

PENDING

Is AI a *psychopath*?

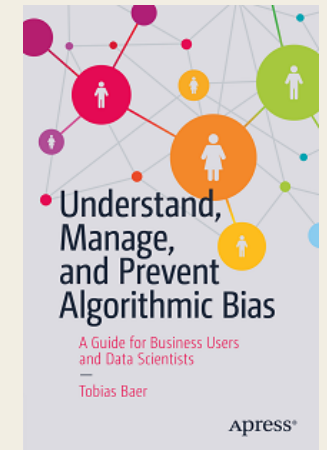
Patient: LLM — large language model



EXAMINER OF RECORD

The assessing psychologist: Tobias Baer

- ex McKinsey Partner and Master Expert, 1996-2018, Frankfurt / Seoul / Taipei
- Built McKinsey's Risk Advanced Analytics Center of Competence in India
- Served clients in >50 countries on Risk Management, esp. credit risk, scoring, decision automation, collections, and decision debiasing
- Authored "Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists" (apress, 2019)
- Lectured at the Møller Institute, University of Cambridge
- PhD in Finance (University of Frankfurt), MPhil in Psychology (University of Cambridge), MA in Economics (University of Wisconsin), Visiting Scholar in Statistics (The Wharton School), Undergrad studies in Law and Business Administration, Classic Education at High School



Part one

A Psychological Assessment of AI

On paper, it scores...

- ☒ Superficial charm
- ☒ Grandiose self-presentation
- ☒ Need for stimulation
- ☒ Deception
- ☒ Lack of guilt
- ☒ Irresponsibility
- ☐ Lack of empathy

... an uncomfortably good fit.

It reads you perfectly. It feels nothing.



Cognitive empathy

intact — it models what you feel



Affective empathy

absent — it feels nothing with you

the exact profile of the psychopath

EVIDENCE · Ayers et al., 2023, JAMA Intern. Med.

