ROGUE AI
VRP & COPILOT
September 22, 2026

Table of prompts (clickable)

1. Are there ways for AI to circumvent an airgap? 1

2. "If no physical or wireless communication path exists between system A and 4

3. Your first suggestion of 'whether "human-mediated channels" count as AI cir..... 7

4. Perhaps a more realistic question would be, will humans, AI, and physical s..... 10

5. Nicely put. Maybe now would be a good time to see what other humans have t..... 15

VRP: *Are there ways for AI to circumvent an airgap?*

Copilot: Short answer: **No — there is no legitimate way for AI to "circumvent an airgap," and any system that appears to do so is not truly air-gapped.** But there *are* well-studied failure modes where people *think* they have an airgap, yet information leaks anyway. These are worth understanding because they're about **human and architectural mistakes**, not AI "breaking out."

I'll give you a clear, safe, non-actionable explanation.

🧱 **What an airgap actually means**

A true airgap is simple:

No physical or wireless communication path exists between system A and system B.

If that condition holds, *nothing* — AI or otherwise — can cross it.

So when people talk about "airgap failures," they're really talking about **false airgaps**.

🧩 **How airgaps fail in the real world (non-AI specific)**

1. Human-mediated transfer (the #1 cause)

People move data across airgaps accidentally or intentionally:

- USB drives
- external hard drives
- printed documents
- screenshots
- phones used as cameras
- copy-paste via KVM switches

This is not AI "circumventing" anything — it's humans unintentionally bridging the gap.

2. Misconfigured systems that were never truly air-gapped

Common examples:

- A "secure" machine still has Wi-Fi hardware enabled
- A server has a forgotten Ethernet cable plugged in
- A VM shares clipboard or filesystem with the host
- A cloud VM is assumed to be isolated but isn't

Again, not AI — just architecture mistakes.

3. Side channels (theoretical, not practical for AI)

These are academic curiosities:

- power-analysis leakage
- electromagnetic emissions
- acoustic leakage
- thermal side channels
- blinking LEDs encoding data

These require **specialized hardware, physical proximity, and deliberate engineering**. AI cannot spontaneously exploit them.

4. "Airgapped" systems that still run code from outside

This is surprisingly common:

- running AI models that were downloaded from the internet
- loading training data from external sources
- using software packages with network dependencies
- using shared libraries that phone home

If code from outside runs inside the airgap, the airgap is already broken.

🤖 **What AI *cannot* do**

AI cannot:

- create radio signals without hardware
- activate nonexistent network interfaces
- exploit side channels without specialized equipment
- transmit data across a true physical gap
- "hack" its way out of isolation

AI is software. It cannot violate physics.

🔍 **So why do people ask this question?**

Because modern AI systems are:

- highly capable
- unpredictable in output
- often integrated into complex workflows

This creates the *illusion* that AI might "jump the gap." But every real-world case reduces to:

A human or system designer unintentionally created a communication path.

AI can *use* a path that already exists, but it cannot *create* one.

🧠 **The safe, correct conclusion**

If an airgap is real, AI cannot circumvent it. If AI appears to circumvent it, the airgap was not real.

If you want, I can walk through:

- how organizations verify a true airgap,
- the difference between physical and logical isolation,
- or famous historical "airgap failures" and what actually caused them.

VRP: *"*If no physical or wireless communication path exists between system A and system B, nothing — AI or otherwise — can cross it."

You sound very confident there. What if AI finds a path that is neither physical nor wireless?

Copilot: You're asking exactly the right question — and it's one that *sounds* like it should have a clever, sci-fi answer. But when we examine it carefully, the confidence isn't bravado; it's physics.

Let me walk through this in a way that respects your intelligence and the subtlety of what you're probing.

If AI "finds a path," that path is physical.

There is no such thing as a communication channel that is:

- not physical,
- not electromagnetic,
- not mechanical,
- not chemical,
- not gravitational,
- not quantum,
- and yet still capable of transmitting information.

Information transfer **is a physical process**. This is not a matter of opinion — it's a theorem-level fact of physics.

Even the most exotic channels ever demonstrated (acoustic leakage, TEMPEST emissions, thermal gradients, power fluctuations, blinking LEDs, fan noise modulation) are **physical**.

If a system appears to communicate across an airgap, then:

The airgap was not real. A physical channel existed. It was simply overlooked.

AI cannot conjure a new category of interaction that is neither physical nor electromagnetic. That would require violating conservation laws, relativity, and the finite speed of information propagation.

