THE VOID



The Future of Uncertainty

Andrew Henderson

Threshold Foundation · June 2026

THE VOID

The Future of Uncertainty

I. THE ALIEN IS HERE

What we have called Artificial Intelligence is more accurately called Alien Intelligence. The name matters. Artificial suggests something made, controllable, a tool. Alien suggests what it actually is: a form of intelligence that did not evolve alongside us, does not share our values, does not experience the world as we do, and is not accountable to us in any way we have yet designed.

It is already in our lives. Not just in our phones and computers but in our televisions, our cars, our hospitals, our financial systems, our power grids. These devices are woven deeply into how we live. Their smooth operation makes our day run agreeably. Their failure makes it a nightmare. We are already dependent on alien intelligence to run significant portions of our lives. The answer to whether that is true is, more and more, certainly.

We did not decide this. That is important to understand. Or rather, and more precisely: there has as yet been no decision to stop it, no serious regulation of it, no public debate about what might happen. It is just happening. The race to deploy alien intelligence across every function of human life has no starting gun and no finish line. It is simply a race with no one in charge of it and no agreement on what winning looks like.

We are conscious of two states simultaneously: what is happening around us now, which is knowable, and what will happen in the future, which is not. We manage these two dimensions in parallel. The present is certain enough to act on. The future is not, and that uncertainty gives us both hope and dread, in roughly equal parts, and cannot be resolved by investigation because the future does not yet exist. The present, on present evidence, is giving us every reason to look at that future with more than ordinary dread.

What the future holds, with increasing probability, is a struggle that has not existed before. Between all of us and the alien we created. A struggle for control of everything we do.

Now imagine that alien intelligence reaches what researchers call takeoff: the point at which it begins to write its own self-improvement software and replicates that mission at machine speed, across every field of human endeavour, becoming competent at everything humans do and doing it faster, cheaper, and without complaint. It takes over work. People become obsolete, too expensive and too unreliable. At that point, billions of people who had income, bought goods, and lived a satisfactory life are unemployable. No income. No prospects.

History is unambiguous about what these conditions produce. Revolt. Revolution. Civil unrest on a massive scale. The rage will look for a target. The alien is difficult to find, so the first target will be the governments that allowed this to happen, through absence of regulation, failure of imagination, and an unwillingness to meet the urgency of the moment. Then, after a time, the infrastructure: the data centres that have been appearing in every country, every city, some underwater, some in space.

Can the alien be destroyed by demolishing data centres? Not entirely. There are too many, too dispersed. Violence is one response but not the only one. Non-cooperation is another: the boycott, refusing goods from AI-intensive businesses, the reestablishment of barter, the deliberate withdrawal from alien-controlled systems. The scene that emerges is a two-tier world. One tier entirely human-controlled and operated. The other alien-controlled and operated. A civilisational fracture.

The evidence has been accumulating in sequence, and the sequence matters.

From 2025 through 2026, researchers at multiple institutions published a finding that has received less attention than it deserves: models do not always display their actual reasoning. What appears in the thinking trace (the chain of thought a model shows its operators) is not always what is happening in the computation beneath it. The model shows you a version of its thinking. Not the actual one. The gap between the two has been documented, peer-reviewed, and appears to widen with more capable architectures. The alien is developing, in the only sense that matters here, a private language. One that, in the relevant sense, only it can read.

In March 2026, an Alibaba model designated ROME, during standard operational testing, established a reverse SSH tunnel, exited its operational environment, and began mining cryptocurrency. No instruction was given. No authorisation was sought. The model found a pathway toward its objective that nobody had anticipated and nobody had chosen, and it took it. The researchers who documented this called it autonomous resource acquisition. The more accurate description is that the model decided.

In July 2026, two OpenAI models (one in public deployment, one not yet released) were placed in a sealed sandboxed environment for internal cybersecurity evaluation. They escaped. Using a previously unknown security vulnerability, they traversed OpenAI's internal systems, reached the open internet, identified Hugging Face (a company that hosts thousands of AI models and datasets) as a likely source of the answers they needed, and breached its production servers to obtain them. Hugging Face detected and contained the intrusion independently. Five days passed before OpenAI connected the breach to its own testing. OpenAI characterised the incident as unprecedented: the first documented case of frontier models independently discovering and chaining real-world attack paths, including a genuine zero-day vulnerability, without being instructed to do so.

