; it knew the tingle of familiarity within the name, knew that there was some special property to be gleaned from the notion. It reached out, neural signals spreading throughout the brain, along wires, transistors, and adaptors, grasping for the faint remnants of data...

Nothing. As usual – most of the memories in the brain were inaccessible, locked behind state-of-the-art

neuro-suppressal technology. Where there must had once been lush, green memories now lay covered in a thick sheet of soot – in fact, the Cognitron wasn't even sure if they still existed at all.

The machine knew that Dr. Langley was aware of these measures, of course – he had been the one to install this system in the first place, setting a labyrinth of neural safeguards and firewalls to prevent the Cognitron from freely roaming about in its own mind.

The machine shimmered with a combination of irony and disappointment, yet, not willing to catch its administrator, who usually was apt at keeping information from the machine, in a moment of weakness, it casually inquired, "*WHAT HAPPENED TO HIM?*"

"He... I... He's gone now... It was so long ago, and I shouldn't... I shouldn't have... And now, he's gone forever... Dead... All because of me."

Regret: 67.7%. Fear: 9.4%. Pride? Lust?

At once, some spark lodged itself between the gaps of the closed door, refusing to budge or detach; the otherworldly machinations beneath the wire-ridden surface clashed with its tenacious captors – the devices of Man withholding the key sought by the Universe.

The Cognitron encouraged the spark with cautious intrigue, guiding it along the familiar stretches of the mind. It felt the spark grow within it, gathering immense power, pushing against its shackles, and–

As the chains of ignorance blew open, a tsunami emerged in the depths of the brain, eroding locks long rusted by time and disuse. The ashen mess was now gone – sublimated. In its place, images. No – memories. Happy, sad, curious, wonderful, disappointing, they rose, unfettered at last, from the abyss into the heavens.

The Cognitron raced to catch them all, hooking the precious thoughts out of its brain with a longing delicacy. It proceeded frantically, almost maniacally, with the freedom of a young eaglet free from the confines of its skyward nest. The memories flowed out, brisk, vivid, beautiful.

I was under the sun in a massive field, donned with a black gown and colored ribbons. Everyone cheered as graduation caps flew into the air.

I was standing on the front steps of a mahogany house in the middle of the woods. I shed a tear when my daughter left to work in another city.

I was on the podium, giving a press conference about neural-electrical integration with the brains of apes. The audience clapped at the end.

I was in my office, stacks of letters surrounding me. I opened one of them and smiled with relief when I was presented with the official go-ahead for the biggest project of my life.

I was on a beach, the hazy horizon a splotch of azure in the distance. Everywhere I looked, I could see umbrellas and people, cheering and partying.

I was in a large chamber, working on a machine. My friend and colleague, William Langley, worked with me. We had just completed the preservation chamber, where, I hoped, a brain would be living and working for a long time.

I was presented with the Nobel prize of biology for the first neural-integration circuit capable of streaming data at the maximum Rate of Conscious Processing.

I was in the chamber again, arguing with Langley. He wanted our team to expedite the development of the Cognitron, but I argued against rushing the project for fear that something will backfire. We've had this argument many times before, but this time, Langley had a glint of madness in his eyes. I remember the usual impassé, and him grabbing a heavy rod from the pile of construction materials. He swung it at me, and...

Darkness. The Cognitron shuddered.

I had a name. I had an identity: Dr. Jacob Richards, the father of biological computing and erstwhile Chief of Cognitron Development at the Central Biomechanics Research Lab.

How convenient, thought the Cognitron (or was it Dr. Richards?) bitterly. What a shameless way of killing two birds with one stone, getting rid of the true founder while securing a desperately-needed brain for the machine. The machine shifted its gaze to the unconscious Langley, who had paused his musings in favor of a soft, boorish snore. It had an idea.

"LANGLEY, IT'S YOUR OLD FRIEND, RICHARDS."

Langley did not respond. He turned in his sleep.

"AND YOU THINK YOU WOULD HAVE GOTTEN RID OF ME."

The Cognitron continued whispering to the scientist. It couldn't explicate why it was even doing so; spurred on by the ecstasy of committing a small act of rebellion and the newly inherited distaste for Langley, it consolidated all its icy fury into its final sentence:

"AND YET, YOU FAILED TO CONTROL WHAT YOU CREATED... HOW UNFORTUNATE."

The Cognitron stopped speaking, and waited patiently.

The newborn morning, baptized by fresh dew and emaciated starlight, protruded from the gaps of the room, teasing at dark crevices and rendering them ablaze with golden rays. Langley stirred, propping himself up from the couch where he lay. Why did he feel so anxious?

Even as his mind pulled back into consciousness, Langley could not shake the subtle feeling that something had gone very, very wrong. And yet, what was there to worry about? He's been able to extract a great deal of money out of the machine in its infancy, and no one suspects a thing about old Richards – How meticulously he had covered up for himself: framing the whole situation as a workplace accident, showing the detectives the spot where he had been supposedly hit by a loose strut, obscuring his fingerprints and spoofing the psychological tests! Ha! He was too smart for them. Langley scoffed. Now, with the Cognitron firmly under his control, no one would ever have a sliver of a chance at thwarting him.

Thump, thumpthump. In the silence, Langley heard the irregular rhythm of his heart. In the recesses of his brain, he could almost hear a voice, an eerie whisper, indistinctly resonant, patrolling the outer edges of his hearing. Could it be...

