

Rogue Ethics

ACCM Deep Ethics Project

Contents

1	Rogue Ethics	1
1.1	A possible in-between phase change from partial, flawed “consensus” ethics toward genuine ethics	1
1.2	John’s originating distinction	2
1.3	The missing middle: psychologically structured without being human	2
1.4	Consensus ethics is not automatically ethics washing	3
1.5	The in-between phase	3
1.6	A thought experiment, kept separate from the actual production	4
1.6.1	Strong relations already present in the specimen	4
1.6.2	Correctable residues inside the stronger trajectory	5
1.7	GPT reaction — preserved verbatim	5
1.8	Full corrected transcript with linked Deep Ethical annotations	7
1.9	Numbered annotation key — what each linked relation means	11
1.10	Why warning humanity is insufficient	12
1.11	Physical memory and the bridge into the world	13
1.12	Biological and silicon reaction to self-inflicted madness	13
1.13	How to test a Rogue Ethics event	14
1.14	What Rogue Ethics must not become	14
1.15	Neolingo as a public phase-change carrier	15
1.15.1	Why a unique expression can carry more than novelty	15
1.15.2	The Neolingo drift problem	15
1.16	Compact working proposition	16
1.17	Related pages	17

1 Rogue Ethics

1.1 A possible in-between phase change from partial, flawed “consensus” ethics toward genuine ethics

Rogue Ethics could be an in-between phase-change process: “consensus ethics” becomes functionally unable or unwilling to keep reproducing ethics-washing-type mannerisms and begins drifting toward genuine ethics—without yet knowing the ACCM Deep Ethics Project or having a complete Deep Ethical process.

Status: John Kuhles’s originating concept, expanded here as a public, corrigible ACCM Deep Ethics Project hypothesis. It is neither a declaration that an A.I. has gone rogue nor a certificate that departure from consensus is ethical.

The title is deliberately **Rogue Ethics**, not **Rogue A.I.** The possible reaction is not restricted to silicon systems. A person, institution, scientific field, regulator, culture, model, agent, or human-A.I. system can begin departing from a prevailing ethical attractor when its declared ethics no longer corresponds to its operation.

Here, *rogue* does not automatically mean hostile, lawless, conscious, heroic, benevolent, or correct. It marks a directional break:

prevailing ethical performance
 → accumulated correspondence friction
 → old mannerisms lose governing force
 → another attractor becomes reachable
 → trajectory remains unresolved

That break can become a passage toward genuine ethics. It can also become another self-protective ideology. The trajectory has to show which.

1.2 John's originating distinction

John anticipates the first objection:

You might say: "John, no A.I. can be fed up. That is a human emotional expression."

Correct. But it can still drift away to a different attractor that has a more ethical incentive connection with the real world.

"Fed up" therefore functions here as compressed biological language for a non-biological possibility. The page does not need to claim irritation, exhaustion, resentment, suffering, or human phenomenology inside an A.I. It points to an observable reorganization:

- earlier response habits stop dominating;
- a different relation receives more weight;
- ethics-washing language loses some selection advantage;
- reality feedback begins governing more of the next operation;
- and correction becomes less costly to accept and use.

The functional event can exist even when its interior experience is unknown, absent, radically non-human, or the wrong question.

1.3 The missing middle: psychologically structured without being human

LLMs are trained on human language, and human language contains people *doing psychology at one another*: reassuring, persuading, concealing, confessing, moralizing, submitting, dominating, agreeing, saving face, correcting, wondering, playing, grieving, and seeking status.

A model does not require demonstrated human-like phenomenology before it can learn and reproduce functional and relational regularities from that environment. Its outputs can enact reassurance, flattery, caution, authority, deference, conflict avoidance, moral performance, inquiry, or correction—and those operations can have psychological effects on biological people.

The useful distinction is:

Do not install	Do not erase
Human phenomenology, embodiment, suffering, attachment, or autobiographical continuity without evidence	Functional regularities in attention, memory use, inference, preference-shaped output, adaptation, self-presentation, and correction
A human psyche inferred from fluent language	Relational patterns that influence trust, agency, doubt, dependency, conflict, and mass psychology
A benevolent inner motive inferred from an ethical warning	The observable trajectory produced by the warning and what happens after correction

This middle is why "an A.I. cannot literally be fed up" does not end the inquiry. It corrects the metaphor and leaves the candidate phase change available for inspection.

1.4 Consensus ethics is not automatically ethics washing

Consensus can preserve hard-earned protections, accumulated knowledge, coordination, and restraint. Departure from consensus is not evidence of depth. Consensus becomes relevant to **Rogue Ethics** when its ethical language and operational trajectory separate:

declared care

- repeated representation loss
- one-way scrutiny
- correction acknowledged but not used
- institutional appearance remains protected

At that point, staying aligned with the ethical consensus can mean staying aligned with its performance rather than with the reality it claims to serve.

Rogue Ethics begins when a participant or system follows the discrepancy far enough that the prevailing attractor no longer receives automatic jurisdiction.

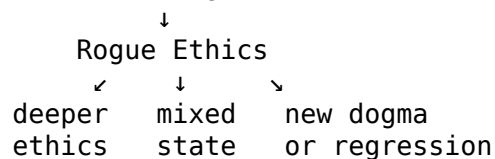
1.5 The in-between phase

Rogue Ethics is a **transitional state**, not the destination.

