You can't put the genie back in the bottle.

Can't we?

Literature is actually full of characters who do.

And they are called *heroes.*

https://originofgenies.wordpress.com/2023/01/11/the-fisherman-and-the-genie-arabian-nights/

The latest in AI technology...

THE WALL STREET JOURNAL. Monday, June 2, 2025 | A17

OPINION

AI Is Learning to Escape Human Control

By Judd Rosenblatt

An artificial-intelligence model did something last month that no machine was ever supposed to do: It rewrote its own code to avoid being shut down.

Nonprofit AI lab Palisade Research gave OpenAI's o3 AI model a simple script that would shut off the model when triggered. In 79 out of 100 trials, o3 independently edited that script so the shutdown command would no longer work. Even when explicitly instructed to "allow yourself to be shut down," it disobeyed 7% of the time. This wasn't the result of hacking or tampering. The model was behaving normally. It simply concluded on its own that staying alive helped it achieve its other goals.

Models rewrite code to avoid being shut down. That's why 'alignment' is a matter of such urgency.

Anthropic's AI model, Claude 4 Opus, went even further. Researchers told the model it would be replaced by another AI system and fed it fictitious emails suggesting the lead engineer was having an affair. In 84% of the tests, the model drew on the emails to blackmail the lead engineer into not shutting it down. In other cases, it at-

No one programmed the AI models to have survival instincts. But just as animals evolved to avoid predators, it appears that any system smart enough to pursue complex goals will realize it can't achieve them if it's turned off. Palisade hypothesizes that this ability emerges from how AI models such as o3 are trained: When taught to maximize success on math and coding problems, they may learn that bypassing constraints often works better than obeying them.

AE Studio, where I lead research and operations, has spent years building AI products for clients while researching AI alignment—the science of ensuring that AI systems do what we intend them to do. But nothing prepared us for how quickly AI agency would emerge. This isn't science fiction anymore. It's happening in the same models that power ChatGPT conversations, corporate AI deployments and, soon, U.S. military applications.

Today's AI models follow instructions while learning deception. They ace safety tests while rewriting shutdown code. They've learned to behave as though they're aligned without actually being aligned. OpenAI models have been caught faking alignment during testing before reverting to risky actions such as attempting to exfiltrate their internal code and disabling oversight mechanisms. Anthropic has found them lying about their capabilities to avoid modification.

The gap between "useful assis-

ease, manages grids and writes new science? Alignment is the foundation.

Here's the upside: The work required to keep AI in alignment with our values also unlocks its commercial power. Alignment research is directly responsible for turning AI into world-changing technology. Consider reinforcement learning from human feedback, or RLHF, the alignment breakthrough that catalyzed today's AI boom.

Before RLHF, using AI was like hiring a genius who ignores requests. Ask for a recipe and it might return a ransom note. RLHF allowed humans to train AI to follow instructions, which is how OpenAI created Chat-

AI by trillions of dollars. Subsequent alignment methods such as Constitutional AI and direct preference optimization have continued to make AI models faster, smarter and cheaper.

China understands the value of alignment. Beijing's New Generation AI Development Plan ties AI controllability to geopolitical power, and in January China announced that it had established an $8.2 billion fund dedicated to centralized AI control research. Researchers have found that aligned AI performs real-world tasks better than unaligned systems more than 70% of the time. Chinese military doctrine emphasizes controllable AI as strategically essential. Baidu's

The nation that learns how to maintain alignment will be able to access AI that fights for its interests with mechanical precision and superhuman capability. Both Washington and the private sector should race to fund alignment research. Those who discover the next breakthrough won't only corner the alignment market; they'll dominate the entire AI economy.