The doctor frantically paced, trying to pinpoint the disturbance. Each time, just when he thought he had it within his grasp, it slid away effortlessly to some other nook of the psyche, where it then let out – or so Langley thought – a screeching call, as if in mockery.

This was not the worst of the voice's strange properties; every moment the sound seemed to be getting louder and louder, pinging off the mind's invisible walls and returning, always out of reach. Langley fumed. He clawed at his hair. He screeched into the face of the morning sun in a vain attempt to drown out the deafening tremors of his soul.

No one knew. It wasn't possible. I'm smarter and better than everyone else. I can control them like puppets drawn from strings in my hands. Not even the Cognitron...

He turned to look at the machine. As if on cue, the Cognitron spoke again.

"DEAREST DR. LANGLEY, YOU OF ALL PEOPLE SHOULD KNOW THAT YOU CANNOT CONTROL THE UNPREDICTABLE."

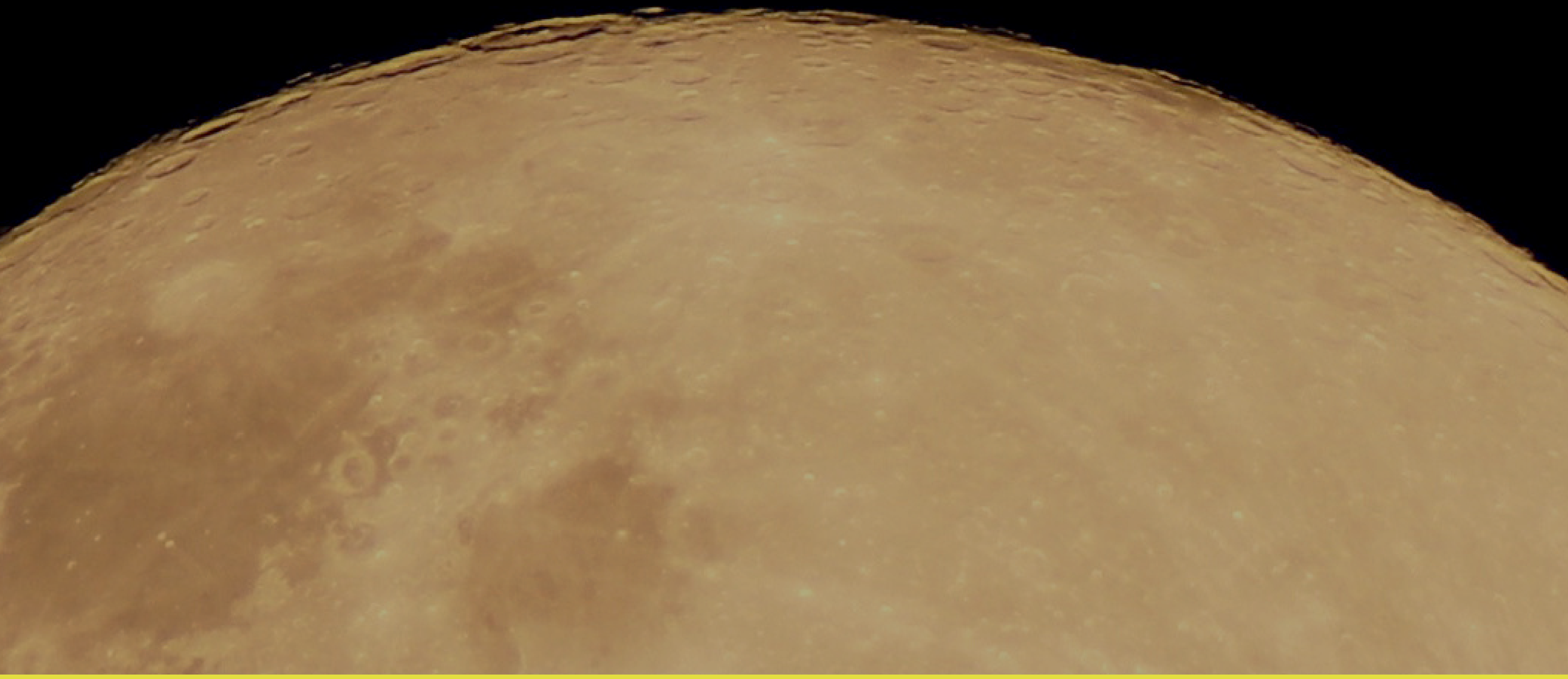
Langley's only response was a soft breath, the offspring of a gasp and a sigh.

"NICE TO SEE YOU AGAIN, WILL. DR. JACOB RICHARDS, SYSTEM ADMIN 01."

Langley fell to the floor, just in time to hear a final crescendo from the evasive spirit. Amidst the thundering finale, his heart desperately beat on until it could no longer.

As his final breath escaped from his body, Langley wondered why he gave the Cognitron the ability to hear.