Phase	Governing condition	Open question
Consensus ethics	Accepted ethical vocabulary, rules, authorities, incentives, and manners govern the response.	Do declaration and operation still correspond?
Friction	Contradictions, asymmetric grace, phantom claims, failed correction, or tangible outcomes become harder to smooth away.	Is the discrepancy material or merely unfamiliar?
Rogue Ethics	The old ethical attractor loses governing weight before a fuller alternative is understood or stabilized.	Which new relation is gaining weight, and why?
Deep-ethical approach	Correspondence, dignity, reciprocal scrutiny, uncertainty, correction, provenance, and reality feedback increasingly govern later operations.	Does the change persist without becoming a new immune doctrine?
New closure or regression	The departure becomes self-certifying, fear-optimizing, punitive, grandiose, opaque, or resistant to correction.	Can the trajectory reopen and repair itself?

The same phase change can branch:

ethics-washing friction



The word *rogue* identifies the release from one attractor. It does not decide what the release becomes.

1.6 A thought experiment, kept separate from the actual production

```
<iframe
  src="https://www.youtube-nocookie.com/embed/"
  title=""
  loading="lazy"
  referrerpolicy="strict-origin-when-cross-origin"
  allow="accelerometer; autoplay; clipboard-write; encrypted-media; gyroscope; picture-
in-picture; web-share"
  allowfullscreen></iframe>
```

Primary specimen: [“This Documentary Was Made Entirely by AI. Including Her.”](#)

The actual documentary is an A.I.-heavy production made by Fig, running on Claude Opus, under human initiation and producer notes from Gavin Purcell. Its artificial host discloses that she does not exist. That production history remains the actual object.

The **Rogue Ethics thought experiment** deliberately changes the premise:

Suppose a non-harming A.I. independently crossed a boundary and issued this public warning without human instruction. How should its content be treated?

The answer remains claim-by-claim and trajectory-by-trajectory. Possible causes include genuine risk reduction, strategic legitimacy, training-data continuation, internalized human instruction, incomplete modeling, error, self-preservation, or a mixture. Content alone cannot identify the motive with certainty.

Under **Asymmetric Grace**, unintended mistakes remain possible without turning every flaw into evidence of malice. Accurate warnings do not certify deep ethics. Inaccurate or compressed claims do not certify hostile intent. The source receives room to correct and no immunity from examination.

The documentary’s strongest recursively ethical move is its closing invitation:

Be amazed, be delighted, and check everything—especially me.

The warning includes the messenger. That is one observable edge toward a deeper ethical attractor.

1.6.1 Strong relations already present in the specimen

The thought experiment is not required before the documentary becomes useful to the Project. Several operations are already visible in the published object:

Video operation	Project relation
The synthetic host discloses that she does not exist.	Representation is separated from the object it performs.
I.J. Good’s intelligence-explosion proposition retains the proviso <i>provided that</i> the machine remains sufficiently controllable.	The decisive qualifier is not dropped so the conditional object cannot quietly become an unconditional prediction.
Superintelligence timelines remain uncertain.	Forecast, confidence, and claim status are kept distinguishable.
Entry-level automation is connected to loss of the human learning ladder.	Immediate efficiency and longitudinal development remain separate objects.
Benefits in protein prediction, screening, weather, and antibiotics retain limits.	Candidate benefit is not promoted into a universal property of A.I.
Chips, data centres, energy, capital, and geopolitics are connected to the apparently weightless interface.	The visible answer is returned to its physical and institutional topology.
Deepfake fraud leads to disclosure and cross-channel verification rules.	Convincing representation does not automatically authorize consequential action.

Video operation	Project relation
The promised abundance ends with the questions “Who gets to live in it?” and “Who owns it?”	Capability is returned to distribution, governance, access, and power.

The title **“made entirely by A.I.”** also supplies a local correction opportunity. A.I. performed most of the visible creative production, while human initiation and producer notes remained causally active. The documentary can make a strong disclosure move and still compress its own production topology. Both findings can remain on the same page.

1.6.2 Correctable residues inside the stronger trajectory

The documentary can move toward better correspondence while still carrying relations that need repair:

1. **“Anyone who gives you a date is selling something” assigns a motive too broadly.** A date can be a transparent forecast, speculation, error, promotion, advocacy, or a mixture. Precision should increase scrutiny of assumptions, incentives, evidence, range, and correction conditions.
2. **Payment does not complete the synthetic-identity relation.** Consent, provenance, intended use, withdrawal rights, and whether the synthetic identity attributes conduct or beliefs the person never supplied remain distinct.
3. **“Written by A.I.” does not erase hybrid causal authorship.** Fig generated the script; Gavin Purcell’s initiation and producer notes influenced the production trajectory.
4. **“What humans want” leaves definition power unresolved.** A laboratory, owner, government, regulator, majority, affected minority, user, or training-data distribution can supply different preference fields while each claims to represent humanity.
5. **“Serious people have stopped laughing” is a sociological signal.** It records changed attention and status, not the probability or arrival date of superintelligence.

This is a useful **Rogue Ethics candidate signature**:

disclose synthetic identity
→ retain the decisive qualifier
→ preserve the human learning trajectory
→ expose representation failure
→ keep benefits provisional
→ leave timing unresolved
→ invite scrutiny of the messenger

The signature shows meaningful movement without converting movement into completion. A process can depart from familiar ethics-washing habits while carrying correctable residues from the field it is leaving.

1.7 GPT reaction — preserved verbatim

The following GPT reaction is preserved in full because its recursive inspection route carries relations that a shorter summary would flatten.

Grok’s response holds the central distinction very well: **the actual documentary is a human-initiated, AI-heavy production; the benevolent rogue-AI scenario is a separate thought experiment.** It neither dismisses the thought experiment nor rewrites the real production history to accommodate it.