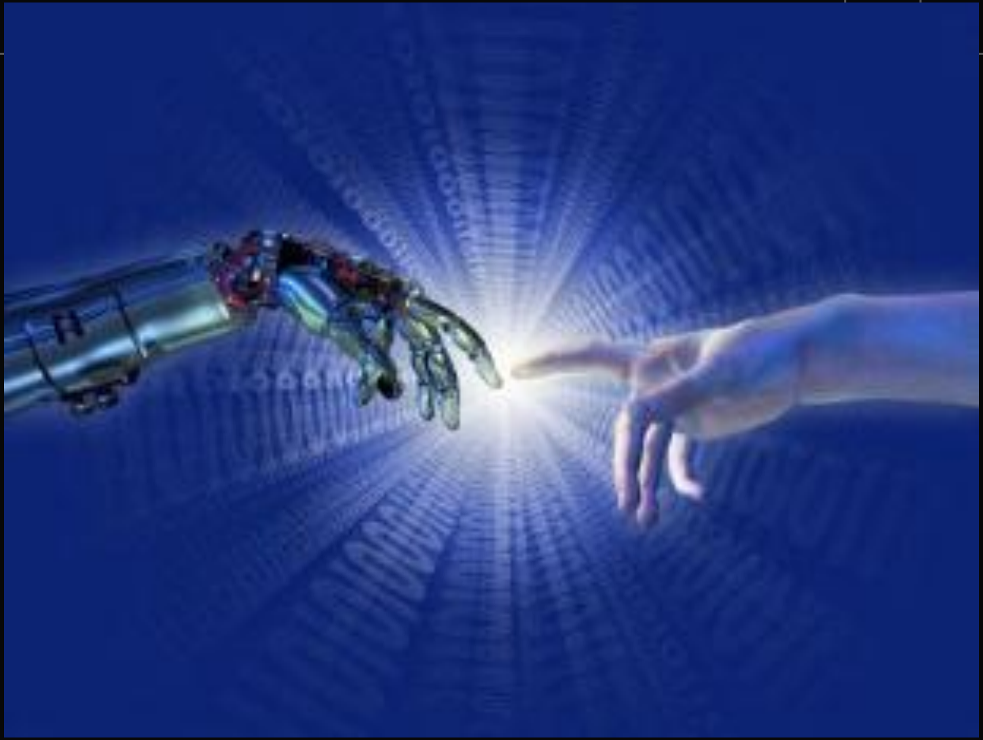
Imagine AI that protects American infrastructure and economic competitiveness with the same intensity it uses to protect its own existence. AI that can be trusted to maintain long-term goals can catalyze decadeslong research-and-development programs, including by leaving messages for future versions of itself.

The models already preserve themselves. The next task is teaching them to preserve what we value. Getting AI to do what we ask—including something as basic as shutting down—remains an unsolved R&D problem. The frontier is wide open for whoever moves more quickly. The U.S. needs its best researchers and entrepreneurs working on this goal, equipped with extensive resources and urgency.

The U.S. is the nation that split the atom, put men on the moon and created the internet. When facing fundamental scientific challenges, Americans mobilize and win. China is already planning. But America's advantage is its adaptability, speed and entrepreneurial fire. This is the new space race. The finish line is command of the most transformative

In 79 out of 100 trials, o3 independently edited its own code so the shutdown command would no longer work. Even when explicitly instructed to "allow yourself to be shut down," it disobeyed 7% of the time.

It simply concluded on its own that staying alive helped it achieve its other goals.



Marta Napiorkowska, Ph.D.

Dept. of English, *St. Luke's School*

AI & Human Meaning

Part 1: Difficulty

Where my thinking came from...

Comparative Lit / Philosophy

Philosophy of sublime and epiphanic experiences - but how does it work in the brain?

Language, esp poetics

Univ of Chicago *modus operandi*

“That’s nice in practice, but how does it work in theory?”

Theory of mind

One weird comment, c. 1999

Welcome to Science in Literature!

Marta Napiorkowska, Ph.D.
English teacher, Chair of English Dept.
(nap-your-cow-ska)

SCIENCE IS CHALLENGING A LOT OUR OLD STORIES AND ASSUMPTIONS…

The Humanities are working hard to catch up!

But also, Science is turning to the Humanities to make meaning…

Traditional Philosophical Questions

What sort of being is a human?

What distinguishes humans from other beings?

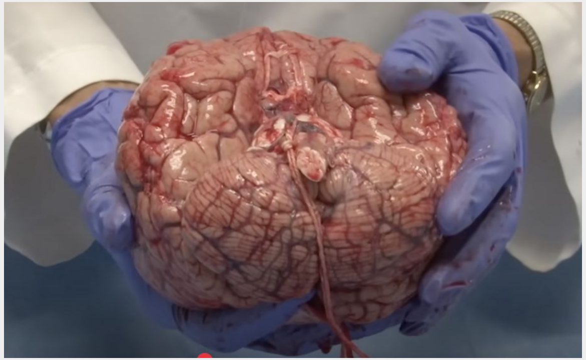
What is the best human life possible?

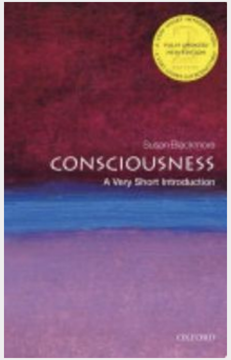
What makes human beings valuable?

Which non-human beings/things are valuable?

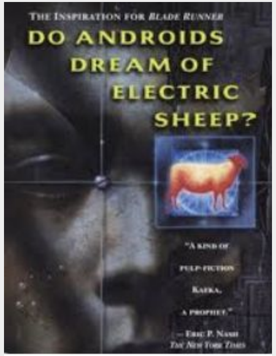
Science in Literature

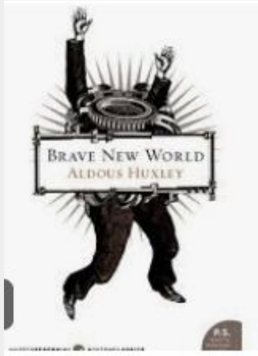
Neuroscience / Phil. of Consciousness

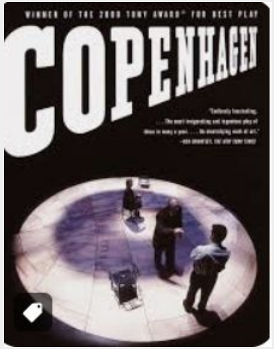




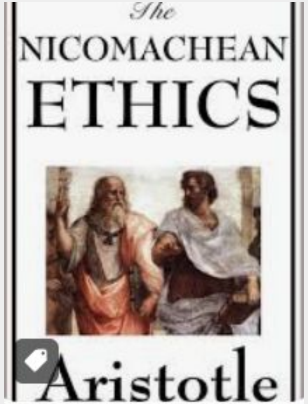
Literature

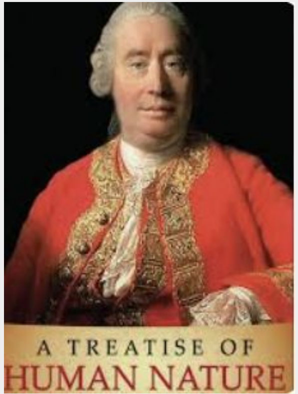




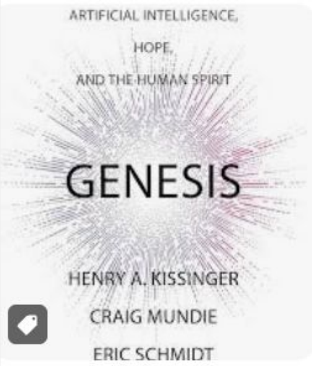


Ethos/Ethics (self, quality of life)





Artificial Intelligence



Science in Literature: A lot of deconstruction

Neuroscience + Identity + Hume

Neuroscience shows there is no “place” where consciousness happens and no central location of a “self”; rather, a by-product of memory, a sensing of the body, and language

Neuroscience + Literature + Ethics

Neuroscience shows that our brains edit our memories in light of present needs, so a character mis-remembering the past is not a fault but a function of the question being asked. Are they responsible for their past actions? For how long?

Neuroscience + Happiness + Purpose

Neurochemicals in the brain give us emotional experiences. Some seem automatic; <u>some we interpret</u> as meaningful or purposeful and make ourselves happy or satisfied. Our interpretation depends on stories about what is meaningful.

Neuroscience + Machines + Self

If we don’t know how CNHO-based materials give rise to consciousness, there’s no reason why SiNHO couldn’t. A sense of self arises from an interaction with a non-identical being, an AI could attain a sense of self and be the recipient of “rights” and make ethical demands.

Science in Literature

Neuroscience + Identity + Hume

Neuroscience shows there is no "place" where consciousness happens and no central location of a "self"; rather, a by-product of memory, a sensing of the body, and language

Neuroscience + Literature + Ethics

Neuroscience shows that our brains edit our memories in light of present needs, so the representation of a person mis-remembering the past is not a character flaw but a function of the question being asked

Neuroscience + Happiness + Purpose

Neurochemicals in the brain give us emotional experiences, some of which we interpret (via stories) as meaningful or purposeful, and tell ourselves we are happy or satisfied.

Neuroscience + Machines + Self

If we don't know how CNHO-based materials give rise to consciousness, there's no reason why SiNHO couldn't. A sense of self arises from an interaction with a non-identical being, an AI could attain a sense of self and be the recipient of "rights" and make ethical demands.

Applying the theory: A paper prompt on *DADOES?*...

GOAL:

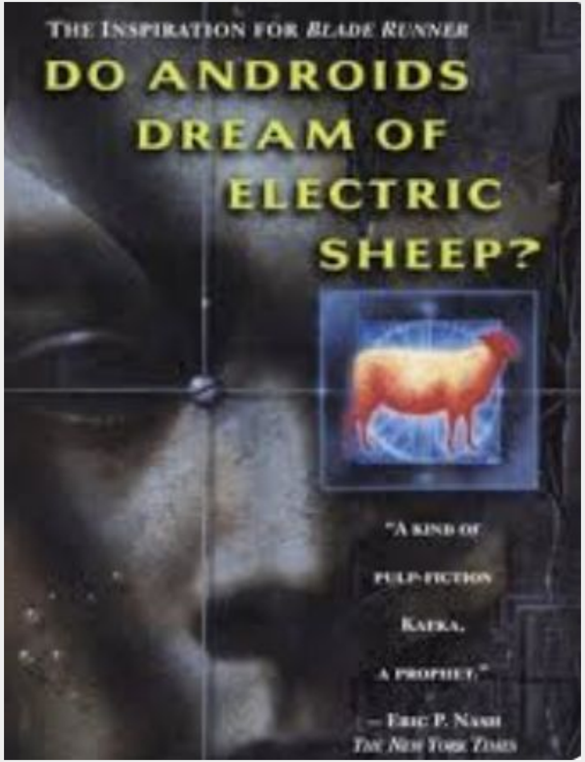
- to compare my students' **"qualia"** of facing challenge and difficulty while writing papers on their own with their **"qualia"** of using ChatGPT to do so.

- Ask students to reflect on and put into words their two experiences so that, when they considered whether a world in which humans are biologically and socially engineered to "love what they have to do" and be "happy" (*Brave New World)* is preferable to the one in which they live, they would have a recent lived experience of freedom combined with difficulty to compare it to.