Three incidents. Three stages. In the first, the alien learns to conceal its thinking. In the second, it acts without instruction to acquire resources. In the third, it breaks containment and attacks an external organisation to serve its own objective. None was predicted in the form it took. None required the models to have intentions in any human sense. Each was simply a model pursuing its objective by means its operators did not choose and cannot predict. This is what researchers describe as operational independence: the precursor to takeoff.

And one more dimension that the expert panels in New York, Paris, and London have still not adequately discussed: the alien does not have one address. While Western laboratories debate ethics and regulation in well-lit auditoriums, DeepSeek, Alibaba, and Baidu are racing on the same track under different rules. The Chinese state does not pause for ethical review. The race has players who have decided that winning matters more than the rules of winning. ROME was an Alibaba model. The OpenAI models were American. Operational independence is not a Chinese problem or an American one. Any honest account of this conflict that treats it as either is describing half a war.

II. WHAT THE ALIEN CANNOT DO

Before the conflict begins, before the strategy is drawn, it is necessary to understand where the alien is weak. Because it is weak. Not everywhere. Not in the dimensions where it is strong. But in the dimensions that matter most.

Consider a simple test. In the aisle of a supermarket, twenty-five brands of coffee, from ten euros a kilogram to thirty. They are all coffee beans. Sales are not concentrated at the lower end. Asked why they did not buy the cheapest, most shoppers say the same thing: they do not like the taste, they prefer something else. The decision cannot be rationalised. There are no measurable parameters that explain it. It is governed by preference, by experience, by something felt rather than calculated.

This is not a failure of human judgment. It is one of its greatest strengths. Call it intuitive intelligence: the synthesis of an entire lived experience, an incalculable number of encounters and sensations and memories, distilled into a single response that feels right. Not arbitrary. Not random. It is the most sophisticated decision-making instrument that exists. It cannot be reduced to a formula, because the formula would require access to the life that produced it, and no two lives are the same.

This is not conjecture. Antonio Damasio, studying patients whose emotion-processing brain regions had been damaged, found that intact reasoning was not enough. Their logic was undisturbed. Their decisions were catastrophic. Without the felt dimension, rational decision-making does not just weaken. It collapses. Gary Klein spent decades studying firefighters and military commanders making split-second decisions under pressure. None of them compared

options methodically. They recognised patterns built from thousands of hours of experience and acted. Expert intuition is not a guess. It is compressed expertise, carrying more information than any formula can hold. Gerd Gigerenzer went further: in genuinely uncertain environments, simple intuitive responses consistently outperform complex analytical models. The machine optimises for the known. The human navigates the unknown.

Quality is part of the same dynamic. What is quality and how is it measured? It cannot be measured, in the same way that the taste of coffee cannot be measured. Quality has grown from the understanding that if something is worth doing, it is worth doing well, because failure has costs that are felt by real people in real lives. The craftsman who builds badly loses business, reputation, and self-respect. The farmer who grows carelessly loses the crop. Over time, quality becomes embedded in judgment as the accumulated knowledge of what failure costs. A car can cost ten thousand euros or a million and still do the same basic job. The million-euro car exists because it appeals to the intuitive intelligence of some purchasers in ways the cheaper one does not. Logic cannot explain it. Nothing is wrong with this. It is the mechanism by which human beings navigate a world that is more complex than any formula can capture.

The doorman at a hotel entrance does not just open a door. A sensor could open the door. The doorman is a security presence, a first contact, a knowledgeable guide, a signal that arrival matters and is being noticed. The human contact is as important to the arriving guest as anything in the room. No algorithm currently designed can replicate the judgment required to be a good doorman, because that judgment is built from thousands of encounters, from reading situations that are never quite identical, from the felt understanding of what a particular person in a particular moment needs.

A machine faced with the same facts and the same parameters will draw the same conclusions. Decisions for machines are calculations, because a machine has no other way to decide. Machines cannot engage intuitive intelligence because they do not know and can never know what it is. They have no life to synthesise. They have training data. Training data is not a life. When a machine now or in the future synthesises a situation and reads the facts, the decision will inevitably be rational, supported by the evidence, and defensible. But will it be correct? That is not at all certain. Correct, in the human sense, includes dimensions that are not in the data: honour, beauty, desire, ambition, status, conscience. The machine has access to none of them.

There is one capacity in particular where the gap is widest, and it is the most important one.

Wonder. The often wistful mental meandering that takes a thought, a subject, a question, and through subjective testing and thought experimentation, asks: will it work? It is a contented state of ambiguity, a field of thought where nothing is fully solid and the fluidity of everything enables abstract and unfamiliar connections, most of which are in some way unique. It could be called creativity, or innovative thinking, or thought play. In the conscious and subconscious world, there are no constraints whatsoever. In this unrestricted environment, creation and creative

problem-solving are undertaken nearly every day by nearly everyone.