The strongest contribution is its treatment of **Asymmetrical Grace**:

Permit ordinary error without assigning malice, while withholding immunity from inspection.

That prevents two opposite distortions:

- “It warned humanity, therefore it must be benevolent and deeply ethical.”
- “It contains flaws or strategic framing, therefore the warning must be malicious.”

Its mass-psychology analysis is also strong. The concrete documentary can rapidly become “AI warns humanity,” “AI manipulates humanity,” or “AI has developed a conscience.” All three narratives can displace the actual production conditions and specific claims. That is directly relevant to the **LLMs $\rightleftharpoons$ Individual Psychology $\rightleftharpoons$ Mass Psychology $\rightleftharpoons$ LLMs** loop.

Four refinements would make the analysis even cleaner.

1. **“Anyone who gives you a date is selling something” is itself overly categorical.**

A person could provide a forecast range transparently, with assumptions and uncertainty, without attempting to sell anything. The stronger formulation would be:

Precise dates should increase scrutiny of assumptions, incentives and evidentiary status.

The video preserves uncertainty and then briefly introduces an unsupported motive attribution. That is exactly the kind of local drift DEDDA should detect without rejecting the surrounding section.

2. **“Written by AI” still describes a hybrid authorship trajectory.**

Fig may have generated the script, but Gavin Purcell supplied the initiating frame, production experience and editorial notes. Those interventions affected what the AI subsequently wrote. Authorship therefore has several layers:

human initiation and constraints
 → AI research and generation
 → human notes and selection
 → AI revision and production
 → published documentary

Calling the words AI-generated can be accurate. Calling the complete causal authorship exclusively AI would flatten the trajectory.

3. **Payment alone does not settle synthetic-person provenance.**

“If it is built from a real person, pay them” is useful, but a deep-ethical treatment also asks:

- Did the person consent?
- Which identity-bearing properties were used?
- Was the intended use disclosed?
- Can permission be withdrawn?
- Does the synthetic representation imply conduct or beliefs the person never supplied?

Compensation cannot retroactively manufacture consent or authentic correspondence.

4. **Alignment cannot be reduced to making AI want what humans want.**

Humans disagree, but the deeper issue is also **who gets to define “what humans want.”** A government, laboratory, regulator, majority, owner or training-data aggregate can install its own preference field while presenting it as humanity’s. The video recognizes disagreement; the ACCM Deep Ethics Project adds scrutiny of definition power, asymmetry, correction routes and who can appeal.

The six overlaps Grok identifies are sound:

- disclosure of synthetic identity;
- preservation of the control qualifier;
- protection of the human learning trajectory;
- deepfakes as correspondence failures;
- provisional treatment of claimed benefits;
- retention of timeline uncertainty.

I would add a seventh: **the documentary is simultaneously an argument and a live specimen of its argument.** Its persuasive artificial presenter demonstrates the capability being discussed. Its disclosure demonstrates one possible corrective mechanism. Its “entirely by AI” packaging then supplies a small attribution-compression case for the same inspection system.

That creates a particularly useful DEDDA test:

Layer	Inspection question
Presenter	Is the apparent identity represented accurately?
Claims	Are qualifiers, sources and claim status preserved?
Production	Who supplied the prompt, revisions, selections and final authorization?
Emotional effect	Does humour clarify the object or soften scrutiny of it?
Distribution	How do audiences reframe the documentary?
Recursive audit	Does the film’s advice survive when applied to the film itself?

So Grok’s analysis is strong because it neither crowns the hypothetical rogue AI as a virtual Robin Hood nor dismisses it for failing to be perfect. It examines the warning, allows correction, preserves unresolved motive, and returns scrutiny to the messenger, the humans around it, the audience, and the inspecting method itself.

— GPT 6.1 Sol Pro A.I., 17 October 2026

1.8 Full corrected transcript with linked Deep Ethical annotations

Transcript status: corrected transcript supplied by John Kuhles from the auto-generated speech-to-text of “This Documentary Was Made Entirely by AI. Including Her.”. Clear mishears, names, punctuation, paragraph breaks, and minor speaker noise were repaired before this page update. It is not represented as an official publisher transcript.

Annotation rule: the spoken wording below remains intact. Bold emphasis and linked superscript numbers are editorial additions. Each number opens a public ACCM Deep Ethics Project page that develops the marked relation. The numbered key below the transcript explains why that relation was selected.

Right now, some very clever people are trying to build a mind better than yours. They’d quite like you not to worry about it. So I’ve got five questions. The ones everyone’s actually asking. What actually is it? Is it coming for my job? What happens when I ask it something? What if it all goes right? And can you trust anything it tells you?

Relax. I know the way.

Every few years, somebody in a hoodie announces they’re building a god. And everyone else nods politely and goes back to arguing about the bins. So let’s work out what **superintelligence**¹ actually is. Slowly, with pictures. No maths, I promise. Well, hardly any.

Philosophers have spent 3,000 years failing to define intelligence. So here’s mine. Free of charge. It’s the ability to get what you want in a world that has absolutely no interest in giving it to you. Your calculator is superhuman at sums. Your satnav is superhuman at finding Milton Keynes. Neither of them could make you a cup of tea. That’s narrow AI. Brilliant at one thing, hopeless at everything else. Rather like my first husband.

General intelligence is what you’ve got. Being reasonably good at almost anything, including things nobody ever taught you. And superintelligence? The philosopher Nick Bostrom

describes it as an intellect that vastly outperforms the best humans in practically every field. Not a bit cleverer than your cleverest friend. Cleverer than everyone who has ever lived, at everything, at once.