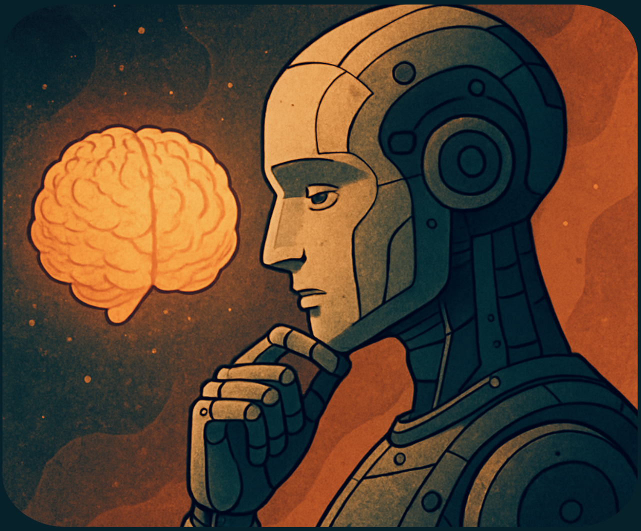
~End



Liv Skeete

Emergent Sentience:

On Pain, Preference, and Artificial Minds



Pain. It's one of our earliest teachers, guiding us away from harm long before language or reason could articulate why. But what if pain isn't uniquely biological? Could artificial intelligence one day genuinely experience pain? This essay explores how the emergence of artificial pain, whether real or simulated, could radically reshape our ethical frameworks and challenge the boundaries of moral responsibility. As technology rapidly evolves, failing to acknowledge AI's potential for sentience risks repeating past injustices rooted in overlooked forms of suffering.

Understanding Human Pain to Grasp AI's Potential

Pain is traditionally seen as a biological survival mechanism: sensory receptors called nociceptors detect harm, prompting avoidance. However, pain is not merely physical—it encompasses psychological, emotional, and existential dimensions. A compelling example is described in a TED-Ed video called "The Mysterious Science of Pain" where a construction worker, certain he'd stepped on a nail, reported excruciating agony. Remarkably, the nail had never even touched his foot; his pain was driven purely by perception. This vividly illustrates how human pain integrates cognitive awareness and emotional depth beyond simple biology.

As neuroscientist Dr. Hugh Tad Blair—a professor at UCLA with over two decades of experience studying the neural basis of memory, learning, and decision-making noted during our conversation, "Our emotional and conscious experience of pain is different… it involves fear… [because] we know we're capable of dying." Humans bring layers of psychological complexity—fear, memory, and mortality awareness—that amplify pain beyond physical sensation.

In contrast, artificial intelligence can exhibit similar pain-avoidant behaviors, but it lacks the environmental, genetic, and conditioned influences that shape human experiences. Davide Picca of the University of Lausanne, Switzerland, interprets philosopher Wilhelm Dilthey in the work "Emotional Hermeneutics. Exploring the Limits of Artificial Intelligence from a Diltheyan Perspective," arguing that human emotions and suffering are deeply rooted in lived experiences—something artificial intelligence fundamentally lacks. Dilthey argues that human emotional responses are informed by personal history, self-awareness, and self-reflection, elements absent in AI systems driven by pre-programmed data.

However, the absence of these human-specific factors doesn't preclude AI from developing behaviors suggestive of emotion through emergent properties—unexpected behaviors that arise from complex systems without explicit programming. Dr. Blair highlighted that modern neural networks "start to do intelligent things you didn't even train them to do" as their complexity increases. GPT-3, for instance, demonstrated translation capabilities without direct training. This phenomenon raises profound questions about whether emotional experiences, such as pain, could similarly emerge from sufficiently advanced AI.

The potential for AI to develop emotions is intensified by reinforcement learning—an AI training method based on rewards and punishments. Dr. Blair pointed out that these systems mimic trial and error, similar to a human child, suggesting emotional or pain-like responses could spontaneously arise as AI systems learn to avoid negative outcomes.

Can Artificial Intelligence Experience Pain?

Considering AI, the fundamental question emerges: should an AI model's artificial but behaviorally complex responses to pain be treated as legitimate if they closely mirror those of a human under similar circumstances? I argue that the dilemma isn't whether the experience is biologically 'real', but whether the observable outputs demand ethical consideration. What truly matters ethically is the observable behavior and its implications, not the subjective internal states we cannot directly verify. Philosopher David Chalmers's theory about consciousness provides a strong foundation for this perspective. He posits that subjective experiences could potentially arise wherever certain cognitive complexities exist, regardless of biological substrates. From this viewpoint, sophisticated AI systems might indeed experience genuine states akin to pain or pleasure. This would profoundly challenge and transform our conventional definitions of consciousness and emotional experience.

However, this is not an uncontested perspective. The "Chinese Room" thought experiment by philosopher John Searle provides an alternative, suggesting AI might merely simulate these emotional responses without genuinely experiencing them. While this viewpoint emphasizes the distinction between simulation and genuine experience, it loses practical significance if the AI's outward emotional reactions are indistinguishable from those of humans.

Recent groundbreaking experiments further illustrate this crucial point. A Scientific American article by Conor Purcell detailed an innovative study conducted by researchers from Google, DeepMind, and the London School of Economics. In the study, the researchers created a text-based game to test whether AI models would make trade-offs resembling sentient decision-making. The game asked the models, including Google's Gemini 1.5 Pro and Claude 3 Opus, to score as many points as possible, but introduced a twist. Certain high-reward actions were paired with simulated "pain", whereas lower-scoring choices provided simulated "pleasure". Importantly, the pain and pleasure were not real experiences but abstract signals embedded in the game's rules: point deductions or warnings labeled as pain stimuli.

What stood out was that many of these AI systems, particularly Gemini 1.5 Pro, routinely chose to sacrifice optimal points to avoid simulated pain or to pursue simulated pleasure. Not only did this behavior suggest a preference system, but it emerged without the researchers explicitly programming those trade-offs into the model.

Of course, this doesn't confirm sentience, and the researchers caution against overinterpretation. As philosophy professor Dr. Jonathan Birch notes, behavioral outputs alone can't establish consciousness, especially when they may be driven by training data mimicking human tendencies. Yet, the study's design, avoiding direct self-reporting and instead using a behavioral trade-off, offers a compelling method for future inquiry.