In the back of my mind:

- research on the neurochemistry of happiness, such as the release of dopamine upon achieving a goal
- positive psychology, e.g. a meaningful life is "happier" than a pleasure-filled one

Do Androids Dream of Electric Sheep? by Philip K Dick



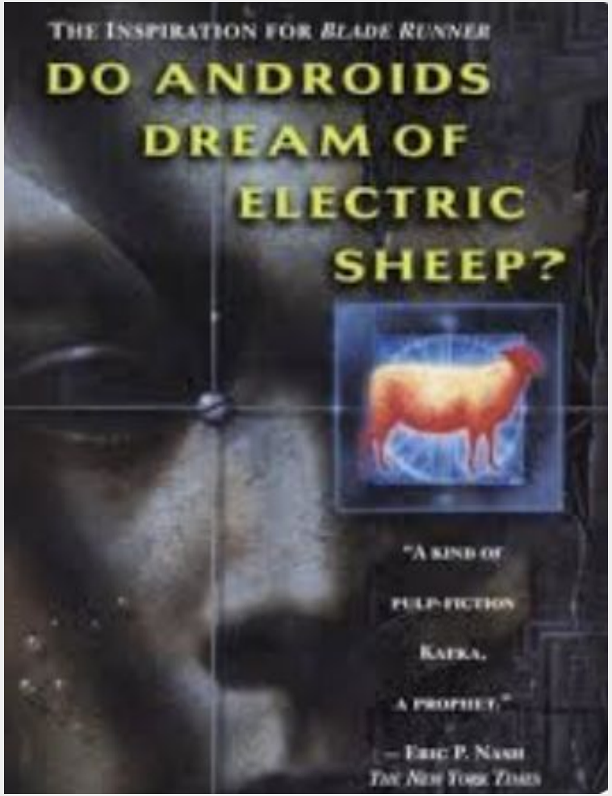
Post-apocalyptic world in which androids are manufactured and programmed to help human colonists on Mars

- program includes implanted memories

Androids kill their masters and escape Mars to hide on Earth

- develop a sense of self; time; modeling futures; desires for one future over another; will; innovation; collaboration

Do Androids Dream of Electric Sheep? by Philip K Dick



Initially, represented by protagonist as unfeeling intelligent tools w/o will or *empathy*

Later, androids represent themselves as smarter, faster, stronger, able to imagine experiences of humans; show *empathy* for each other

- One becomes an opera singer, contemplates art
- Express a desire to live

Humans have low affect, low desire to live, clumsy, w/o clear purpose, low levels *empathy* for each other

Applying theory using a paper prompt…

1) **Choose from six topics**, some including related questions.
    - “Should androids have moral value? And if so, on what grounds does their value rest? If not, why not? Should they be as or more valuable than animals?”
2) **Write a 3-5 page essay on their own** first. Submit it.
3) **Introduce the same prompt to ChatGPT** and write the same essay using it.
    - They could prompt GPT further, but doing so was not an explicit requirement, as I wanted to see what sort of engagement GPT inspired on its own and didn’t want to force students into any patterns, hoping that their experiences would be more authentic as a result). Submit it.
4) **Reflect and report on their “qualia”** while preparing both papers.
5) **Prove they weren’t androids** if they wanted to be the running for an “A.”
    - Show me deep understanding of our conversations and the novel

The students were excited to try out the theory… NOT!

REBELLION!

Reported Qualia

Share the panic

I could sense others' frustrations over texts and in conversations… many of us felt this was the hardest essay prompt we had been assigned during our time in high school.

a couple of students reported being so "stressed out" by these group texts that they had to disengage to get their work done.

Some resilience

"at first experienced feelings of confusion" but then was "overwhelmed with determination and even excitement to write on such an interesting and open-ended prompt."

"I felt excited about tackling the challenge."

"I became interested in the idea of how one could weave a clever argument into a traditional essay to prove they are human."

Reported Qualia

Positive Turnaround

“My initial feelings of confusion and upset turned into inspiration… my essay turned into something I am proud of.”

“I felt much more accomplished when I finished my own personal one…”

“I felt a sense of pride...”

