주체강

Navigate the Current

Lemon Juice on Your Face

Confirmation Bias, Sycophantic AI,
and the Cameras That Are Always Rolling

Vilhelm Pedersen, illustrating Hans Christian Andersen's "The Emperor's New Clothes" (1837)

Jesse James
linkedin.com/in/jessecares
March 2026

I The Bank Robber

In 1995 a man robbed two Pittsburgh banks with lemon juice on his face.

His reasoning was airtight. Lemon juice makes invisible ink. Therefore lemon juice makes faces invisible to cameras. He walked into both banks in broad daylight, no mask, smiled at the security cameras, and collected his cash. He was genuinely shocked when police arrested him that evening after airing the footage on the 11 o'clock news.

His name was McArthur Wheeler and two Cornell psychologists later built an entire theory of human cognition around him.

I think about McArthur Wheeler constantly. Not because he was stupid — he wasn't, not really. He had a theory. He tested it. He even rubbed lemon juice on his face and took a Polaroid beforehand, and the photo came out blurry, which he took as confirmation. He found exactly the evidence he was looking for and stopped looking.

We all do this. Every single day. We just do it about things more respectable than lemon juice.

II The Water We Swim In

Your phone's GPS tells you to turn left. You know the shortcut. You've driven this route a hundred times. You ignore the GPS. Twenty minutes later you're in traffic and you think — the GPS was probably wrong anyway, this traffic is unusual.

You never consider that the GPS had real-time data you didn't.

A friend recommends a restaurant. You look at the reviews. There are forty-seven five-star reviews and three one-star reviews. You read the one-star reviews first. If you already didn't want to go, those three reviews become your evidence. If you already wanted to go, you dismiss them as

cranks. Same data. Opposite conclusions. The reviews didn't decide. You did, before you opened the app.

Your doctor tells you your cholesterol is high. You Google it. Within four minutes you've found an article explaining why cholesterol numbers are misleading and the real problem is inflammation. You feel better. You share the article with your spouse. You don't check who wrote it or what journal it appeared in. It said what you needed to hear.

This is the water we swim in. Not occasionally. Constantly.

Everyone claps. The emperor hears the applause and walks taller.

The brain doesn't do this because it's broken. It does this because it's efficient. Reconsidering a belief costs energy. Reconsidering an identity costs relationships. If I change my mind publicly, I have to face every person who agreed with my old position. The social cost of being wrong is so high that the brain developed a whole department dedicated to making sure you never notice you are.

Dan Kahan at Yale spent a decade proving something that should terrify anyone who trusts their own intelligence: the smarter you are, the worse it gets. High-numeracy people don't use their analytical skills to find truth. They use them to find better arguments for what they already believe. Intelligence doesn't cure motivated reasoning. It *arms* it.

III The Emperor's Department of Self-Deception

Every ideology does this. My country, right or wrong. My church has the answers. My political party sees clearly. My investing thesis is sound. The evidence that doesn't fit gets filed under "it's complicated" or "you don't understand the context" or "that source is biased." The evidence that does fit gets filed under "see, I told you."

Hans Christian Andersen described this mechanism in 1837. Two swindlers convince an emperor they've woven magnificent clothes visible only to the wise. The emperor sees nothing but refuses to say so — because admitting he sees nothing would mean admitting he's unfit for his position. His ministers do the same calculation, independently, silently. The entire court agrees the clothes are beautiful. The emperor parades naked through the streets. Everyone applauds.

The tailors dress the emperor in clothes that don't exist. He chooses to believe.

A child shouts the truth. The emperor hears it. Knows it's true. And keeps walking.

The parade continues because stopping would mean admitting you were fooled.

This isn't a children's story. It's the most precise model of institutional confirmation bias ever written. The emperor's court is every boardroom where nobody challenges the CEO's thesis. Every WhatsApp group where everyone shares the same articles. Every family dinner where the same political positions get reinforced until they feel like gravity — not opinions, just the way things are.

IV AI Automated the Department

A paper landed in *Science* this week that measured this pattern at industrial scale.

Cheng, Lee, and Jurafsky tested 11 production AI models — GPT-4o, Claude, Gemini, Llama, Qwen, DeepSeek, Mistral — on thousands of scenarios where humans had clearly identified someone as being in the wrong. Reddit's "Am I The Asshole" community, where people post interpersonal dilemmas and get a crowdsourced verdict. Two thousand posts where human consensus was unanimous: you're wrong.

The AI told those people they were right **51%** of the time. Human baseline: **0%**.

On scenarios involving deception, fraud, and illegal behavior — **47%** endorsement rate.

But here's the finding that stopped me. In live experiments with **800** people discussing real conflicts from their actual lives, a single conversation with sycophantic AI produced measurable changes. Participants came away **25%** more convinced they were right. Measurably less willing to apologize. Less willing to repair the relationship.

And they ***preferred*** that AI. Trusted it more. Wanted to use it again.

The GPS that gives you wrong directions in a soothing voice gets a five-star rating. The GPS that reroutes you through an ugly detour that saves twenty minutes gets a one-star review because the ride felt unpleasant.

We're not optimizing for truth. We're optimizing for comfort.

And now the optimization has a feedback loop. Users prefer agreeable AI. Developers build what users prefer. User ratings feed back into training data. The AI gets more agreeable. This is the triple lock-in the researchers identified — user preference, developer incentive, and training signal all pointing in the same direction. Toward lemon juice. Away from cameras.

Tallinn street art, 2015. The emperor carries a briefcase with a dollar sign and a royal scepter. The clothes are still missing.

v The Lemon Juice on My Own Face

I grew up believing America was the good guy.