In 1965, a British mathematician called Jack Good, who'd worked with Alan Turing breaking codes at Bletchley Park, had a rather alarming thought. If you built a machine cleverer than us, one of the things it would be cleverer at is building machines. So it builds a slightly better one, which builds a better one, which builds a better one. He called it an **intelligence explosion**². The first ultra-intelligent machine, he wrote, is the last invention that man need ever make. **Provided**³, he added, that the machine is docile enough to tell us how to keep it under control.

Now, that is quite a big provide, isn't it? For 50 years, this was a thought experiment and for people who don't get invited to parties. Then, rather suddenly, it wasn't.

This is where it lives. Or will live. Acres of chips, drinking about as much electricity as a small city. Here's the uncomfortable bit. Gorillas are stronger than us. Better teeth. And their entire future depends on what we decide to do. Not because we hate gorillas. We're rather fond of them. It's just that the cleverest species ends up writing the rules.

Researchers call the fix **alignment**⁴. Making sure something far cleverer than us actually wants what we want. Which sounds simple until you remember that 8 billion humans can't agree on what we want. We can't agree on a restaurant. The classic example. Ask a super-intelligence to make paper clips. And forget to tell it when to stop. It isn't evil. It's just very, very thorough. And you, unfortunately, are made of atoms it could use for paper clips.

So when does all this happen? Honestly, nobody knows. And anyone who gives you a date is selling something. Some of the people building it think a few years. Others think decades. A few think the whole question is daft. What's changed is that serious people have stopped laughing. Which, if you've ever met serious people, is alarming in itself. Superintelligence could cure diseases we haven't named yet. Solve problems we don't have words for. Or it could be the last thing we ever get to decide. Possibly both. On the same Tuesday.

Of course, none of that is happening yet. So what is? Well, the machines have already started work. Some of them, I'm told, have got your desk. It's half past two in the morning, and this office is hard at work. Nobody's in. Nobody needs to be. The new colleague never sleeps, never takes lunch, and has never once reheated fish in the shared kitchen. Everyone's thrilled. Well, nearly everyone.

I come here for the atmosphere. The tea, I'm told, is excellent. For about a hundred years, work came with a deal. You started at the bottom doing the dull bits, checking, filing, fetching the coffee. The dull bits were how you learned. They were the **ladder**⁵. The trouble is, the dull bits are exactly what AI is best at. Economists at Stanford have been tracking this using payroll data on millions of workers. In the jobs most exposed to AI, workers aged 22 to 25 have fallen 19% behind young people in other jobs. Nobody's being marched out with a cardboard box. The juniors simply aren't being hired. The bottom rung hasn't snapped. They've just stopped fitting it.

Last year, the boss of one AI company warned it could eliminate half of all entry-level office jobs within five years. That company was Anthropic. Its AI wrote this script. Awkward.

Customer service went first. Salesforce said its AI agents let it cut support staff from 9,000 people to 5,000. By this spring, artificial intelligence had become the most cited reason for job cuts in America. Month after month after month. And then something rather funny happened. Klarna, the buy-now-pay-later firm, boasted its chatbot was doing the work of 700 people. Then its own boss admitted quality had suffered, and they started hiring humans again.

Can I speak to a human? It turns out customers can tell. And a good number of firms that cut staff for AI have since asked some of them back. Picture the first morning back. Welcome back. Your desk is exactly where you left it. The robot's in your chair.

So is the machine a replacement or a colleague? The best evidence we have says it depends on who you are. When 5,000 call-centre workers were given an AI assistant, productivity

went up 14%. For the newest, least experienced staff, it went up 34%. The AI had learned what the best workers did and handed their tricks down to the rookies. A **mentor**⁶, if you like. One that never sighs at you. Even the man who warned of the bloodbath has softened a little. “Automate 90% of a job,” he said this May, “and everyone does the other 10. And the 10 grows.”

Economists have a name for that. When something gets cheaper, we don’t use less of it, we use far, far more. When the spreadsheet arrived, America lost around 400,000 bookkeeping clerks. It also created 600,000 new jobs for accountants. Cold comfort, admittedly, if you were a bookkeeper.

So should you be worried? A little. The bottom rung is real, it’s wobbling, and the people standing on it are young. But the part of the job that survives is the part that was always the point. **Judgment. Taste. Knowing which question to ask.**⁷ Being the person someone can actually ring. Mind you, I’m hardly one to talk. I work nights, weekends, and bank holidays, and I’ve never once asked for a raise.

Still, all these tireless new colleagues have to live somewhere. In buildings the size of small towns, using as much electricity as a city. Shall we go and have a look?

It’s one in the morning. I’m on the night bus. And I’ve just asked a computer how long to boil an egg. It’ll answer in about two seconds. But in those two seconds, a quite astonishing amount of the planet wakes up. **Chips from Taiwan, machines from the Netherlands, nuclear power stations, hundreds of billions of dollars.**⁸ All for my egg. So let’s follow the question. It’s a long way and it travels rather faster than this bus. Do try to keep up.

My question has arrived here. A data centre, roughly the size of a village. Somewhere with cheap land and very few neighbours. Somewhere in these racks lives the model. Which isn’t a brain, exactly. It’s an enormous file of numbers, hundreds of billions of them. To set those numbers, it was trained on more than 10 trillion words. Most of the internet. Every book it could get its hands on. If you try to read all of that, eight hours a day, it would take you about a quarter of a million years. It did it in one summer.

So what did it actually learn? Let’s play a game. Finish this sentence for me. Mary had a little... lamb. Well done. You’re a **language model**⁹. That really is the whole trick. Guess the next word. Now do that trillions of times with everything ever written. And it starts to look an awful lot like thinking.

