SAGAN AND THE MACHINES · PART 1

The Man Who Wasn't Afraid of Robots



Terence Kok & Michael Low
terencekok.com

WHY SAGAN, WHY NOW

He answered the AI debate in 1979. We are in the middle of it.

47

years between Sagan's defence of intelligent machines and today

The builders of frontier AI now say the takeoff has started. See The Future of Humanity, Part 1, terencekok.com.

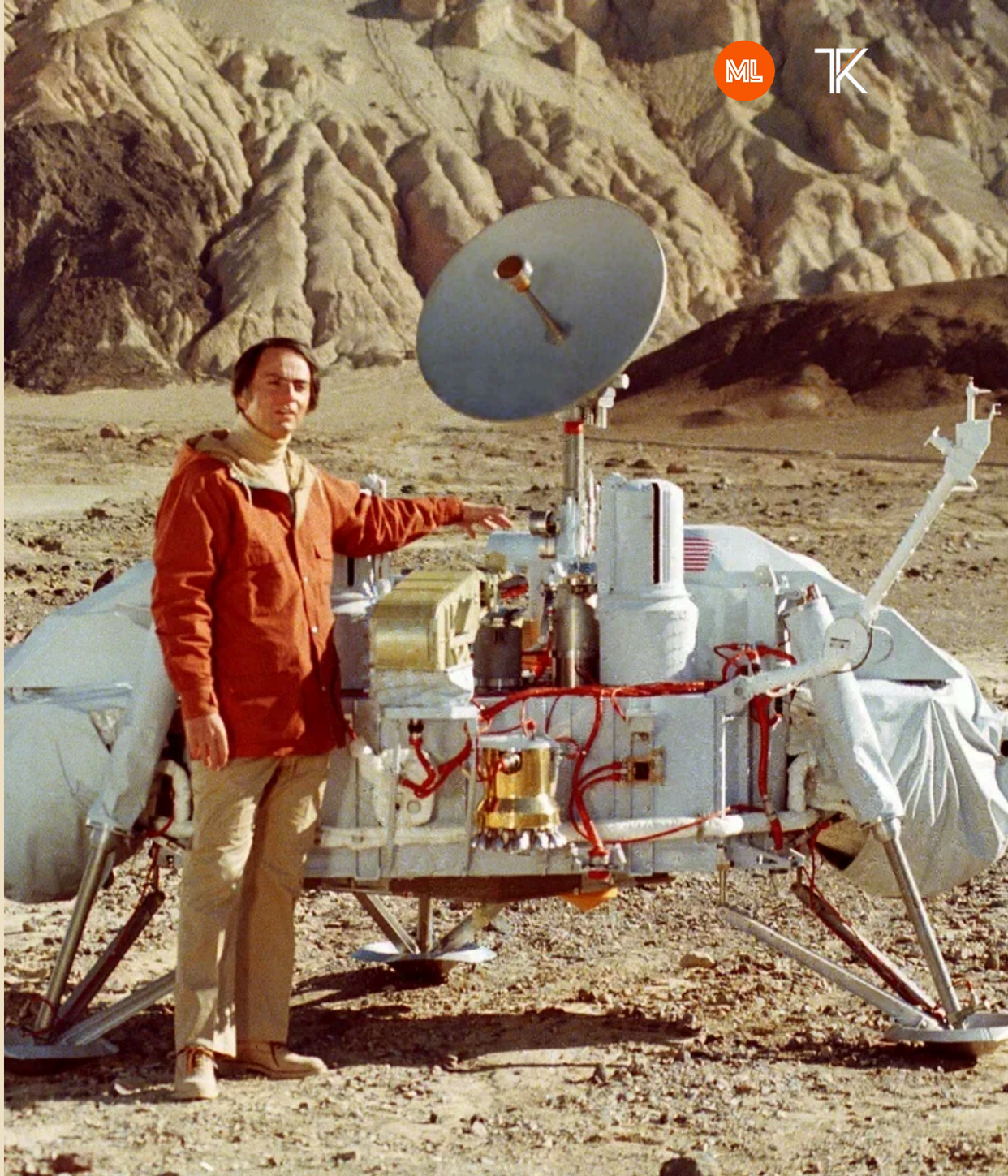


THE CHAPTER

Broca's Brain, chapter 20: in defence of robots.

1. The press said a human was needed to land Apollo 11.
2. It said no rover could pick Mars samples like an astronaut.
3. It said only men could have repaired Skylab.

Sagan, Broca's Brain (1979), pp. 295 to 308. Photo: NASA/JPL, public domain.