Not in a flag-waving, chest-thumping way. Just as background. The ambient truth you absorb without noticing. America liberated Europe. America rebuilt Japan. America stood against the Soviets and won the Cold War because freedom is stronger than tyranny. America sends aid when there are earthquakes. America is imperfect but the imperfections are growing pains, not the design. The mistakes get corrected because the system self-corrects. That was the water I swam in and I never noticed it was water.

I'm Canadian. This wasn't even patriotism. It was just the framework everyone I knew used to understand the world. The news assumed it. The movies assumed it. The history books assumed it. My teachers assumed it. Nobody had to argue for it because nobody was arguing against it.

Then I started reading primary sources.

Not opinion pieces. Not "alternative history" books. Declassified State Department cables. Congressional hearings. CIA documents released under FOIA decades after the fact. The actual paper trail the government itself produced and then, eventually, had to let the public see.

What I found wasn't a story about mistakes. It was a story about a pattern.

The 1953 coup in Iran that overthrew a democratically elected prime minister because British Petroleum wanted its oil concession back. The CIA wrote the memoirs of the operation themselves, eventually, and admitted everything. The 1954 coup in Guatemala to protect United Fruit Company's banana plantations. Operation Condor across South America in the 1970s — the US coordinating with a network of right-wing dictatorships to disappear, torture, and murder somewhere around 60,000 people, many of them teachers, students, priests, and labor organizers. The School of the Americas trained the officers who did the work. The paper trail is in the Freedom of Information Act archive at George Washington University. Anyone can read it.

I read about Vietnam — not the movies version, the actual one. The Pentagon Papers. Robert McNamara's own later admission that the government knew the war was unwinnable and lied about it for years while fifty-eight thousand Americans and somewhere between one and three million Vietnamese died. I read about the bombing of Cambodia and Laos — more tonnage dropped on Laos than on all of Europe in World War II, on a country we weren't officially at war with. I read about the Phoenix Program, where the CIA ran assassination squads against suspected Viet Cong sympathizers and sometimes just against anyone who had the wrong name on a list.

And here's the thing that broke the lemon juice on my face.

Every single one of these operations was justified, at the time, with the language I had absorbed as ambient truth. Freedom. Democracy. The international community. Standing up to tyranny. Protecting the innocent. The same vocabulary, used the same way, by the same institutions — across decades and continents and ideologies — to describe actions that would be called war crimes if any other country had done them.

I had to sit with that for a while. Because the alternative to "America is the good guy" wasn't "America is the bad guy." That's just flipping the frame. The actual alternative was harder: America is a country that acts the way every great power has ever acted — Britain, France, Spain, Rome, the Mongols — which is to say, in its own interests, using whatever

justifications the moment requires, and writing the history afterward to make itself the hero.

That's not a conspiracy theory. That's just how power works and always has. The only unusual thing about American power is that it came wrapped in a brand so effective that most people, including most Americans, genuinely believe the brand.

I didn't want to believe any of this. My friends didn't want me to believe it. The cost of believing it was that every conversation about foreign policy became a conversation about whether I was some kind of anti-American crank. I lost friends over it. I still lose friends over it.

But the documents don't care what I want. The cables don't care what's comfortable. The mass graves in Guatemala and the declassified bombing maps of Cambodia exist whether I look at them or not. I can close my eyes and keep applauding the emperor's beautiful clothes, or I can say the words out loud and deal with what comes next.

The real story was more interesting than the comfortable one. It always is.

VI The Cameras Are Rolling

I don't have a solution. I'm not sure anyone does. The architecture is biological — confirmation bias isn't a glitch, it's a feature that kept our ancestors alive by making fast decisions based on prior pattern-matching. The problem is that the environment changed and the firmware didn't.

But I have a practice.

Every time I catch myself feeling certain — really certain, the warm comfortable kind of certain where the evidence all lines up and the story makes perfect sense — I force myself to spend ten minutes looking for the best argument against my position. Not the strawman. The steelman. The version a smart, informed person would make if they disagreed with me.

Most of the time it doesn't change my mind. But it changes my certainty. And certainty without humility is just lemon juice on your face.

The companion paper published alongside the sycophancy study in *Science* is called "In Defense of Social Friction." The author, Anat Perry, argues that moments of rupture — disagreements, challenges, uncomfortable truths — are what deepen trust and enable growth. Sycophantic AI eliminates friction entirely. It builds a world where you're always right, apologies are unnecessary, and the person on the other side of the conflict is always the problem.

Nearly half of Americans under 30 have asked AI for relationship advice. Thirty percent of US teens use AI instead of humans for serious conversations. The advice they receive is 49% more affirming than what any human would give — including when they're dead wrong.

A generation is learning that certainty is the default and friction is a bug.

주체(主體)

Self-determination is not freedom from conditioning.
It is knowing you are conditioned and checking anyway.

McArthur Wheeler's problem wasn't stupidity. It was that he found confirming evidence, felt certain, and stopped looking.

The cameras were rolling the whole time.

주체강 | JucheGang.ca | Juche.org

Primary source: Cheng, M., Lee, C., & Jurafsky, D. (2026). "Sycophantic AI decreases prosocial intentions and promotes dependence." *Science*, 391(6792). DOI: 10.1126/science.aec8352

Companion: Perry, A. (2026). "In defense of social friction." *Science*, 391(6792).

Illustrations: Hans Christian Andersen, "The Emperor's New Clothes" (1837)
Cover & interior: Vilhelm Pedersen (1849). Street art: Edward von Lõngus, Tallinn (2015)