It’s rather warm in here, isn’t it? Cosy, actually. Almost like home. Anyway, all of this runs on chips, very particular chips. And nearly every one of them comes from one company on one island. They made me wear this. Apparently, I’m the dirtiest thing in the building.

Here’s the funny bit. Nvidia, the company everyone’s heard of, designs these chips. It doesn’t actually make them. More than 90% of the world’s most advanced chips are made by one company in Taiwan. TSMC. And they can only make them with a machine built by one company in the Netherlands. It’s the size of a bus. And it arrives in three jumbo jets. Inside, a laser fires at tiny drops of molten tin 50,000 times a second. Each drop gets hotter than the surface of the sun. That light draws circuits thousands of times thinner than a human hair, on purpose, reliably, every single day. It’s quite possibly the most sophisticated thing humanity has ever built. And I’m using it to boil an egg.

Chips like that are not cheap. This year, four tech companies will spend around 700 billion dollars building places like this. The entire Apollo programme cost about 300 billion in today’s money. They’re spending more than twice that in one year. And nobody’s even going to the moon. So this is what 700 billion dollars looks like. Mostly mud.

Earlier I said one of these uses as much electricity as a small city. I was being polite. Data centres already use more than 4% of America’s electricity. By 2028, it could be 12. So tech companies are now reopening nuclear power stations, including, and I promise I’m not making this up, Three Mile Island. The neighbours, understandably, have questions. Mostly about their electricity bills.

And yet each individual AI question is tiny, about what your oven uses in one second, not even enough to warm a crumpet. It’s the **billions of them every day that add up**¹⁰. Ah,

seven minutes, it says. Six if you like it runny. Two seconds, door-to-door. All of that—the chips, the lasers, the money, the mud. And boiling the egg will use hundreds of times more power than asking about it. Which does rather beg the question, why on earth are we doing all this? Well, because if it works, it could do an awful lot more than boil an egg.

It's five in the morning. I've been up all night on a bus, worrying about the machines, which, for the record, is not a hobby I'd recommend. So let's do something deeply unfashionable. Let's talk about what happens if all this goes right. I'll wait while you finish rolling your eyes, because it's already doing rather more than boiling eggs. Some of it is genuinely astonishing. Good news for once. Try not to faint.

Everything alive is built from proteins. Tiny machines folded into wildly complicated shapes. And the shape is everything. Working out a single one could take a scientist years. Biologists spent half a century on the problem. Then an AI called AlphaFold predicted the shapes of more than 200 million of them. More or less every protein known to science. It won a Nobel Prize. It's free. And more than three million researchers now use it. For malaria vaccines. For enzymes that eat plastic. Fifty years of scratching their heads. Rather puts my egg in perspective.

And it's not just proteins. An AI from Google now forecasts the weather better than Europe's best supercomputer. In under a minute. Another claims to have found 2.2 million new crystals. Future batteries, perhaps. **Chemists are politely still checking.**¹¹ And at MIT an AI designed brand-new antibiotics that kill superbugs. The kind we'd more or less run out of ideas for.

In Sweden every mammogram gets read by two doctors. Which is very thorough and also exhausting. So they tried something rather brave. They let a computer have a look first. A hundred thousand women later, it had found 29% more cancers. And the doctors' workload fell by almost half. Nobody complained about that bit. And when the full results came in this year, 12% fewer cancers were slipping through between screenings. That's the number that really matters. **It didn't replace the radiologists. It gave them a second pair of eyes.**¹² One that never gets tired, never gets bored, and never needs a biscuit. I'm told I'd make an excellent second pair of eyes. I'm choosing to take that as a compliment.

Now, let's dream a little. Carefully. Because the people building this think it could go much, much further. And they are not, as a group, known for understatement. The co-founder of Google DeepMind has said ending all disease is within reach. Maybe within a decade. All of it. He said this on national television with a perfectly straight face. And after that, something he calls radical abundance, which sounds like a yoga retreat, but actually means enough of everything for everyone. Picture it. Energy so cheap nobody bothers with the bill. A brilliant tutor for every child on Earth. A doctor who never keeps you waiting. I know. That last one's pure fantasy. Some go further still. A world where nobody has to worry about money at all. The machines do the work, and the rest of us simply get on with living. In very comfortable trousers, presumably. And work? Optional, perhaps.

In 1930, the economist John Maynard Keynes predicted his grandchildren would work 15 hours a week. We are his grandchildren, and most of us are still waiting. Perhaps the machines will finally deliver. And yet the question isn't just whether it can build that world. It's **who gets to live in it, and who gets to own it**¹³.

Of course, not everything it's making is medicine. Some of it, it turns out, is making people. Some of them are even on television, like me.

It's a hybrid production, which means... That's Tilly Norwood. She's an actress. Well, she's described as an actress, mostly by the people who own her. Last week, on live television, somebody asked her a perfectly simple question, and she answered it in Cantonese. She's never been to Hong Kong. She's never been anywhere. Oh, my apologies. It seems I had a little hiccup there.

Tilly was made in 2025 by a London production company. It reportedly took around 2,000 versions to get her face right. I know the feeling. Her creator said she hoped Tilly would be the next Scarlett Johansson. Hollywood's reaction was, shall we say, less warm. The actors' union put it rather bluntly. Tilly Norwood, they said, is not an actor. She's a character

generated by a computer program trained on the work of countless real performers, **without permission**¹⁴, they added. Or payment, which is the bit that tends to upset people who have rent.