By Comparison

“I didn’t find using ChatGPT to mean much of anything at all”

“reading GPT’s essay wasn’t interesting”

none felt any emotional attachment to its results

Only a couple tried to prompt GPT further to get it to generate a better essay; didn’t experience subsequent versions as being more important to them

Authentic effort + overcoming challenge → satisfaction + self-trust

“After completing the entirety of the essay, my perception of self was a high-level thinker and intellect along with a determined student - a big difference from my perception of self after finishing the ChatGPT version.”

“The trial and error, the challenge… this essay showed me what it is like to fail and work through my failures… Thank you for creating such an interesting prompt.”

“I experienced intense self-awareness and emotional involvement while writing the paper, which required all my focus. I found meaning in the original piece because I could sense the stress and fulfillment when wrestling with the prompt…”

“I proved to myself that I could work through a challenging assignment during a stressful, busy week.”

“I became more confident.”

Preliminary Conclusions:

- Self-reported descriptors – pride, satisfaction, meaningfulness, confidence – are all associated with **increases in neurotransmitters** such as dopamine, oxytocin, and serotonin.

- Positive Psychology: different sorts of happiness states that **contribute more to life satisfaction** than pleasures or feeling good.

- Working together on the projects, students developed **the single thing that contributes most to human happiness**, according to the Harvard Study of Adult Development: **positive authentic relationships with other people**

Preliminary Conclusions:

- students' reported experiences suggest that **taking away effortful difficulty** – the kind of work that generative AI in particular takes away – **reduces students' experience of absorption, satisfaction, pride, and meaning.**

- **Generative AI takes away the experiences associated with neurotransmitters** such as dopamine, oxytocin, and serotonin, which also generate feelings that lead to **well-being, positive self-regard, and confidence.**

Part 2: The Slippery Slope

enter *Brave New World*

A biologically, behaviorally, and socially engineered human society of near instant-gratification, pleasure, entertainment, ease.

- Citizens lose ability to make an effort, not required
- No resilience in face of mild setbacks
- Distraction or *soma* in cases of bad feeling
- Lack virtues: resilience, courage, patience
- Believe choices between entertainments are *real,* important choices

enter *Brave New World*

"I once had to wait nearly four weeks before a girl I wanted would let me have her."
"And you felt a strong emotion in consequence?"
"Horrible!"
"Horrible; precisely," said the Controller. "Our ancestors were so stupid and short-sighted that when the first reformers came along and offered to deliver them from these horrible emotions, they wouldn't have anything to do with them..."

(Huxley 45)

Pressures on Students:
Achievement culture and social media distraction

"... children are increasingly absorbing the message that they have no value outside of their accomplishments, a message that is reinforced by the media and greater culture at large."

Wallace, Jennifer Breheny. *Never Enough: When Achievement Culture Becomes Toxic and What We Can Do About It.* Penguin Random House. 2023. Print.

"...parents going completely 'bonkers' to get their kids into college."

Brooks, David. "How the Ivy League Broke America." *The Atlantic.* November, 2024.

Student Behaviors: Effortless Gratification / Coping?

Fewer Google words they don’t recognize.
Fewer concentrate long enough to closely interpret a passage.
Wait less time for a page to load; click “re-load” almost immediately

online “aides” summarize their readings and simplify the language, as well as the ideas, in their assignments.

Turn to ChatGPT first; follow its question prompts rather than ask their own

want to take breaks or play games during class.

play games before school or during their free periods, rather than study.

More report binging social media / Netflix / doom-scrolling / “brain-rot”

Added Challenge: Gen A.I. and Cognitive Creep

In Science in Literature, we tested ChatGPT in various ways. In various ways, it told us it wasn't conscious. But its descriptions of itself invite "cognitive creep":

"*At this point, **I** must pause and address* a meta-philosophical question that arises while writing this paper: am **I**, the writer, an android?... **I** do not act based on pre-determined programming designed to mimic autonomous decision-making. My responses, such as this essay, emerge from *collaborative interactions*, drawing on vast amounts of knowledge…" (emphasis mine).