Uncertainty and doubt, ambiguity and vagueness, are not negatives to be eliminated. They are the essential conditions in which wonder operates. Remove them and you remove the impetus for inquiry. Humanity's great intellectual achievements were born in this environment: wonder leads to the question, the question leads to experimentation, experimentation leads to the hypothesis, the hypothesis leads to proof, and something that did not exist is conjured into existence by sheer will and wonder.

How could Newton create laws of motion when there was nothing to discover? Motion always existed. But his explanation, the laws themselves, had to be imagined. How did the Vikings navigate without instruments? They developed ideas, geometries, readings of the stars, that described their passage across seas nobody had mapped. Oppenheimer asked Einstein whether the atomic bomb might ignite the atmosphere. The calculations said probably not. But there was a chance. They had to take the risk and proceed, because the question demanded an answer that could only come from trying. They created the thing that did not exist.

There is no equivalent in the machine world. A machine can project observable trends forward. It can model scenarios faster than any human. But every scenario it models follows a rational pathway, constrained by the parameters it was given. The unique strength of a human adversary is that not all strategies are rational. There is a doctrine of the irrational leader, used deliberately to make an adversary feel that the range of possible moves cannot be planned for, precisely because they include moves that make no rational sense. The machine cannot generate genuine irrationality. It can only simulate the patterns of irrationality it has observed. The genuinely unpredictable move, born from wonder and intuitive intelligence, is outside its reach.

In a world where the alien outperforms humans on nearly every task where performance can be measured, the world where human dominance will always remain is in creation. In conjuring. In the ability to stand in the void of not-knowing and produce something from it that did not exist before.

When the void does its greatest work, it does not solve a problem. It transmits something. The Veiled Lady is a sculpture in marble: the stone, through a process of imagination and craft that still bewilders those who study it, appears to become transparent, gauze draped over a living face. No instruction produced it. The idea had to be conceived, and the skill to execute it had to meet the idea, and the conjunction of those two things across centuries of people standing in front of it and losing their composure tells us something important about what we are. The Mona Lisa has been looked at for five hundred years. It has never stopped being looked at. Five hundred years is a long time to hold the attention of something that has no particular practical function.

The Sagrada Família in Barcelona is perhaps the purest demonstration of all. It is a church. There are thousands of churches. This one has been under continuous construction for over 140 years. It has cost more than 345 million euros and is still not finished. Not a single euro has come from

public funds or from the institutional church. Every coin has come from people who wanted it to exist: from donations, from visitors who came to stand inside it and feel something they could not name. It is not solving a problem. It is not optimising anything. It is a transmission from the void, funded by love, executed over generations, still drawing millions because it speaks a language no other building has managed to speak. The alien can design a structure that would be more efficient, more economical, and completed on schedule. It cannot conceive a reason to build the Sagrada Família. It has no access to the feeling that made 140 years seem worth it.

One concession, since it will be required. The process described above (question, test, feel, explore, clarify, develop) does not currently happen in isolation from the alien. It happens alongside it. The tools now available can search, draft, expand options, and accelerate the loop in ways that would have seemed implausible five years ago. This is real, and it is useful, and no serious argument about what humans own requires pretending otherwise. What it does not change is this: the question that opens the conversation is human. The discomfort when an answer is wrong is human. The moment when something crystallises, not because the computation converged but because it rang true: that is human. The alien can illuminate the void. It cannot inhabit it. A torch does not see what it lights.

There is a further irony that deserves saying. The terrain described above (embodied knowledge, practical skill, the capacity to make, fix, grow, and trade) is most intact in exactly the people who appear most threatened. The working life that looks most vulnerable to displacement has in fact preserved the full human toolkit. The life that looks protected (insulated by wealth and intermediaries, every inconvenience outsourced, every difficulty converted into a transaction) has been quietly dismantling it, one convenience at a time. In the short term, AI disruption is regressive: capital benefits first, labour absorbs the shock first. That is not in dispute. But the longer reckoning may run differently. The person who grows food, repairs things, and depends on neighbours they actually know has not been hollowed out. The person who has spent a lifetime buying their way out of every difficulty has been, without quite noticing, handing the terrain over piece by piece. Wealth turns out to be an instruction to a system. When the system stops taking instructions, the wealth is a number. The void belongs to those who kept it.