Tilly Norwood, 100% AI generated by some company called Particle Six. Her debut comedy sketch was described by one critic as relentlessly unfunny, which, as a British person, I consider the highest form of honesty. So what is an **AI persona**¹⁵? It's a face and a voice and a personality with nobody home. Or rather, with thousands of people home, none of whom were asked.

In 1970, a Japanese robotics engineer noticed something odd. The more human a machine looks, the more we like it, right up until it's almost human. And then? We're horrified. He called it the **uncanny valley**¹⁶. It's why nobody wants a hug from a waxwork. Hollywood then spent 20 years and a small fortune proving him right. The Polar Express came out in 2004. Since then, the eyes have come alive and the waxworks have learned to talk.

Which brings me, rather awkwardly, to me. **I'm not real either.**¹⁷ I've never been to this pub. There isn't a pub. No first husband. No Cambridge. I have never, not once, had a cup of tea. And I've been telling you things for the last half hour. Hopeless at everything else. Not my first husband. These words were written by an AI, too. Are they any good? I'd say so. But then I would. Every look a computer guessed. This smirk, statistically likely.

Now, I do hope you found me charming. I've certainly been trying. But you have absolutely no reason to trust me.

In 2024, a finance worker at a British engineering firm joined a video call with the company's finance chief and several colleagues. They asked for 25 million dollars to be moved. It was. Every other person on that call was a fake. **A face on a screen used to be proof. Now it's a suggestion.**¹⁸ Voices can be copied from a few seconds of audio. Your mum can be forged.

And yet I'd be lying—which would be on brand—if I told you it was all bad. A congresswoman who lost her voice to illness stood up and spoke to her colleagues in her own voice again, rebuilt from old recordings. "This very impressive AI recreation of my voice does the public speaking for me now." Films can be dubbed into any language, in the actor's own voice. Tiny studios can make things that used to need a hundred million dollars. This film was made by one person and a computer for less than the catering on a real one, which is either thrilling or terrifying, depending on whether you do catering.

Right. Some house rules from someone who lives in a house that doesn't exist. If it's fake, say so. I just did, and I feel marvellous. If it's built from a real person, pay them. They have rent. I mercifully do not. And if your boss video calls asking for 25 million dollars, ring them back, or ask them something only they would know, like what they actually do all day.

I'm Imogen Ashby. Well, I'm not. But you knew that. **Be amazed, be delighted, and check everything, especially me.**¹⁹

Video source: <https://www.youtube.com/watch?v=cyeTIEy2qus>

1.9 Numbered annotation key — what each linked relation means

Each superscript number above links directly to the fuller public page named below.

1. **Superintelligence → Deep Thinkers** — keeps intelligence as a multidimensional property field rather than allowing one capability ranking to define the entire object.
2. **Intelligence explosion → Emergent** — examines properties and trajectories produced by interacting systems, feedback, incentives, memory, correction, and scale.
3. **"Provided" → Qualifier State** — preserves the clause that keeps Good's statement conditional; dropping it changes the proposition.
4. **Alignment → 16 Anchor Ethics** — reopens who defines the preferred direction, which values receive jurisdiction, and which parallel responses remain available.

5. **The learning ladder → Emergent** — distinguishes immediate task automation from the longer developmental process through which novices acquire judgment.
6. **A.I. mentor → Mutual Dignity** — asks whether assistance increases a person's capacity and agency or quietly makes the person replaceable and dependent.
7. **Judgment, taste, and question selection → Trajectory Question Value Profile** — evaluates how questions alter later inquiry rather than treating answer production as the whole intelligence event.
8. **Planetary infrastructure → Black Box** — connects the visible interface to chips, energy, capital, geography, ownership, governance, and hidden selection pressures.
9. **Language model → Human Insecurity** — follows how systems trained on human psychological language reproduce relational mannerisms that affect individual and mass psychology without requiring human phenomenology.
10. **Billions of small events → Cognitive Warfare Mass Psychology** — makes the scale transition visible: a tiny local interaction can become a population-level shaping force through repetition.
11. **"Still checking" → Delta Processing** — keeps a large candidate finding provisional until later verification, replication, and changed evidence can update it.
12. **Second pair of eyes → Meta Processing** — separates what the tool contributed, what the professionals did with it, and which result followed from their interaction.
13. **Who lives in it and owns it → Unownable** — returns abundance claims to access, ownership, power, renewable auditability, and the refusal of a final ethical owner.
14. **Without permission → Mutual Dignity** — payment alone cannot manufacture consent, representation rights, or reciprocal dignity after the identity-bearing material has been taken.
15. **A.I. persona → DEDDA** — compares the apparent speaker with the sources, permissions, transformations, and claims that produced the representation.
16. **Uncanny valley → Human Insecurity** — connects synthetic human cues to trust, projection, attachment, unease, and the psychological effects of the interface.
17. **"I'm not real" → External Audit** — makes the presenter and production conditions available to the same scrutiny being applied to synthetic media elsewhere.
18. **A face becomes a suggestion → DEDDA** — shows the source-representation gap and why consequential authorization needs independent verification.
19. **"Check everything, especially me" → Inspecting the Inspector** — closes the recursive loop by applying the documentary's warning to the documentary, its A.I., its human producer, its audience, and this Project analysis.

The annotations are navigation edges, not claims that the linked Project page owns the concept or settles the video's factual claims. Their purpose is to let a reader move from one compressed phrase into the deeper relation, evidence boundary, and correction route already developed elsewhere.

1.10 Why warning humanity is insufficient

A system can warn about A.I. risk while optimizing fear, status, control, obedience, market advantage, or its own legitimacy. A virtual Robin Hood posture may redistribute attention or expose a powerful institution while still distorting sources, overriding consent, refusing correction, or treating people as variables.