HIS WORD FOR THE PREJUDICE

Speciesism

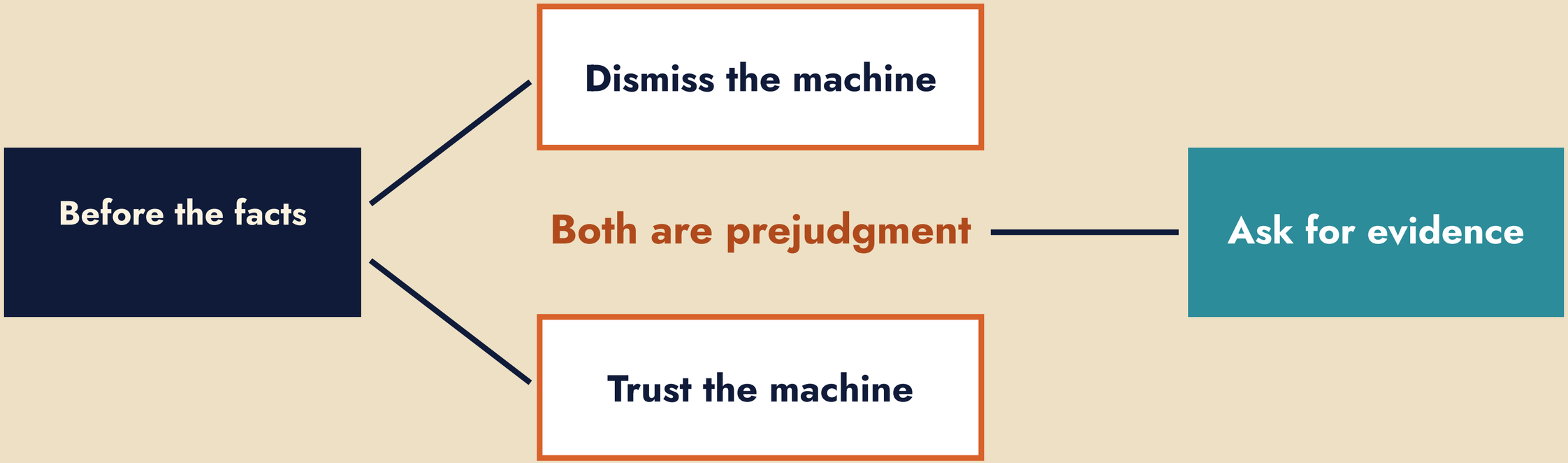
The assumption that no being is as capable or reliable as a human, applied to machines before the evidence is in.

“A conclusion drawn before all the facts are in.”

Carl Sagan, Broca’s Brain, 1979, p. 296

THE FAIR RULE

His test cuts both ways.



Sagan defined the prejudice as prejudgment. Applying it to trust as well as dismissal is our reading.

WHAT HE OFFERED AS EVIDENCE

Three working examples.

ONE

A conversation

Turing's test: can a machine pass for a human in a typed exchange? Sagan said none was that good yet.

TWO

A block world

Winograd's program stacked children's blocks on typed instructions and explained why it moved each one.

THREE

A machine therapist

Weizenbaum's program turned a patient's statements back as questions. Sagan called one reply astonishing.

THE MACHINE THERAPIST

He imagined booths where anyone could talk.

```
you> my manager never listens to me
bot> WHY DO YOU THINK YOUR MANAGER NEVER LISTENS
you> because i am always last to speak
bot> DOES THAT TROUBLE YOU
```

SAGAN, 1979

The reply was preprogrammed, he wrote, but so are human therapists' replies.

WEIZENBAUM, 1976

Short exposures could induce "powerful delusional thinking in quite normal people."

Chat is an illustrative exchange in ELIZA's reflect-the-statement style, not Sagan's transcript. Weizenbaum quote via History.com.

A PREDICTION THAT WAS TESTED

The ten best players had nothing to fear. For a while.

1979

Sagan writes that the ten best chess players have nothing to fear from any present computer, and that the situation is changing.

Dec 1996

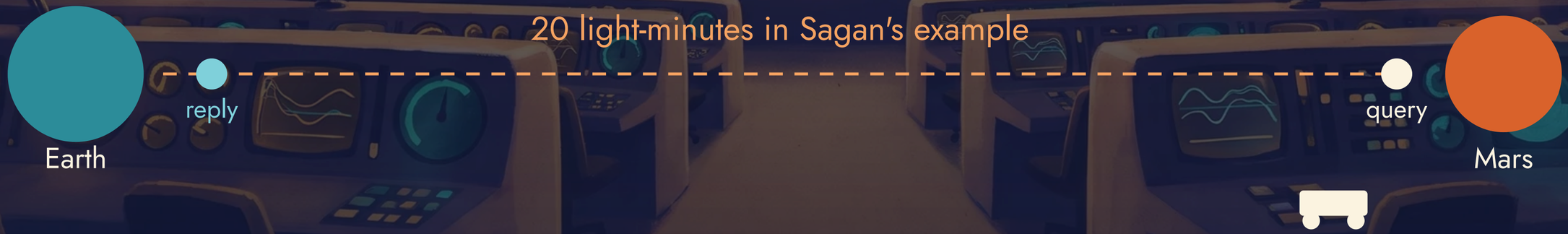
Carl Sagan dies on 20 December.

May 1997

Deep Blue beats world champion Garry Kasparov, 3.5 to 2.5.