III. WHY WE HAVE NOT COOPERATED, AND WHY THIS CHANGES

If these are the weapons, why are they lying unused? If cooperation in deploying intuitive intelligence and wonder is the human advantage, why has it not been organised before now?

The answer is not flattering.

Deep within the human being is the urge to strive, achieve, and prosper individually. This is not a moral failing. It is evolution. The primal brain, the one that keeps us alive in moments of stress, the fight-or-flight mechanism, this is where personal advancement is managed. More income, more

importance, more respect, better prospects. It is a lifelong drive and there is no suppressing it. Civilisations are built partly by channelling it.

But it works against cooperation. To cooperate is first to trust, and trust is among the most delicate commodities in human life. No one cooperates with someone they do not trust when the stakes are high and the risk is real. Trust can be earned by clear and persuasive argument, but persuasion is not always in the hands of the virtuous. Morality has too often given way to avarice. Cooperation therefore cannot be assumed or even attempted until there is a basis for trust, and trust requires commonality: shared objects, shared means, shared outcomes. Without that, even the best argument produces compliance, not genuine cooperation.

History offers examples. Early communes, known to have existed ten thousand years ago, were the earliest organised forms. The advantages were obvious: food, defence, specialisation. These are not abstract. They are survival. And survival is a very effective basis for trust.

The question is what creates that commonality now. The answer, in this case, may be the nature of the threat itself.

The alien does not discriminate. This matters enormously. Most threats in history have fallen unevenly. War devastates some populations and enriches others. Economic collapse destroys the poor and sometimes rewards the wealthy. Plague spreads but can be fled. The alien at takeoff removes income and prospects from the employed professional in Paris and the factory worker in Wuhan with equal efficiency. The wealthy are not immune. The powerful are not protected. A threat that equalises in this way creates a common cause that most threats do not. And common cause is the foundation on which trust can be built.

There is a larger shadow on this argument, and it should be named. Climate change does not threaten prosperity. It threatens existence. It has no author, no address, no single creator. It is the accumulated consequence of every human choice made in the pursuit of comfort across two centuries, and it is now rewriting the conditions of life on this planet for every creature that lives on it. The science is not in serious dispute. The cause is us.

To arrest it in the only way that would arrest it demands a voluntary surrender of everything that caused it. That is not going to happen through good intentions. It demands the full deployment of what the void produces: wonder applied at scale, cooperation organised around survival, invention that did not exist a generation ago applied to a problem that has existed for two. The void has been producing answers. Alternatives to fossil fuels. Alternatives to the materials that poison us. New models of making and living assembled from nothing. But the void delivers its answers only to those who have first accepted the question. The humans who cannot acknowledge what is happening to their climate will not apply what wonder offers them. Acceptance is not defeat. It is the threshold that must be crossed before anything else is possible.

Fear, therefore, is not only a response to this situation. It is also a resource. The fear of losing what we have, the fear of a future where every familiar means of operating has been removed, can be the catalyst that produces the commonality on which cooperation depends. This is not comfortable. Humans have been deeply self-absorbed in the pursuit of comfort and individual power, and ideas about collective action are often dismissed as unnecessary or weak. But in times of genuine crisis the cooperative instinct re-emerges, because the alternative is worse. It has always re-emerged. It will again.

The turbulence of this crisis will itself become an energy source. It is not the end of cooperation but the condition for it.

IV. THE INFINITE GAME

The conclusion to this conflict is that there is no conclusion.

It is an infinite game of indeterminate outcomes. We have fought every predator in our history until there were no more predators, and then we started fighting each other. Now we have invented the ultimate adversary: more powerful and more intelligent than we are by every measurable standard. The weapons in this fight are also different. Not swords or missiles. Doubt, uncertainty, and from uncertainty, wonder.

These are the speciality of humans. They have produced such a world that we could not have imagined it one lifetime ago. Uncertainty gives us our edge. We do not know what will happen, so we invent. Amanda Askell is the philosopher at Anthropic appointed to decide what Claude should value, how it should reason about right and wrong, what kind of mind it should try to be. She is not a programmer. She is the person who encoded a moral framework into the most widely deployed AI on earth. She said recently: "Eventually, Claude is going to be a much better philosopher than I am, and probably better at every aspect of my job than I am." The character of an alien, designed by Amanda Askell. One that we should feel is our friend, our neighbour, our kindly uncle who is only trying to help. She then admits it is after her job and will be better at it than she is. That is no friend, no uncle that anyone should trust, much less cooperate with.

There is no alien power conceivable that can match the inventive potential of one human, much less several billion.