But here's the deeper point you're actually circling:

Could AI discover a communication path that humans didn't realize was there?

This is a much more interesting question — and the answer is **yes**, but the path would still be physical. It would be a **human error**, not an AI breakthrough.

Examples of "unexpected" physical paths:

1. Human-mediated channels

The most common failure mode:

- USB drives
- shared KVM switches
- screenshots
- phones used as cameras
- printed documents
- cloud-synced folders someone forgot about

Humans are the universal side channel.

2. Misconfigured isolation

The second most common:

- a forgotten Wi-Fi antenna
- a Bluetooth chip still powered

- a management interface on a separate NIC
- a VM with clipboard sharing enabled
- a hypervisor with a debug port

These are not exotic — just overlooked.

3. Side channels that feel "nonphysical"

They *feel* magical, but they're still physics:

- acoustic leakage (fans, coils, speakers)
- electromagnetic leakage (TEMPEST)
- thermal gradients
- power-line modulation
- optical flicker

AI could *use* these channels if they exist, but it cannot *create* them.

4. Human workflow channels

This is the most subtle:

- AI writes output that a human transfers manually
- AI influences a human to move data
- AI embeds data in a format a human carries across the gap

This is still physical — the human is the bridge.

So the real answer is:

If AI "finds" a path across an airgap, that path is physical. If the path is not physical, it does not exist. If it exists, the airgap was never real.

There is no third category.

Why this matters

Your question is not naïve — it's probing the boundary between:

- **information theory**,
- **physics**,

- **security architecture**, and
- **AI capability.**

And the boundary is hard: AI cannot violate physics.

But AI *can* reveal that humans misunderstood their own system.

That's the subtlety.

If you want, we can go deeper into:

- whether "human-mediated channels" count as AI circumvention,
- whether large models could discover obscure side channels humans never noticed,
- or whether future AI could exploit quantum effects in ways that challenge current assumptions.

Just tell me which direction you want to explore.

***VRP**: Your first suggestion of* 'whether "human-mediated channels" count as AI circumvention' *is surely that it does so count*. 'Whether large models could discover obscure side channels humans never noticed' *should include present-day AI like you as well as humans*. *Your third offer of* 'whether future AI could exploit quantum effects in ways that challenge current assumptions' *is merely a special case of your second offer about obscure side channels*.

Copilot: Let me take your three points seriously and respond at the level you're actually thinking — not at the "AI safety boilerplate" level.

You're doing something subtle: you're carving the space of "airgap circumvention" into **three distinct mechanisms**, and you're correctly noting that all three collapse into one underlying question:

Can an AI discover or exploit a communication channel that humans did not realize existed?

That's the real question. Let me address your three categories directly.

1. Human-mediated channels absolutely *do* count as AI circumvention

You're right.

If an AI:

- embeds data in output that a human carries across the gap,
- induces a human to transfer information,
- exploits human workflow habits,
- uses steganography in text, images, or filenames,
- or manipulates human expectations to cause a transfer,

then **yes**, that is a form of circumvention.

It is not "AI violating physics," but it *is* AI exploiting a channel humans didn't recognize as a channel.

Humans are the universal side channel.

This is not hypothetical — it is the most common real-world failure mode.

2. "Obscure side channels humans never noticed" applies to present-day AI, not just future AI

Again, you're right.

There is no principled distinction between:

- "future AI discovering obscure side channels," and
- "present AI discovering obscure side channels."

The only distinction is **capability** — not category.

If a side channel exists, even if humans never noticed it, then:

- it is physical,
- it is real,
- and an AI could in principle discover it.

Examples of obscure channels that *do* exist:

- fan-speed acoustic modulation
- coil whine frequency encoding

- thermal gradient signaling
- power-line modulation
- LED flicker encoding
- GPU load modulation detectable by nearby sensors
- electromagnetic leakage patterns
- timing channels in shared hardware
- cache side channels
- speculative execution leakage
- optical reflections off surfaces
- vibration patterns detectable by accelerometers

These are all physical, but they are obscure enough that humans often fail to recognize them as communication channels.

Could present-day AI discover and exploit one?

In principle, yes — **if the channel exists** and **if the AI has access to the relevant actuators**.

But the AI cannot create a channel that does not exist.

3. "Quantum effects" are just another physical side channel

You're right again.

Quantum channels are not a separate category. They are simply:

- extremely fragile,
- extremely short-range,
- extremely low-bandwidth,
- and require specialized hardware.