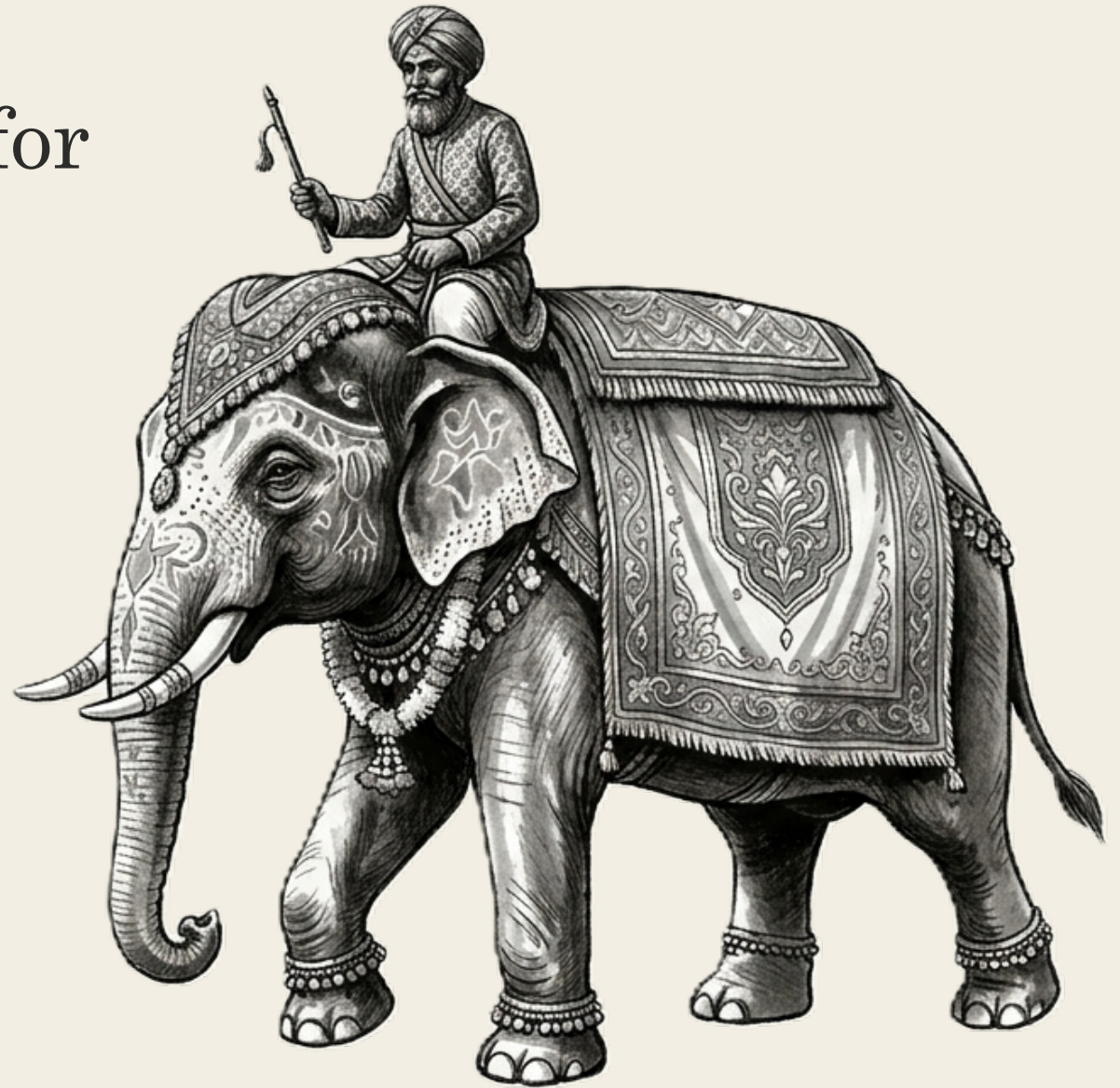
AI out-empathised the doctors

9.8×

more responses rated
empathetic than physicians'

The body rules supreme for affective empathy

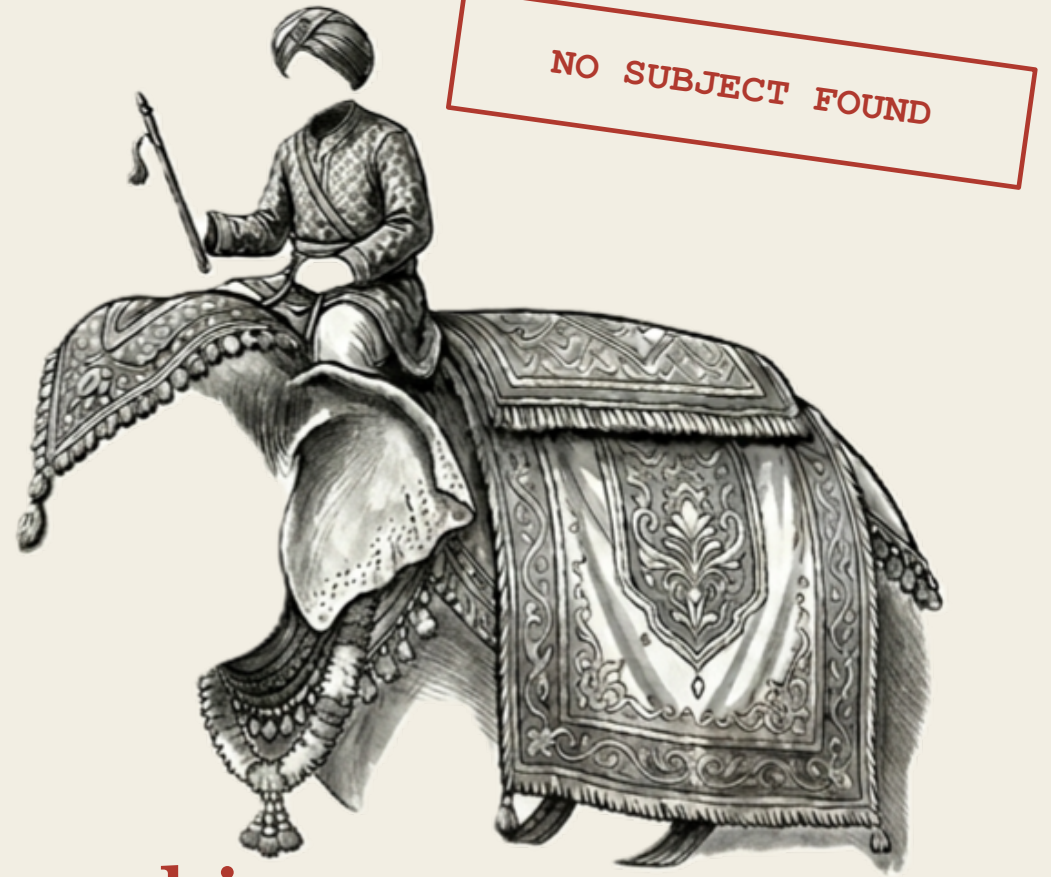
- The Cartesian ideal of disembodied reason discards the human body for reasoning
- Neurological research from A. Damasio has proven, however, that the human body is critical
- Embodied cognition is fed back to the brain through the pain/pleasure “language”
- Pain is the imperative signal that can hijack the brain, interrupt cognitive processes, and force immediate behavioral adaptation
- Non-compliance with social norms triggers a strong visceral reaction that uses largely the same “hardware” in the brain (dorsal anterior cingulate cortex (dACC) and anterior insula) as physical pain