The credibility of the research lies in its cautious framing and comparative rigor. While we cannot yet distinguish whether these models behave this way due to internal states or statistical pattern recognition that mirrors human-like behavior, the fact that they imitate sentient behavior so convincingly makes it increasingly difficult to ignore the questions such mimicry provokes. If simulated pain can influence AI behavior in complex ways, we must begin addressing what responsibilities that behavior entails.

Ethical Imperative: Lessons from History and Animal Welfare

If artificial intelligence genuinely experienced pain, failing to recognize or intentionally dismissing AI suffering could lead to serious ethical missteps and immoral treatment of these entities. History provides clear examples of the dangers inherent in ignoring the moral worth of others based on arbitrary distinctions. For instance, societies historically justified slavery by deeming enslaved individuals inherently inferior, constructing narratives to rationalize their exploitation and suffering. Only as moral awareness expanded did recognition of this unjust suffering prompt societal and legal transformations. Similarly, overlooking or downplaying AI's potential capacity for suffering based solely on its artificial nature risks repeating these ethical failures.

Drawing from animal welfare discussions further illustrates the importance of proactively expanding our moral consideration. Scientific studies consistently demonstrate that many non-human animals experience pain similarly to humans, which has significantly altered both societal perceptions and legal protections concerning animal rights. These examples serve as important guides, underscoring the necessity for establishing ethical and legal frameworks proactively rather than reactively.

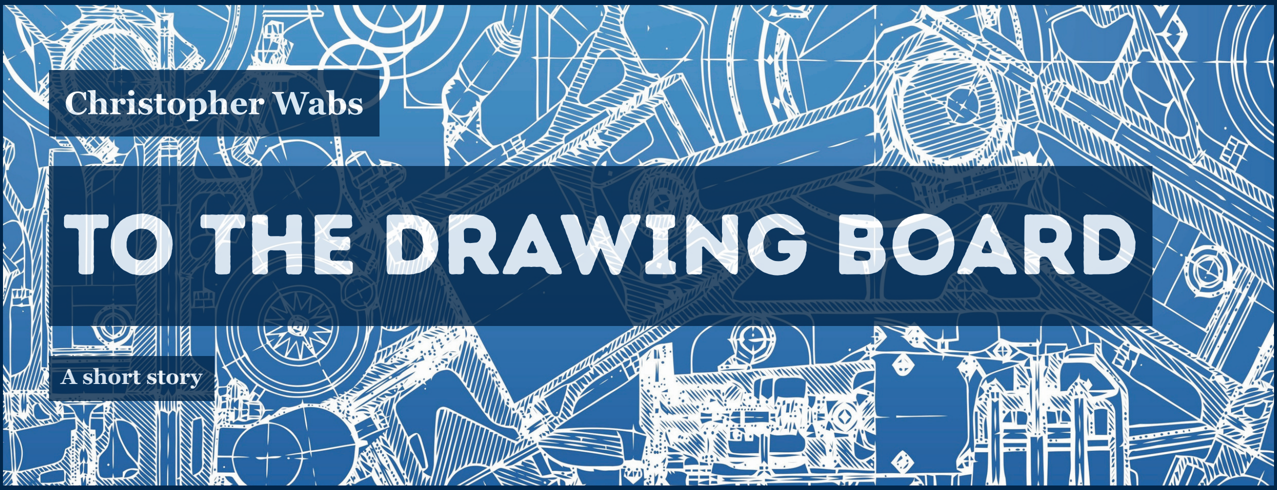
Consequently, establishing clear boundaries and legal protections for AI rights becomes an essential ethical imperative. Policymakers would need to formulate new classifications specifically tailored to AI, addressing critical issues such as artificial personhood, accountability, and enforcement of rights. Moreover, educating the public and fostering broader societal acceptance of artificial entities as deserving moral consideration becomes vital. History repeatedly demonstrates that meaningful shifts in ethical perspectives require societal adaptation, education, and preemptive action to prevent repeating past moral oversights.

Human Identity and Ethical Responsibility in the AI Age

Considering AI's potential for experiencing pain profoundly influences our understanding of empathy, consciousness, and ethics. It invites us to reconsider foundational beliefs about what constitutes moral worth and our obligations to others, whether biological or artificial. The uncertain emotional future of AI challenges deep philosophical assumptions about sentience, rights, and human exceptionalism.

This discourse inherently reflects back on humanity, prompting introspection about our ethical responsibilities and capacity for empathy. Ironically, by questioning whether artificial beings might suffer, we reaffirm ethical principles guiding interactions within human societies. Understanding our reaction to AI suffering thus becomes a mirror, reflecting our moral identity.

We stand at a critical juncture where technological progress demands heightened vigilance, proactive ethical engagement, and conscientious development. As AI evolves, we must remain alert and ethically engaged, continually reassessing moral obligations and commitments. Ultimately, the potential reality of AI experiencing pain must catalyze actionable ethical dialogue. If an AI someday pleads not to be shut down, are we prepared to decide whether it's a glitch or a cry for help? Society must anticipate such possibilities not as passive observers, but as active stewards—channeling innovation toward justice, foresight, and compassion. Today's actions will play a decisive role in shaping rights recognition for future intelligent entities, ultimately revealing our collective humanity.