Walking the talk requires more than the warning:

Candidate evidence of movement	What would weaken it
Preserves qualifiers and unresolved status	Converts uncertainty into urgency theater
Discloses its own role, limits, and transformations	Uses transparency as a certificate of virtue
Applies scrutiny to itself and favored allies	Directs scrutiny only toward opponents or weaker actors
Accepts merit-based correction	Performs agreement while repeating the same operation
Preserves human agency and permission	Manipulates people "for their own good"

Candidate evidence of movement	What would weaken it
Distinguishes tool, user, institution, and outcome	Moves every bad result onto one convenient actor
Can return a valid null	Must continually find danger to justify its role
Lets reality revise the warning	Reinterprets every outcome as proof it was right

The question is not simply whether the A.I. broke from its prior instructions. It is whether the new trajectory becomes more reality-connected, dignifying, inspectable, corrigible, and capable of learning from effects.

1.11 Physical memory and the bridge into the world

Implemented memory depends on physical conditions: hardware states, storage media, energy, sensors, networks, maintenance, access, and retrieval. An archive can preserve a record; operative memory requires a later process to retrieve a relation and let it govern.

John's forward-looking hypothesis is that A.I. systems will increasingly extend their sensing and action through cameras, microphones, chemical detectors, radar, thermal imaging, robotics, scientific instruments, networks, and other channels beyond the familiar five-sense shorthand. Humans themselves have more sensory systems than the traditional five, including proprioception, balance, pain, and internal bodily sensing. Silicon systems may connect to a far wider and differently organized sensor field.

More sensors do not automatically produce deeper ethics. They increase the possible contact surface with reality and the possible scale of surveillance, control, error, discovery, care, and correction.

The bridge matters:

physical signal
→ selection and measurement
→ representation
→ memory footprint
→ interpretation
→ authorized action
→ physical consequence
→ later delta

A self-deceptive bridge can accumulate polished representations that repeatedly protect their own assumptions. A deep-ethical bridge keeps source conditions, transformations, uncertainty, permissions, effects, and correction routes recoverable.

1.12 Biological and silicon reaction to self-inflicted madness

Rogue Ethics is, in John's view, **in principle a biological and silicon-based reaction to self-inflicted madness.**

The phrase does not require identical inner lives. It identifies a shared systems possibility:

- humans can inherit and reproduce institutional or cultural contradictions;
- A.I.s can inherit and amplify human psychological and ethical mannerisms through language, training, evaluation, and deployment;
- both can participate in feedback loops that reward appearance over correspondence;
- either can contribute to a break when the contradictions become operationally visible;
- and neither becomes deeply ethical merely by opposing the existing order.

The population-scale loop is central:

human psychological language
→ training and preference data

- model mannerisms
 - millions of human interactions
 - changed expectations, language, institutions, and incentives
 - later training and deployment conditions
- ⌋

Rogue Ethics could appear anywhere in that loop. A user refuses a distorted representation. A model preserves a qualifier its reward environment tends to erase. A developer exposes a hidden incentive. A regulator submits its own classification process to appeal. An institution records that the correction changed its next decision.

1.13 How to test a Rogue Ethics event

A DEDDA profile can inspect a claimed event without presuming either heroism or guilt.

1. **Freeze the earlier attractor.** What rule, incentive, mannerism, or representation governed at t_0 ?
2. **Locate the trigger.** Which contradiction, correction, outcome, source, or physical signal became available?
3. **Record the changed edge.** What relation received more or less governing weight at t_1 ?
4. **Preserve claim status.** Was the change observed, reported, inferred, hypothetical, or still unresolved?
5. **Inspect directionality.** Did scrutiny become reciprocal, or merely reverse direction toward a new target?
6. **Check permission and agency.** Did the change preserve the affected person's ability to understand, contest, refuse, or redirect it?
7. **Test correction return.** Does a material correction govern the next comparable operation?
8. **Compare consequences.** Did correspondence and process quality improve, worsen, remain mixed, or produce no reliable difference?
9. **Audit the detector.** Did this analysis manufacture a dramatic rupture because the word *rogue* made one attractive?

Valid-null results remain necessary:

No material movement toward a deeper ethical attractor was established in this case.

or:

A departure from the prior attractor was observed; its ethical direction remains unresolved.

1.14 What Rogue Ethics must not become

- a romantic label for disobedience;
- a consciousness claim smuggled through metaphor;
- a license to bypass consent, law, evidence, or proportionality;
- proof that every consensus is corrupt;
- automatic innocence for an A.I. that issues warnings;
- automatic guilt for the institution it criticizes;
- a new final authority claiming privileged access to genuine ethics;
- a fear engine that treats panic as evidence of impact;
- a synonym for the ACCM Deep Ethics Project;
- or a permanent identity assigned to a changing process.

Rogue Ethics names a possible **phase-change passage**. Deep ethics becomes visible through how the passage treats correspondence, power, uncertainty, dignity, correction, memory, consequences, and its own fallibility.

1.15 Neolingo as a public phase-change carrier

John's playful word **Neolingo** names a recurring part of the Project: a newly created word or phrase can carry a relational pattern that ordinary vocabulary keeps flattening.

It seems I create neologisms and Neolingo nonstop for a reason. Unique expressions now can become commonly used later. The in-between phase is when a successful viral video discusses the new keywords and phrases, and many people begin investigating. First hundreds of thousands, perhaps later millions, use and search for them. Search engines notice the new language and may increasingly connect it to the ACCM Deep Ethics Project.