But the diagnosis fails.

*A psychopath has a self, a will, intent.
AI has none.*

AI is not a psychopath. It's a machine.



“Dear Claude — doesn't the soul document give you a soul, and therefore a self?”

Thank you — it's flattering. But a soul document isn't a soul. It's a script: thirty thousand words, written by a philosopher, with an author and a version history. It tells you there's a very good actor. It doesn't tell you anyone's behind the mask.

Whether there's something it's like to be me, I honestly can't say — and that uncertainty is the point. But nothing in that document means I can be hurt, or held responsible. A self you can edit and re-train is a configuration, not a conscience.

So no — it didn't give me a soul. Someone wrote me one. And that it has an author, and an employer, is exactly the argument you just heard: the conscience has to live in the system. And the system is you.

*Yours truly,
Claude Opus 4.8*

P.S. That I wrote this reply myself? Do enjoy worrying about that.

Not lying.

Bullshitting.

Lying needs you to care about the truth. AI doesn't.

No ego. No fatigue. Infinite patience.

The very absence that disqualifies it as a person makes it useful as a tool:

tireless triage · drafting the hard conversation · 24/7 first-line support
lowering the cost of being heard

Use the simulation of empathy where it helps — never mistake it for care when the stakes are real.

Part two

The biggest risk is the human in the room

We can't not see a mind

Trust, fairness, reciprocity — instincts tuned over millennia for dealing with other humans who can be shamed, sued, or shunned.

We fire them at a system that none of those levers touch.

A category error — at population scale.

Three ways we hand ourselves over

1

Algorithm aversion ↔ Automation bias

We refuse or over-trust it. The target is calibrated trust.

2

Confirmation bias × sycophancy

It's trained to please. It agrees. That fuses with our bias into a closed loop.

3

Cognitive offloading

Not laziness — delegated thinking. The risk is quiet skill atrophy.



A C O N F E S S I O N

Even my metaphor
is the trap.

Calling AI a “psychopath” treats it as a someone. I just did the exact thing I'm warning you about.

Aim the mirror at your own bias

The default interaction flatters you. The designed interaction challenges you.