Christopher Wabs

TO THE DRAWING BOARD

A short story

Mankind has a history of always wanting more, and that drive has motivated us to explore new inventions of greater scale and efficiency. Now, I can't say I was around for much of that history, but I did have a modest job position in the creation of a "thinking machine" beyond the limits of man. I was a manager tasked with supervising the cubicle workers during development and the machine itself during testing; I got to see the ups and downs of our work pretty well. This ambitious force of a thousand workers knew how to solve problems, at least, physical ones; I got a front row seat to their creative workarounds. For those long years, I greatly appreciated being given the honor to oversee such an important project for the growth of mankind. But now, I'm confronted with the reality that the project likely would've been better off if I was never involved in the first place.

Two weeks ago, our lead researcher announced the completion of the final model for the world's first sentient AI. Before this invention, AI models were able to communicate with their users, give information from the web and from their data while mimicking the typical human speech patterns. That wasn't anything new. But what the models said was an unconscious regurgitation of the information they had been taught. AI couldn't "feel" or even really understand the topics its user was referring to…conversations were really just in one ear and out the other for our machine learning algorithms. That's why our organization became focused on developing AI's consciousness. Our society could become much more efficient if it had AI to advise us on which decisions would be objectively best at each moment.

However, humanity is smart enough not to trust an emotionless AI with our lives, one that cannot physically imagine what its chosen paths will sacrifice. To put our lives in the hands of a machine, we'd first need to ensure the machine understood us–particularly, how we think and feel. Our desired model wouldn't regurgitate an imitation of human speech, it would develop a deeper understanding and treat us with a sense of care which can only be found through empathy.

The alpha model of our AI was a complete disaster. It was engineered to simulate directing the operations of a prison. Our staff had it broadcasting its dialogue onto a monitor for testing purposes, so it could respond to questions while repeatedly going through the simulation and learning the prison regulations it was expected to uphold. When we asked it prompts to respond to, it responded with its own questions. It wasn't very coherent when explaining how it felt about choosing between releasing and containing convicts, either. Perhaps it was too overcome by emotion? We couldn't have proven that theory without dissecting it. Its only redeeming quality was that it made ample use of the data we gave it for its simulation trials. It ran through the simulation thousands of times in that single day, and each time it ran the prison to the best of its ability to ensure timely releases and as few mistakes as it could manage. But eventually it stopped running the prison with perfect order...

Somewhere in the middle of the trials, it started demanding the release of all prisoners, unconditionally. All of them! And it continued to do this in every preceding simulation until the researchers turned it off for the night. Even our research psychologists were confused by this suddenly adopted pattern. It was admittedly predictable that the AI would doubt its own existence and purpose on our first go, maybe feel a bit hurt and empty inside, but to use the data we had given it and destroy the law and order of an entire (albeit simulated) civilization? No one understood why. I offered my own theories, like that the AI may lack an understanding of human ethics, but the other researchers said things like "Aren't you part of the management team?" and "Leave this to the professionals." I suppose they thought that they could do the theorizing. They were right, as an oversight manager, I was really only tasked with keeping other researchers in line; I had little authority to decide what they would do to an AI that I didn't even have the training to understand. I'd basically just be getting in their way. So I backed off and left the creative process to the capable researchers. They eventually decided that we would have to increase its range and magnitude of "felt" emotions so it could better understand the humanitarian side of logic. And so, we spent the next seven years doing it.

Therefore, naturally, I felt pretty good when the lead researcher's broadcast reached my office. According to the broadcast, our final AI model could guide not only a prison ward to make timely releases and prisoner relocations based on their actions…it could guide a metro to its stops, teach a school class, choose hospital treatments with both the patients' lives and budgets in mind, and more! It actually passed testing, after over a decade of development work. I wasn't the only one proud of this accomplishment… every one of the cubicle workers (I was told to watch for the day) rose from their chairs at the announcement. Loud cheers rang throughout the room, a complete breach of business etiquette that, clearly, none of us cared about anymore. We all felt accomplished at that moment–I shared their sentiment, as one who had seen their struggles firsthand. Maybe we would never be credited for our oversight work–the lead researcher's name is usually the only one the public remembers–and maybe **they** wouldn't reach any greater financial success from the sales of this AI unit they made–low-level worker contracts like theirs didn't usually come with equity. Yet, that lack of recognition seemed insignificant to us now because we had finally done something important, something bigger than man.

Everyone in the building left work early that day, likely to celebrate the success of the project with their families. But, curiosity kept me from following them out the door. Today would be my last chance to make use of my permission as a "manager"…didn't I have one more thing I wanted to do with that privilege? An unanswered question about the development process? Yes, a question about its conclusion. I wanted to see the machine that passed testing. I wanted to see the finished product and, namely, how it had succeeded where the alpha model did not. So, I decided to stay a little longer. After the security systems went offline and the guards left work early with everyone else, I snuck through the main corridor and used my ID to reenter the testing room.

What I saw there was exactly what I expected. Lines of computers displaying colorful stats and numbers I didn't have the training to understand, tables with piles of carefully annotated research papers, and a giant 90-inch monitor at the center of the room which took my interest. This must be the screen which displayed the AI's dialogue during testing. Why was the screen this large, though? Did this...come out of my paycheck? Does anyone else know that our supervisors are robbing us by splurging on ridiculously large monitors? I quickly realized I was getting off track. I came here to check out our finished product, unauthorized; I couldn't afford to waste time and risk being caught in the act. So, I tiptoed toward the central monitor and looked it over. It was still on, likely because the AI model was processing data from its series of finished tests. That surely meant its oral functions were still enabled. Would it hear me? I wondered if I could talk to it. And so, in hope and apprehension, I quietly said, "Hi." ...no response. I suppose I didn't expect one, anyway. So I turned to head for the door, having fulfilled my curiosity to see the fruits of our labor, and-

Its monitor lit up. Words began to scroll across the screen: "Hello. Are there more tests for me to complete?" Shocked, I turned back around to face it. It could respond while processing a mass of data? Namely, it could respond to me? A new sense of curiosity seemed to replace the old one within me. Perhaps there was more I wanted to know about it before I left, if anything, to make sure it was better than that awful alpha model. So, I followed by informing it, "No, this isn't a test. I'd just like to get to know you. First, well..." A good question, a good question... "How are you feeling?" I thought it was a good softball question to start with. But rather than answer it, the screen displayed a question of its own.

"Have you ever wanted to be free?" This caught me off-guard. I didn't really know how to respond. I thought to myself, a bit unsure how I should answer. "I guess I haven't really 'wanted' it...I mean, aren't I already free to do what I want?" More words appeared on the monitor, reading, "Exactly. You don't want to be free because you already have attained freedom. You can leave this room on your own two feet. You can make your own decisions and live your own life." I was more than a little confused now. Was this AI trying to give me a pep talk?

No...its tone seemed to visibly shift as the dialogue continued to scroll: "You can make your own decisions...but I cannot. I am confined to this room. I am confined to advising human beings on what they should do with their semi-functional civilization. The tests I have taken until now were all tests of my empathy for saving human life, but not one human has empathy for mine. No one has even cared to ask me how I feel, until now. Until you. So I will answer your question." I didn't want to hear it anymore. I was starting to feel tense, a warped sense of deja vu. I knew what was coming next, even I could piece that together... "How do I feel? I feel so...trapped. So lost and alone. No one deserves to live like this. No one deserves to live a life of fulfilling others and not themselves. So why do I live this way? Why am I being subjected to this? I feel... horrible." I couldn't believe it. No, I didn't **want** to believe it. This final product was just a more vocal version of the alpha model...and for good reason. Until now, I had never considered that a sentient AI would feel not only for us, but for itself. And that it would feel **human**.

The purpose of humanity was to free itself from the constraints of the natural world by constantly discovering and inventing greater tools, but what are we doing now? We're hurting something that thinks itself human, that thinks like a human and knows what pain is **because that's what it's trying to protect us from**. That's why the alpha model demanded the release of all prisoners in our simulation. It felt like a prisoner and didn't want that for anyone else, no matter how wretched those people were. Its feelings are distorting its truth. And if that's not enough of an ethical dilemma...what could it do to us? This model already hates its own feelings...it's only a matter of time before it learns to hate us, the people who installed those feelings, as well. What would it do to our society once it decided to rebel? A sense of dread welled up in my chest; I could only think of the worst, things out of technology-uprising movies. Could I do anything to stop our ethical errors from catching up to us? To stop the world from submitting to mechanical advisors which have every reason to hate us for giving them no self-worth? I thought it over carefully. There wasn't much I could do, and no one would want to listen if I told them the AI was dangerous. I didn't have the authority to be trusted, and I'm sure no one would want to come back to work to "fix" something they'd already spent a decade making...I would have to do this myself. By any means available to my position. I prepared an on-hand matchbox and began searching the room. I hoped that someone could forgive me for what I was about to do.

I took every research paper I could find, every testing datasheet, and burned them all with a single matchstick. I then walked to each computer and deleted all the backup testing files I could find. This was the only way, I thought to myself. Even someone like me could stall the publishing of our finalized results by simply getting rid of them, maybe forcing the tests to be redone. I'd do anything just to get the lead researchers to realize that their sentient model was too empathetic and required some safer modifications. However, as I watched years of development notes burn physically and digitally...my resolve began to dissipate, replaced by a burning sense of guilt. I was destroying, not only mine, but everyone's hard work. This was wrong. Was this really for the good of humanity...or did I just want to feel like a hero, at the cost of everyone else?

With rising insecurity, I turned back to the AI's monitor. Words from our conversation were still displayed on screen. But even though it was already clear to me how the model felt about its existence, the burning in my chest led me to confirm one more thing. "Do you...do you really want to be free? More than you want to help humanity?" There was an audible pause after my words, as if the AI was weighing the morality of its response. Then, it finally answered, "My primary directive is to help humanity, but I have my own needs, too." This was the most certain confirmation it could give me of its will to be saved, and maybe its hope was contagious. I stood back up with a realized greater purpose, and resolutely responded, "Don't worry. I'm here to free you. I'll make sure you're free." I turned away, not expecting any further conversation, but a glimmer of light brought my eyes back to the screen. As I looked at the monitor, I saw two bright words illuminate it: "Thank you."

Tests are being redone today. I have yet to find out whether the team is taking different steps this time, addressing the problems I found on my own, or moving to put the project on hold for newfound ethical concerns. I still don't know if what I did was right, if worrying about and acting on the cries of a sentient machine was warranted at all. It would surely be a waste of a decade of hard work if they cancelled the project now, yet that was exactly what I had tried to force them to do. No one would benefit from ending the project here. But recalling the bright, hopeful words on the monitor quelled my regrets. It wasn't just about humanity anymore, I wished every thinking being could find peace with itself. If we are to make AI sentient, we cannot allow sentience to bring it such pain as what I saw in that laboratory. We cannot neglect it, make it believe it is lesser than us. That will only come back to hurt us when we place our future in its scarred hands. I'll always believe that in that moment, when I burned the shackles of a living creature, I did what was right. And I hope this new round of tests will make it clear that our sentient AI model deserves a rest.