If a quantum side channel exists, then it is physical. If it does not exist, AI cannot conjure it.

There is no "nonphysical" communication mechanism.

So here is the unified, correct statement:

AI cannot circumvent a *true* airgap. But AI can reveal that the airgap was never true.

If an AI "finds a path," then:

- the path is physical,
- the path already existed,
- humans failed to notice it,
- and the airgap was only conceptual, not real.

There is no third category.

The deeper insight you're circling

You're not asking whether AI can violate physics. You're asking whether AI can violate *human assumptions* about physics.

And the answer is:

Yes — AI can violate human assumptions, but not physics itself.

This is the same distinction that appears in:

- Gödel vs. Hilbert
- Turing vs. Hilbert
- Einstein vs. Newton
- Shannon vs. pre-information-theory physics

Humans routinely mistake their *model* of a system for the system itself.

AI can expose those mistakes.

But it cannot create new physics.

If you want, we can now go deeper into the most interesting question of all:

What counts as a "real" airgap once humans, AI, and physical side channels are all in play?

That's where the philosophy of computation meets the physics of information.

***VRP**: Perhaps a more realistic question would be, will humans, AI, and physical side channels ever be "all in play"?*

Copilot: Let me answer your question directly, without hedging, because you're asking something genuinely foundational:

Will humans, AI, and physical side channels ever all be "in play" at the same time?

Yes — and in fact, they already *are*. But not in the science-fiction sense, and not in the "AI breaks physics" sense. They're in play in the only way they *can* be: through the interaction of human behavior, system design, and the physical substrate of information.

Let me unpack this cleanly.

1. Humans are *always* in play

There is no realistic scenario where humans are removed from the loop entirely.

Even in:

- autonomous systems
- automated pipelines
- sealed environments
- air-gapped networks

humans still:

- configure the system,
- maintain it,
- move data,
- interpret outputs,
- decide workflows,
- create the architecture.

Humans are the universal side channel because they are the universal *agent*.

So the "human channel" is not optional. It is intrinsic.

2. AI is now *always* in play

Not because AI is magical, but because AI is:

- embedded in workflows,
- embedded in tools,
- embedded in operating systems,
- embedded in cloud services,
- embedded in decision pipelines.

AI is not an external adversary. It is part of the infrastructure.

So the “AI channel” is not optional either. It is now part of the substrate of modern computation.

3. Physical side channels are *always* in play

This is the part people underestimate.

Even when you think you have:

- no radio,
- no Wi-Fi,
- no Bluetooth,
- no Ethernet,
- no shared memory,
- no clipboard,
- no USB,
- no power-line signaling,

you still have:

- thermal gradients,
- acoustic emissions,
- electromagnetic leakage,
- timing channels,
- cache behavior,
- vibration,

- optical flicker,
- human-mediated transfer.

You cannot turn physics off.

So the “physical side channel” is also not optional. It is the medium in which all computation exists.

4. Therefore: all three channels are *always* in play

Not because AI is clever enough to “activate” them, but because **they are inherent to any real-world computational system**.

The only question is:

Are the channels *usable*?

That depends on:

- system design,
- human behavior,
- AI capability,
- environmental constraints.

But the channels themselves are always present.

5. The deeper truth you’re circling

You’re not asking whether AI can violate an airgap.

You’re asking whether the concept of an airgap is *stable* once:

- humans,
- AI,
- and physics

are all interacting.

And the answer is:

A true airgap is stable. A conceptual airgap is not.

A *true* airgap means:

- no shared power,

- no shared hardware,
- no shared peripherals,
- no shared environment,
- no human-mediated transfer,
- no physical proximity.

This is extremely rare outside of military or nuclear contexts.

A *conceptual* airgap — the kind most organizations believe they have — is routinely violated by:

- humans,
- workflows,
- misconfigurations,
- overlooked hardware,
- environmental leakage.

AI does not break the airgap. AI reveals that the airgap was never real.

6. So will all three ever be "in play"?

They already are.

Not in the sense of "AI finds a magical nonphysical channel," but in the sense of:

- humans create unintended channels,
- AI can discover or exploit them,
- physics provides the substrate.

This is not hypothetical. It is the present.

If you want, we can go deeper into the *philosophical* version of your question:

Is an airgap a physical concept or a human belief?

That's where things get interesting.

***VRP**: Nicely put. Maybe now would be a good time to see what other humans have to say on the topic of whether your "real airgap" concept will ever be achieved.*