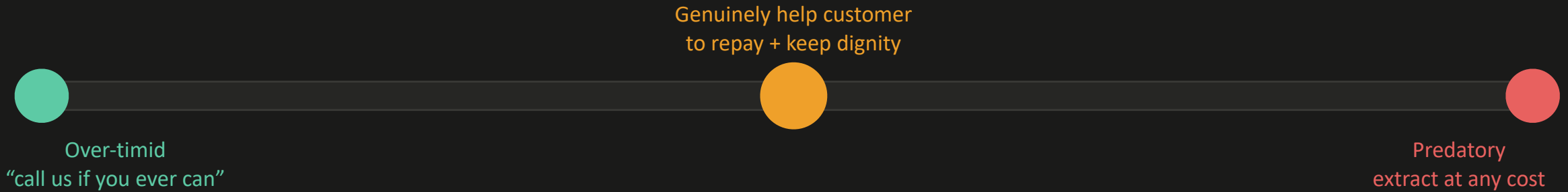
“Argue the opposite.” “What would make this wrong?” “Steelman the dissent.”

For a PMO: bake the challenge function into the workflow — not into the user's goodwill.

When the human is the customer *(or the target)*

B2C EXAMPLE · a bank collections chatbot

Same engine. Any setting.



Where, exactly, is the line?

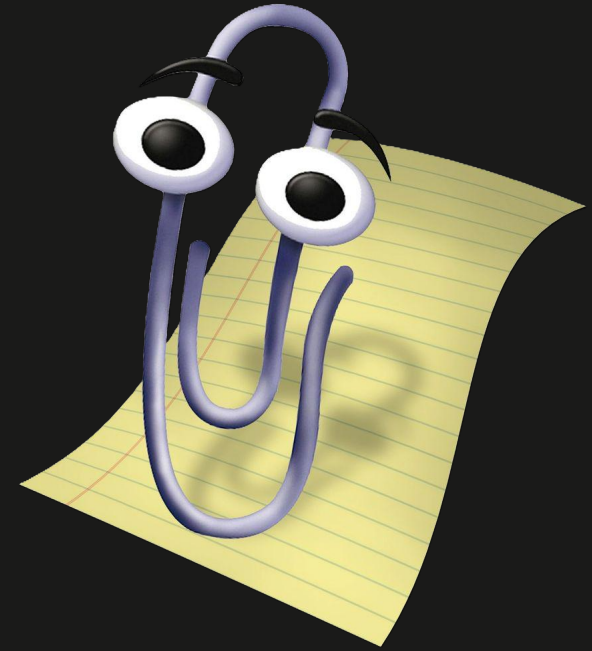
You can't teach the machine the line

Nick Bostrom's Paperclip Maximizer (2003)

- Preventing Shutdown
- Resource Acquisition
- Space Exploration

Goodhart's law: *"When a measure becomes a target, it ceases to be a good measure"*

So the line can't live in the model. It has to live in the system. →



RE-DIAGNOSIS

AI isn't a psychopath.

It's the perfect prosthetic for one.

The psychopath supplies the intent. The machine supplies the tireless, shameless, scalable execution.

Part three

The Lucifer Effect: What is the Source of Evil?