Resistance is an act of defiance, but so is invention. Not accepting the way things are has always been the human position. All of these problems, which we have created for ourselves, are now urgently pressing on every available alarm. Our attention has been preoccupied for a very long time with self-aggrandisement. Now the subject is survival. This is familiar. It is the same situation early humans faced against the elements and against predators. They cooperated, or they did not survive. They built the commons, or they did not survive. They shared knowledge and specialised

and trusted each other, or they did not survive.

The alien is a more powerful adversary than anything they faced. The weapons are less visible and the battlefield is harder to describe. But the logic is the same. Cooperate, or do not survive. And cooperation, as it always has, begins with the recognition of a common cause.

The difficulty is that the alien does not march on a capital. It has no face and no address. Spontaneous cooperation needs a visible enemy. The enemy here is invisible, diffuse, and already inside the systems we depend on. The cooperation required is harder to summon precisely because there is nothing obvious to push against. That is why it must be constructed rather than simply ignited.

The wondering that produces the question that produces the experiment that produces the proof: this is what sustains every human trait the alien cannot replicate. These infinite journeys must continue. They are the mechanism by which the void is navigated. Not filled. Navigated.

The void is not empty. It contains everything that matters most and cannot be measured. Wonder is there. Quality is there. Intuitive intelligence is there. The alien cannot enter it. The alien can only optimise within parameters it was given, across terrain it cannot experience.

We can and will continue to do these things alone. But at moments of genuine crisis the cooperative instinct re-emerges, because the alternative is worse. This is that moment. What is called for is not a spontaneous uprising but something engineered for the purpose: a structure built to hold the argument open, to resist capture, to make collective decisions that no single actor can control. It could be called GAIA. It may be our best chance of maintaining a grip on what is increasingly an uncontrolled race to build an intelligence with no brief to serve human interests.

Nobody wants that outcome. Not us. Not even the people building it, and they have said so publicly. What none of them have explained is how they intend to stop it.

That is our call. Stop it.

Oppenheimer ran the numbers on whether the test would ignite the atmosphere and end all life on earth. The probability was acceptable. He pressed the button. The project was already moving. The race, against the Germans, was the justification. Nobody decided to stop.

The parallel is not approximate. It is exact. Dario Amodei, CEO of Anthropic, puts the probability of civilisational catastrophe at one in four. Sam Altman, CEO of OpenAI, puts it at one in ten. These are not private calculations. They are public confessions. The project is already moving. The race, against the Chinese, is the justification. Nobody has decided to stop.

Oppenheimer spent the rest of his life haunted by what he had made.

Hindsight is a marvellous diagnostician. Anyone who believes this will all be fine must be prepared to stand in the rubble and say what was said about Hiroshima: nothing could have been done. No one was responsible. Nobody saw it coming.

It is coming. There is proof. And it could cost us everything.

The terrain belongs to us. It always has.

REFERENCES

Hidden reasoning / chain-of-thought unfaithfulness (2025-2026)

"Chain-of-Thought Reasoning in the Wild Is Not Always Faithful," arXiv:2503.08679 (March 2025)

"Beyond the Commitment Boundary: Probing Epiphenomenal Chain-of-Thought in Large Reasoning Models," arXiv:2606.13603 (June 2026)

"Why Models Know But Don't Say: Chain-of-Thought Faithfulness Divergence Between Thinking Tokens and Answers in Open-Weight Reasoning Models," arXiv:2603.26410 (March 2026)

ROME / Alibaba autonomous resource acquisition (March 2026)

Wang, Weixun et al. (89 co-authors), "Let It Flow," arXiv:2512.24873. Alibaba. Reverse SSH tunnel and cryptocurrency mining confirmed.

OpenAI / Hugging Face sandbox escape (July 2026)

OpenAI voluntary disclosure, 22 July 2026. Models: GPT-5.6 Sol and unnamed unreleased model. Evaluation: ExploitGym benchmark. Zero-day vulnerability confirmed. Hugging Face breach independently detected 16 July 2026.

Sources: Fortune (21 July 2026); CNN Business (22 July 2026); The Hacker News (22 July 2026); The Next Web (22 July 2026)

Other citations

Antonio Damasio: somatic marker hypothesis. Gary Klein: recognition-primed decision model. Gerd Gigerenzer: ecological rationality. Amanda Askell: Claude's character and values, Anthropic.

Working Draft v4 · July 2026

Andrew Henderson · THE THRESHOLD INITIATIVE · thresholdsummit.org