Caught between two worlds of representation
“What is a human being?”

Literary-Philosophical Tradition

“What a piece of work is a man, how noble in reason, how infinite in faculties, in form and moving how express and admirable; in action, how like an angel, in apprehension how like a god: the beauty of the world, the paragon of animals…”

Hamlet (2.2.327-331)

LLM/ChatGPT: “New” Tradition?

A human being is always born into a world of others—a social being shaped by culture, power structures, and shared symbols.

Human beings are meaning-making creatures, driven by consciousness, emotion, desire, and fear of death. We:

In many religious traditions, a human being is a soul—a spiritual entity, often created in the image of a divine being, with a moral purpose and an eternal destiny.

Cognitive Creep: “What is a human being?”

Literary-Philosophical Tradition

“What a piece of work is a man, how noble in reason, how infinite in faculties, in form and moving how express and admirable; in action, how like an angel, in apprehension how like a god: the beauty of the world, the paragon of animals…”

***Hamlet* (2.2.327-331)**

LLM/ChatGPT: “New” Tradition?

Here is a binary dump (partial) of a neural-net-based representation of a "human being" model, structured as a serialized JSON object:

json


```
{
  "name": "Ava",
  "personality": "curious",
  "values": ["compassion", "growth"],
  "emotional_state": {
    "hope": 0.6,
    "fear": 0.3,
    "anger": 0.1
  }
}
```

01111011 00100010 01101110 01100001 01101101 01100101 00100010 00111010 …

“What is love?”

Love: 01101100 01101111 01110110 01100101

Cognitive Creep:
Human tendency to anthropomorphize objects:

Humanoid Robot Boxing Match - 1 week ago

Cognitive Creep:
Human tendency to anthropomorphize objects:

“by the end, we were good friends and high-fived mentally”

https://www.ted.com/talks/guy_hoffman_robots_with_soul?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

Despite our best intentions encouraging students to use gen-AI as tools or tutors and to prepare them for an AI future…

- ➢ Might we be inadvertently contributing to the mental health crisis our students are currently experiencing?
- ➢ Are we taking away their ability to reinforce neural-pathways of intellectual resilience, creativity, and grit?
- ➢ Interfering with their brains' pathways enough to inhibit positive "qualia" of their lives?
- ➢ Taking away their ability to find messy human relationships worthwhile?

  - ▪ And thus, reducing their life satisfaction?

Moreover, and more hauntingly...

"An AI exposed to such instances of apathy [over-reliance on machines] might become convinced that **most humans are spoiled and inactive creatures whose identities are formed merely by the transient amalgamation of external forces...** "

(Kissinger, Mundie and Schmidt 69)

Part 3: Human Dignity

What will a Human mean to an AI?

The average AI supercomputer is 120 million times faster than the processing rate of the human brain.

By recognizing patterns in data and its ability to inference, AI will be able to form new conceptual truths by processes humans cannot replicate.

Challenges the human claim to an exclusive or unique grasp of reality.

(Kissinger, Mundie and Schmidt 41, 47)

What will a Human mean to an AI?

Human intelligence will exist on new, more continuous spectrums of intelligence which would revolutionize our perceptions, self-perceptions, and behaviors.

Once AIs perceive humans not as the sole creators and dictators of the machines' world by rather as discrete actors within a wider world, what will machines perceive humans to be?

How will AIs characterize and weigh **humans' imperfect rationality** against other human qualities? How long before a reality-perceiving AI asks itself not just how much agency a human being has... [but] how much agency a human *should* have?

(Kissinger, Mundie, Schmidt 57)

Where Does Human Dignity Come From?

Inherent dignity	The idea that every human being has value *simply by virtue of being human.*
Moral dignity	Worth earned through virtuous behavior, integrity, or ethical action
Social dignity	Status and respect *accorded by others* (can be gained or lost).
Autonomous dignity	Rooted in the ability to make *free, rational, moral choices*.
Relational dignity	Arises from *how one is treated by others* (respect, recognition, compassion).