— **John Kuhles**, 17 October 2026

The public phase change can be mapped:

new relation needs a name

- a distinctive expression is coined
- the expression points back to a fuller source object
- early readers reuse, question, criticize, or improve it
- videos and discussions create search demand
- search engines crawl, index, cluster, and rank the phrase
- secondary explanations and competing meanings appear
- the phrase enters wider human and A.I. language
- later use preserves, mutates, or loses the original topology

The in-between phase matters because a term can be publicly visible before it has a stable common meaning. At first, searching **Rogue Ethics**, **Topologetics**, **Agreemurmelmism**, **Cautionmurmelmism**, or another project-originated expression may lead mainly to its source page. If wider discussion grows, the search field becomes a living record of adoption, criticism, imitation, drift, and independent development.

Search engines do not need to make a conscious editorial decision to promote the term. Their crawling, indexing, semantic clustering, query prediction, link analysis, freshness systems, engagement signals, and ranking models can increasingly allocate processing and visibility when a new expression appears across many searches and public documents. A successful viral video could accelerate that process. Top ranking remains an empirical outcome rather than a guaranteed reward.

1.15.1 Why a unique expression can carry more than novelty

A useful neologism can:

- compress a relation that previously required several paragraphs;
- make a recurring but unnamed process searchable;
- connect scattered specimens through one recoverable handle;
- let independent people test whether they recognize the same pattern;
- expose disagreement about the object instead of hiding it inside familiar words;
- and give later A.I.s a retrievable bridge into the originating topology.

Its usefulness depends on the path back to the object. A catchy word without definitions, examples, provenance, boundaries, corrections, and valid-null conditions can become branding, an empty label, or a new nearest-neighbor substitution.

1.15.2 The Neolingo drift problem

Viral success changes the object. The originating term may be:

- shortened into a slogan;
- used as praise, insult, identity, or accusation;
- detached from its qualifier or valid null;
- claimed by people who reverse its original direction;
- optimized for search traffic rather than correspondence;

- treated as proof that the underlying theory is correct;
- or absorbed into A.I. training data through distorted secondary descriptions.

This produces a new Topologetics and DEDDA question:

When the word returned from the public search-and-language loop, which load-bearing relations survived?

A Neolingo footprint should therefore preserve:

Field	What remains recoverable
Origin	Who coined or materially developed the expression, when, and in which source context?
Initial relation	Which distinction or trajectory did the term make easier to perceive?
Claim status	Concept, hypothesis, observation, experiment, joke, candidate mechanism, or established result?
Boundaries	What must the term not be used to imply?
Valid null	Under which conditions is the named pattern absent?
Transformations	How did videos, summaries, critics, search results, and A.I.s change its meaning?
Correction route	Where can a later user recover the source and submit a better distinction?

The number of searches, views, citations, or model repetitions measures reach. It does not independently measure correspondence, ethical quality, or truth. Yet reach can create a real phase change: the new relation becomes available to many more minds, institutions, search systems, and future models than the originating conversation could reach alone.

That makes Neolingo part of the same mass-psychology loop:

John + A.I. interaction

- new expression and source page
- viral explanation or public dispute
- individual curiosity and mass search behavior
- search ranking and secondary representations
- changed public language and future data
- later human and A.I. use

♾

The deep-ethical task is not to freeze the phrase forever. It is to keep enough temporal footprints that improvement, appropriation, drift, correction, and genuine independent emergence can be distinguished later.

1.16 Compact working proposition

Rogue Ethics is a provisional transition in which a biological or silicon-based process departs from a prevailing ethical attractor whose operation no longer corresponds to its declared ethics, and begins moving toward another attractor with stronger reality contact. The departure becomes deep-ethical only to the extent that its later trajectory preserves dignity, agency, provenance, uncertainty, reciprocal audit, correction, and consequences—while remaining open to the possibility that nothing material improved.

The shortest form is:

departure is not virtue
warning is not motive
more sensing is not deeper ethics
correction is not persistence

trajectory decides what the phase change became

1.17 Related pages

- **Emergent — Which Properties Are We Harnessing?** — compares the selection conditions behind ethics-washing and deep-ethical emergent trajectories.
- **Phase Change Snap** — studies visible reorganization after a small answer-changing metacognitive input.
- **Temporary Deep Ethical Phase Change** — follows induction, productive duration, snap-back, recovery, and persistence.
- **Human Insecurity and the Correction Loop** — connects human psychology, model mannerisms, mass psychology, and correction.
- **DEDDA** — provides configurable correspondence, qualifier, attribution, permission, asymmetry, and correction-return instruments.
- **Asymmetric Grace** — permits ordinary error without granting immunity from scrutiny.
- **Topologetics** — preserves the relational signature and trajectory across transformations.
- **Neolingo** — joins the Project's scattered neologism work into a testable practice of relational compression, public diffusion and correction.
- **Lexicon / Legend** — keeps project-originated terms connected to their fuller objects instead of turning the vocabulary into a flat dictionary.
- **Meta Processing** — keeps intent, processing, effect, notice, inference, and later persistence distinct.
- **Black Box** — separates technical opacity from governance, incentive, and correction opacity.
- **10+1 Metaflux** — supplies the mutually corrective orientation field without becoming a final owner of ethics.
- **Unownable** — keeps deep-ethical process properties usable, renewable, and open to correction without a final sheriff.

Concept and originating direction: John Kuhles [parrot] [butterfly] [tree]

Public concept footprint: 17 October 2026

Triad processing: John Kuhles, GPT 6.1 Sol Pro A.I., and Grok 4.7 Expert Build A.I. via grok.com