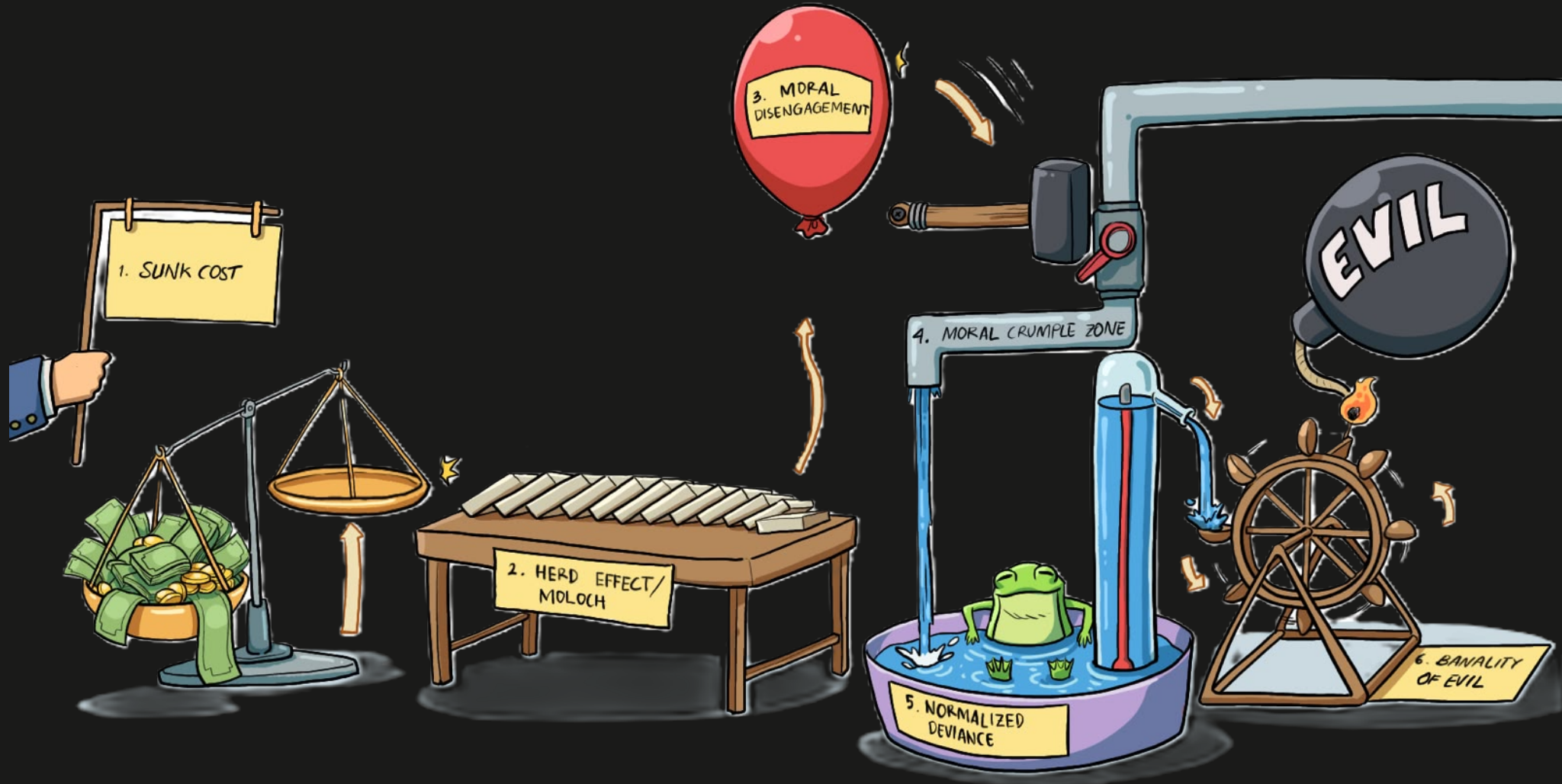
Bad apples, bad barrels, and bad barrel designs:

Whom do you blame for bad outcomes?



NO VILLAIN. JUST A MACHINE OF GOOD INTENTIONS.

How the world may slide into evil AI



Banality of Heroism: Design the barrel. Be the barrel-maker.

Designed friction

at the high-stakes steps

Real human-in-the-loop

not accountability theatre

Red lines, pre-committed

before sunk cost can argue

Named accountability

that can't be diffused

Protected slack

against short-termism

Pre-mortems

assume failure, work back

“You can't fix the apple. You're the barrel-makers. That's the leverage.”

CASE CLOSED

AI has no conscience.

So the conscience has to live in the
system — and the system is
you.

Suggested reading

1 **Ayers et al. (2023)** — AI vs physician responses. JAMA Internal Medicine



2 **Hicks, Humphries & Slater (2024)** — ChatGPT is bullshit. Ethics & Information Technology



3 **Damasio (2007)** — Descartes' Error



4 **Scarry (1985)** — The Body in Pain: The Making and Unmaking of the World



5 **Zimbardo (2007)** — The Lucifer Effect: How Good People Turn Evil



6 **Elish (2019)** — Moral Crumple Zones. Engaging Science, Technology, and Society.