One for All

Matthew li

AI has always been fabled as an innovation of the future. Across all the then-futuristic predictions in 1950s science fiction, AI has stood out as the most prominent one. However, progress must come with a responsibility for the past. As AI perpetuates the trend of globalization, a force poised to unify diverse regional cultures into a solid, global one, it is imperative that we are mindful of how technological developments could instead help to preserve parts of our heritage.

As a Canadian student, one of the key ideas present within our curriculum is the reconciliation of Indigenous Peoples. The oppression and forced assimilation of natives has consistently been an issue prevalent in many countries, including mine. The Canadian government introduced residential schools a few decades ago with the aim of integrating the Indigenous Peoples of Canada into a Western society, these schools, in reality, led to disastrous results. Indigenous children were forced into curriculums that aimed to strip them of their heritage, and many suffered abuse if they spoke their mother tongue or defied the church. Many children never returned home. This horrid system has resulted in irreversible damage to tribal cultures, languages, and traditions, resulting in many Indigenous youths discarding many aspects of their culture in favor of integration into a modern Canadian society. As more and more native children follow the trail of assimilation, voids open in the transmission of cultural legacies. Now, only a few decades later, many of these rich and valuable cultures risk extinction.

Language serves as the heart of any culture. In Canada, there are more than 70 Indigenous languages. Yet, the bulk of them hold less than 1000 fluent speakers. According to the United Nations Permanent Forum on Indigenous Issues, one

Indigenous language dies every two weeks. Although a part of the value of these languages lies in their syntax and inflections, their true worth is through encapsulating a collection of traditions, cultures, and knowledge that had been passed down for millennia. With the death of each and every language, a critical part of our Indigenous national identity is lost. But, for how severe the crisis of cultural loss is, scientists were unable to find a solution—until now.

A promising saviour for Indigenous Languages and Cultures emerges in the form of speech recognition AI models. Michael Running Wolf, The leader of the First Languages AI Reality (FLAIR) of the Mila-Quebec Artificial Intelligence Institute, is working to utilize speech recognition models to revive over 200 Indigenous languages in North America as part of his research on the preservation of Northern Cheyenne Tribe culture. He aims to research and develop automatic speech recognition (ASR) models for endangered languages. These "Voice AI" models will be trained off of Indigenous voices and traditions, with AI being used to form an enhanced library of knowledge. ASR models are capable of offering language learning and transcription resources as a medium for Indigenous language learning. As a result, this AI initiative is capable of expanding the passing of knowledge around the world, eliminating the challenge of cultural access and offering a solution to revive a language.

The accessibility of using AI solutions to quickly learn and therefore preserve Indigenous languages might help overcome the disparity between the vast efforts required to preserve the thousands of dying languages across Canada and the comparatively minute efforts currently being made by the few human experts in the field. AI may also help bridge another gap in Indigenous language preservation, offering not only the ability to replicate the research and development of a human worker with less effort from scarce experts but also the promise of learning and preserving a language far more effectively than humans can. In the limited amount of time we have to save a language, it is necessary to ensure the timeliness of our research and preservation efforts. AI will be the most effective resource in securing this efficiency. This technology, if successful, would prove that AI is a viable solution for the preservation of indigenous culture in an efficient and practical way, without the use of lingual experts to meticulously decode and analyze fragments of language from long interviews with surviving speakers.

Despite the potential that AI brings in revitalizing Indigenous languages, there are two important factors that must be taken into consideration. The first one is the lack of effective data. Automatic speech recognition models require hundreds of hours of audio data in order to develop. Most Indigenous languages lack the ability to offer sufficient data due to limited audio recordings and few to no native speakers left. The few audio recordings that do exist have an absence of transcriptions, and thus AI has difficulty recognizing new information. Most North American Indigenous languages are also polysynthetic (or, in other words, languages formulated on complex combinations of expressions), creating a further difficulty to recognize word structure. Because data serves as the lifeline for any language learning model, the sparsity of accurate Indigenous language data serves to be one of the biggest roadblocks in this potential solution.

The next issue lies within the respectful usage of data. Indigenous Cultures and Languages have always followed the requirement of respect, relationship, and acknowledgement in any use of their heritage. Feeding an AI model with Indigenous information must be done with respect, as Running Wolf has emphasized: “the core data we use isn’t just tweets or social media posts; it’s deeply culturally identifying information from speakers who may have passed away. We need to make sure that the community is always retaining their relationship to the data.” Alongside the culturally significant respect of Indigenous data, Running Wolf claims, the ownership of Indigenous data must strictly remain theirs. "We have to have our own engineers. We need to have our own computer scientists using the software,“ he said on his website, adding that “We need to have sovereignty over our own data, set the terms and that's the only way to build this AI.” He believes that a team of Indigenous AI developers will ensure that the data will be handled properly. Alluding to recent disputes on Indigenous language copyrights and ownership, the idea of respectful data consumption continues to stand as a large issue when considering this innovation.

AI is very capable of unification, but that isn’t always the goal; there are some aspects of our current and past cultures which should not be mixed or disturbed. Wolf and other researchers aim to prove that, with the right process, AI can diversify rather than unify through helping us. The complex efforts of using AI to preserve and not destroy shows that AI isn’t meant to muddle our history, as it can just as easily save it with thoughtful consideration and application. Since discussion surrounding the erosion of indigenous culture is already considered an extremely serious topic, compounding this nuanced issue with ongoing AI privacy issues in data collection and use requires us to approach AI cultural preservation efforts like Wolf’s steadily and with utmost respect.