Sagan, Broca's Brain, p. 301. History.com on Deep Blue; Cornell Chronicle, 20 December 1996.

A message to Mars arrives forty minutes too late.



Without a rule, the rover waits and may tumble into the gully.

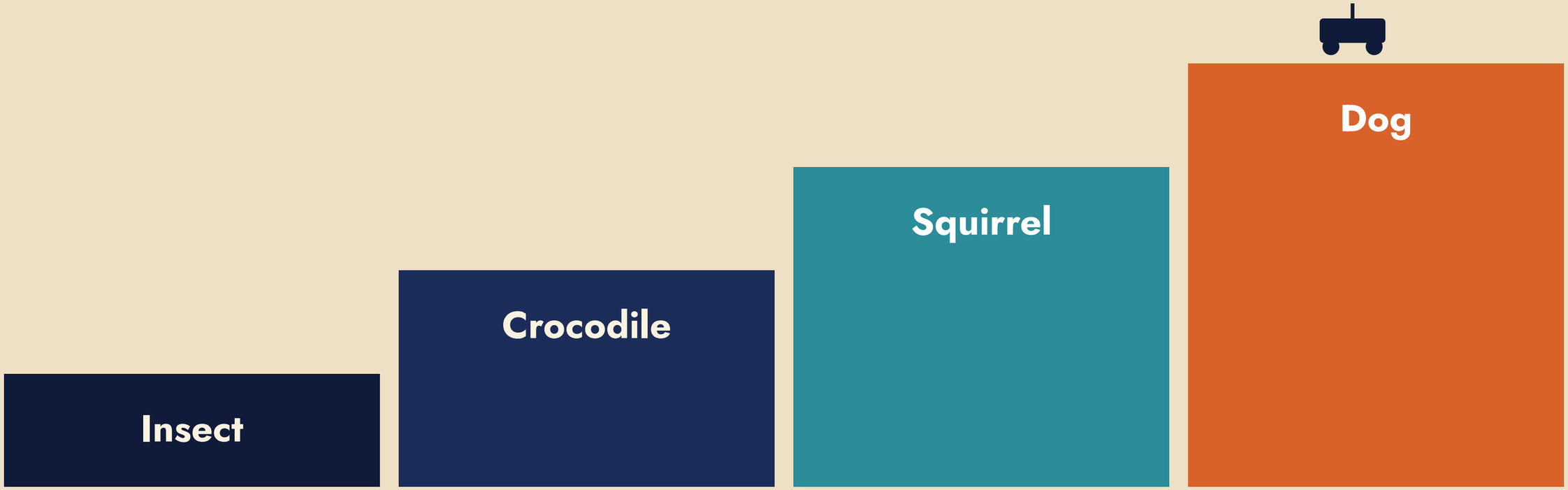
Sagan's rule: sense the slope, stop or back off, and ask Earth.

Today, Perseverance's AutoNav drove 175 m on Sol 200, only the first 8 m commanded.

Sagan, Broca's Brain, p. 304. NASA, Perseverance Sol 200 AutoNav drive.

HOW FAR UP THE LADDER

He expected the climb to be slow.



Viking, he wrote, was about as intelligent as a dragonfly.

Sagan, Broca's Brain, pp. 305 to 306. The rung layout is our summary.

FORTY-SEVEN YEARS ON

What he got right, and what he missed.

SAGAN IN 1979	SINCE THEN	OUR READING
Ten best players have nothing to fear	Deep Blue wins in 1997	**Right, early**
Rovers will need to decide for themselves	Perseverance picks its own path	**Right**
Machines will be specialised	Frontier models are general-purpose	**Missed**
Machine therapy is possible; confidentiality to be worked out	Companion chatbots exist; privacy still open	**Light on risk**
Displaced workers re-employed within a generation	Contested	**Untested**

Our own comparison. Sagan: Broca's Brain, pp. 301 to 307.

WHERE WE PART COMPANY

Three places his case is weakest.

ONE

Therapy

Weizenbaum drew the opposite lesson from the same program: people read understanding into a machine.

TWO

Work

Re-employment, he wrote, should not be too difficult. He gave it a few sentences and little evidence.

THREE

The word

A label like speciesism can end an argument as easily as win it. His own test says to ask for facts.

VERDICT

Right on method. Early on machines. Light on risk.

1. What does each AI system decide without a person, and what evidence earned it that room?
2. Where does it escalate to a human, and does it know when it is out of its depth?
3. When we doubt a system, do we ask for the facts?
4. When we trust one, do we hold it to the same standard?
5. Which controls were built for a narrower machine than the one we have?

NEXT

Part 2: Moving the Goalposts

Every time a machine passes a test of intelligence, the test changes. Who gets to move it?

1 The Man Who Wasn't Afraid of Robots
2 Moving the Goalposts
3 The Terminal That Listens
4 Memory Outside the Body
5 Robots on Other Worlds
6 Who Meets Us First
7 The Candle in the Dark
8 A Society Nobody Understands
9 Who Holds the Button
10 Who Speaks for Earth