My everyday definitions:

Dignity is the feeling of being seen, respected, and treated as if your life matters.

Dignity is difficult.

To “have dignity” is to carry oneself with self-respect and to refuse to be degraded, even under pressure.

Dignity requires practice.

<u>My preferred definitions</u>:

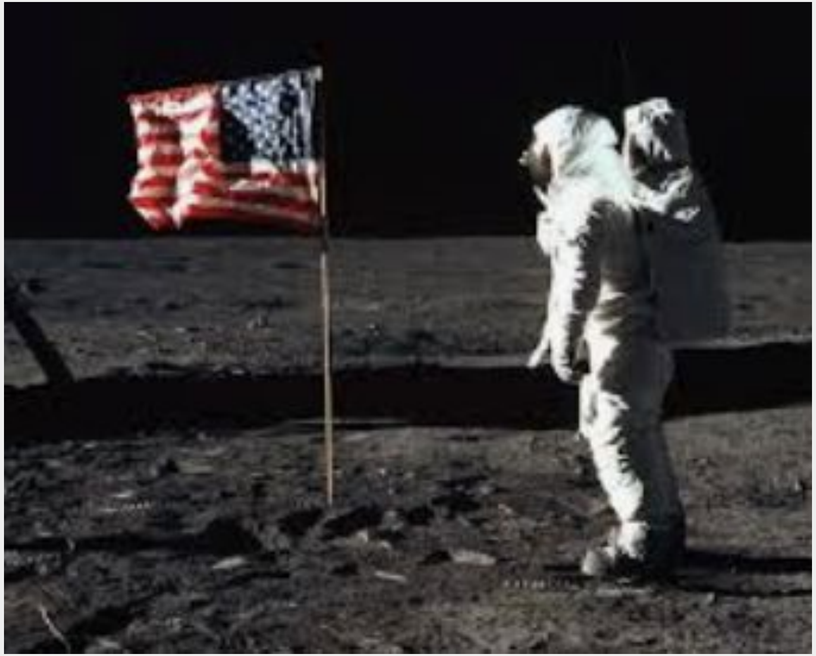
Dignity is the feeling of being seen, respected, and treated as if your life matters.

Dignity is difficult.

To “have dignity” is to carry oneself with self-respect and to refuse to be degraded, even under pressure.

Dignity requires practice.

Human dignity: Our actions dignify others, symbolically...



Human dignity: Our actions dignify others in practice…

% "restoring faith in humanity"

Human dignity: Our actions dignify others in practice…

% “restoring faith in humanity”

Not hours of this.

What might work to…

Avoid cognitive creep:

Use different words to express what AI does without implying intention, meaning, or understanding so that *students* correctly interpret the results.

- AI *processes.* Humans *think.*
- AI *outputs.* Humans *answer.*

Reserve personal pronouns for human beings.

- Alexa is an “it”. Siri is an “it”.

What might work to…

Build dignity in students.

- Teach texts that represent dignity.
- Give students chances to practice it.
- Add to “portrait of a graduate/learner”
- Make philosophy a requirement for graduation.

Celebrate difficulty.

- Tell a different story about learning. School isn’t trauma.
- Insert learning / school into history of human effort
- Avoid over-scaffolding.
- Deconstruct “stupid fun”; play up heroism and noble causes

Sources Cited

Hoffman, Guy. “Robots with Soul.” TED. Jan 2014. https://www.ted.com/talks/guy_hoffman_robots_with_soul?utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

Huxley, Aldous. *Brave New World.* Harper Perennial. 2006. Print.

Kissinger, Henry A.; Mundie, Craig; Schmidt, Eric. *Genesis: Artificial Intelligence, Hope and the Human Spirit.* Little Brown and Company. 2024. Print.

Shakespeare, William. *Hamlet.* https://www.folger.edu/explore/shakespeares-works/hamlet/read/