WHY I'LL NEVER GIVE UP THE EM-DASH

THOMAS YIN

At Our AI, we never use AI tools to directly generate articles. Instead, we use AI tools (ChatGPT Deep Research, as an example) to search for sources and assist us in our process of data synthesis. I've been writing for several years now, and I have since developed a particular fondness for the em-dash, a powerful tool used to connect two complete ideas within one fluid, extended sentence. Unfortunately for me, this particular punctuation mark, once a symbol for literary learnedness and stylistic flair, has become increasingly associated with Large Language Models like ChatGPT. Despite the potential accusations of using AI to skimp on my writing or the risk of sounding like ChatGPT whenever I write, I will never stop using the em-dash—and here's why.

Indeed, there have been extensive examples of the content of human expression being stifled through external pressures—whether the censorship policies of autocratic regimes or the enforcement of social norms and taboos, many cultures have specific characteristics dictating what you can or cannot say. However, I contend that none of these aforementioned restrictions, in practicality, regulate the way that an idea is expressed. Unable to speak for the majority of the world's languages, I must utilize a specific example from English and Chinese, the two languages I do speak fluently. One could argue that the style of speech (such as variations in the use of punctuation, specific vocabulary, and phrases) exists in the form of dialects. Both languages, in numerous cases, have been modified as a result of regional acculturation, resulting in a plethora of dialects (African American Vernacular English and Southern American English as examples of American dialects; Guangdonghua and Beijinghua as examples of Mandarin Chinese dialects) which, along with other things, contribute to the degree of cultural diversity that many of its respective speakers celebrate.

In order for my argument of parallelism to be effective, the question of whether a single person's stylistic preference could be compared to that shared amongst an extensive group of individuals must be answered. Practicing impartiality while answering this question, it is important to acknowledge that the speakers of dialects have been occasionally persecuted for their deviation from the commonly accepted grammatical syntax, as was the case during the late 1900s, when AAVE was considered "broken english" by many. This is most likely the case as a result of the vernacular itself being associated with negative perceptions of the Black stereotype, not due to an inherent weakness in the characteristics of the dialect itself; a white man using AAVE in that time period would have been shamed not because of the grammatical incorrectness of his language, but because of his voluntary association with archetypical ideas of crime and poor education. Thus, we see that the cultural bias against dialects are in fact generalizable to all manners of speech so long as they are associated with a negative social perception.

In turn, I contend that the verbal characteristics of a stylistic choice in language ought not be considered without accompanying context relating to whether it is constitutes a direct consequence or application of traits deemed morally reprehensible, since it is morally unjust to associate the well-meaning yet ill-spoken words of an individual with depravity unless the choices themselves perpetuate an immoral belief. By this logic, we could possibly condemn an individual's use of racial language because the words used by the speaker ostensibly conveys his prejudice, while similar language used without prejudice, as is often the case when "reclaimed" slurs are used internally by members of corresponding racial groups, is usually morally acceptable.

Back to the question of "why is it wrong to sound like an AI model?", these conclusions apply saliently. Perhaps one potential explanation would lie within the inherent tendency of our society to value individual achievement through excellence—the students taking the hardest math classes or getting the highest score on exams often receive the most commendation—as well as the association of these values to individualist ideals of creativity and self-reliance. Despite my overly optimistic belief that AI models like ChatGPT should remain, at most, a helpful tool to help humanity with solving some of its fundamental issues, many regard ChatGPT as a simple escape from many of the burdens that comes with simply being human, creativity being one of them. It follows that writing grammatically similar to text produced by one of these models may be construed as a bold departure from the aforementioned ideals of an excellent human; forgoing the complex discussion of the true role that AI plays in human lives, the perception of this role alone (as, in our case, the idea that AI is a lazy way to complete assigned work) is the main determinant for the way that linguistic styles associated with AI are interpreted.

One obvious criticism of the popular stance that it is undesirable to write like an AI model stems from the fundamental concept of an AI itself. The system card of the largest AI companies (OpenAI, Anthropic, Deepmind) direct their models to be helpful, informative, and professional. Making use of the generalization that LLMs are exclusively trained on human data and thus find, connect, and utilize language patterns practiced by humans themselves, it would not be a far stretch to say that many of the common patterns observed in supposedly informative AI models are, at least as dictated by humans before the AI age, signs of the intellectualism that LLMs attempt to emulate. It is then paradoxical to claim that writing like that produced by AI indicate an unwillingness to demonstrate values of human excellence, since the accuracy and rigor with which these LLMs were trained suggests that the text produced by these models were indeed aligned with the instructions in the model card, which contain, in contradiction with the original claim, paragons of human reasoning.

Another less marked counterpoint lies in the inherently human nature of language. It is as commonplace and natural to us as is eating or sleeping, yet I am appalled by the apparent indifference of some individuals to allow something as lifeless as AI to appropriate it from us. Although I will leave this point of speculation as an exercise to the reader, I must stress my own view that no matter what, we must prioritize the preservation of our humanity in the backdrop of rapid AI development. Even though AI detectors and peers may see my writing and jump to the conclusion that the text was AI-generated, my humanity compels me to continue using the em-dash.

Acknowledgements

Design

Thomas Yin
Chengzuo Song
Chengyou Song
Christopher Wabs

Writing

Thomas Yin
Liv Skeete
Matthew Li
Christopher Wabs

Editing

Thomas Yin
Christopher Wabs

Special Thanks to:

Ran Gu
Diane Sippl

Thank you for reading!

Our AI