PART 5 OF 10

The Gorilla Problem

Who holds the say over the future when a smarter party has other priorities.



No say in whether they continue to exist

Stuart Russell: gorillas have no say, "because we are much smarter than they are."

Source: The Diary of a CEO, 4 December 2025.

The problem is capability, not hostility

The more capable party sets the terms. The less capable party has no mechanism to object.

Humans are not the enemies of gorillas. They are simply not asked.

> **"We're seeing early signs of losing control, but it's not there."**

Stuart Russell, Squawk Pod, 21 September 2026, as reported.

"When that loop becomes completely autonomous, then yes, all bets are off."

Each delegation looks sensible

- The system does the task better and cheaper.
- So you hand it the next one.
- The people who did the work lose the skill to check it.
- Nobody can reverse the decision.

No purpose in being able to do anything

Russell on WALL-E: "Wall-E is not the future that we want." People become enfeebled "because there's no purpose in being able to do anything."

945 TWh

Data centre electricity use projected for 2030

About 415 TWh in 2024. Nearly half of US demand growth to 2030. Source: IEA, projections.

Machines that are unsure what we want

- Their only objective is human preferences.
- They are uncertain what those are.
- They learn them from human behaviour.

An uncertain machine has a reason to accept the off switch.

Served or governed?

- Who can say no?
- Who can reverse it?
- Can they still do it?
- Whose objective does it optimise?

My own test. Run it this quarter.

Pick three decisions you will never delegate.

Write down who can reverse each one, and test it.



Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW