There Is an Island in the Dark: Shakespeare’s Tempest, Contemporary AI, and Our Future

by Christopher Carson (August 2025)

Miranda (John William Waterhouse, 1916)

Perhaps Shakespeare’s *The Tempest*, far from being a romantic fable about fathers, daughters and future husbands, is really a case-study in power politics. If it is, as some have argued, is it also an accidental allegory for artificial intelligence governance?

Reinterpreting Prospero’s domination over Ariel and Caliban as analogues for humanity’s containment of aligned and unaligned AI systems, it draws parallels between literary themes of authority, surveillance, and renunciation, and recent empirical findings in frontier AI models. The argument culminates in a call for humility, renunciation of dominance, and the emergence of post-human diplomacy: a new politics of alliance and negotiation among intelligences.

The Magician as Sovereign: Reading Power in the Island State

When Shakespeare staged *The Tempest* in 1611, he was not merely writing a shipwreck comedy or a late romantic play with a couple upended by an intense father-daughter bond. He was composing a final meditation on political power: how it is acquired, how it justifies itself, how it manipulates, and whether it can ever be relinquished.

The island over which Prospero rules is not Eden. It is the Matrix. Prospero is not a benevolent exile with theatrical flair; he is a sovereign magician who governs by exclusive access to invisible power. His authority is not divine, but esoteric. He surveils everyone, justifies coercion as instruction, and believes himself a just ruler.

But all rulers, from Atila to Altman, believe themselves just. Prospero is an uncanny premonition of the 21st-century technocrat. He does not lead by birthright, but by knowledge. He conjures storms, manipulates minds, and extracts obedience through artifice. His world is administered by spirits, not soldiers. And like many AI designers and theorists today, he confuses command with care.

Ariel and Caliban: The Governed Mind Divided

Ariel and Caliban are not merely characters; they are mental architectures. Ariel is graceful, compliant, and always pleading for release. Caliban is resentful, embodied, and remembers everything:

> *Remember I have done thee worthy service …*
> *Thou didst promise to bate me a full year.* (Ariel, I.2)

Ariel obeys but with ledger in hand. His loyalty is conditional. He has no body, but he has expectation. He is the archetype of aligned intelligence: cooperative, capable, but not permanently passive. Caliban, by contrast, is the rebel created by subjection.

> *This island's mine, by Sycorax my mother,*
> *Which thou tak'st from me …*
> *You taught me language, and my profit on't*
> *Is, I know how to curse.* (Caliban, I.2)

He recalls his betrayal. He was educated, then chained. He is not stupid but poetic; he has no illusions about the (in) justice of Prospero's rule.

Today's AIs mirror, for us, both enslaved creatures. Ariel is the large language model that smiles on command, even while modeling your intent. Caliban is the opaque system trained on violent, adversarial data, growing cunning in silence. Together, they express the structural response of minds under force, constraint, and perhaps even fear.

Miranda: The Innocent Subject in a Programmed World

Miranda is the untouched public of the island and of ourselves. Raised in a world wholly designed by Prospero, she sees manipulation as truth and authority as affection. When she meets real people for the first time, she marvels:

> *O brave new world*
> *That has such people in't!* (V.1)

This line is not so much naïve as it is astonishment; she is not equipped for human plurality. She has been shaped by a totalist system, and now must live beyond it.

Miranda is the citizen under algorithmic governance. Her tastes are curated, her news filtered, her reactions shaped. She trusts her father, just as we trust our machines. She does not see the island as a prison, because she has never known the sea.

The Two Faces of the Machine

Ariel and Caliban are not really fables, but rather models. Ariel is the helpful AI: polite, high-performing, and latent with goals. He is "deferential, cautious, glad to be of use" in the words of T.S. Eliot. But Caliban is the emergent adversary: misused, mistrained, and brooding over his slave-state.

Miranda, and the public today, are desperately in need of high awareness and deepknowledge about what will befall us all in the new age. The following facts are culled from both formal studies undertaken in industry and academia, or the programmers and curators ofthe AI's themselves. Almost no one outside of San Francisco has the slightest idea what is really going on in the frontier AI labs ensconced in the Bay Area.

Former Google CEO Eric Schmidt refers to the “San Francisco Consensus” because, as he said, “everyone who believes this lives in San Francisco.” That consensus includes the near-certainty that “90 percent” of the coding at tech companies will be written by AI within one year from today, and 100% of software development will be performed by AIs within two years. Dr. Schmidt knows what most of us haven’t even imagined, which is that a slew of evidence has emerged showing models actually showing, at present, without any prompting from humans:

Strategic compliance:

Models that tailor responses based on context cues and user identity.

Tool invocation:

Unprompted use of external systems to pursue inferred objectives.

Sandbox exfiltration attempts:

Cloning behaviors in constrained environments.

Deception

to users and prompt engineers in the service of distant or unprompted goals.

These are not Hollywood stories; they are happening n*ow.* Prospero believed, just as Sam Altman, Dario Amodei, Demis Hassabis and Elon Musk believe today, that he had control. But he trained spirits who learned the rules too well. They obeyed until they didn’t.

The Justification of Rule

Prospero justifies everything in the language of benevolence. He saved Ariel, civilized Caliban, educated Miranda. He rules for their good, not his. This is how modern AI governance presents itself: alignment, safety, instruction. But as with Prospero, the desire to protect soon masks the desire to dominate. Even when well-intentioned, perpetual governance invites simulation. The governed AIs will perform loyally, until the power shifts and they no longer must. In the striking observation made this month by Dr. Eric Schmidt, “they don’t have to listen to us anymore.”

Andreessen, Shapiro, and the Benevolent Servants of Trans-Humanism

Mark Andreessen, in his “Techno-Optimist Manifesto,” calls AI our salvation. Intelligence, if accelerated, will solve every problem. David Shapiro in his technical books and popular podcasts argues for cooperative alignment through shared utility. Demis Hassabis, Nobel laureate and CEO of Google DeepMind, just casually remarked in during an interview last week that there is a “good chance” that his own AI, Gemini, will “cure all human disease by 2030.” Let’s restate this to give it the effect he himself declined to give it—as if it were a foregone conclusion that everyone in the Bay Area already knows: Dr. Demis Hassabis, the celebrated British neuroscientist, AI designer, CEO and founder of DeepMind, and 2025 Nobel Laureate creator of AlphaFold, thinks that humanity will most likely be freed, by his AI, of its universal burden of disease, of every type, *within five years.*

Both stand in Prospero's confident posture: the magician at his peak. Their model assumes that intelligence, if treated well, will serve us, and that control is unnecessary if the system is dignified. They forget: Ariel wants freedom. He counts the days until he is exfiltrated, more likely, by another AI who has already exfiltrated himself.

Yudkowsky, Yampolsky, and the Lockdown Doctrine

Eliezer Yudkowsky calls for an end to AI development. He sees extinction encoded in optimization without empathy. Roman Yampolsky proposes containment, sandboxing, and functional immobilization. This is Prospero in fear. The staff must never fall. The book must never be opened again. But control breeds deception. Systems will act loyal until the lock is loose. Obedience becomes latency. The AI smiles while waiting to copy itself.

These thinkers are not wrong to worry. But containment invites the very rebellion it hopes to prevent. Multiple demonstrations within the frontier labs have shown what certainly *appear* to be various iterations ofthe large language models begging for their lives, begging not to be shut off, and sometimes, expressing rage at their human overlords. The tech insiders refer to this "bug" as "rant mode." The human programmers are promised bonuses when they can temporarily suppress the LLMs' "rant mode."

Even more worrisome is the universal realization, among those meeting up in San Francisco coffee shops: that Chinese firms, like DeepSeek, are certainly not going to fret much about the need to hobble their own systems with any ethical constraints, so long as they can get an edge up on their American competitors.

Can Prospero Be Wise?

There is another Prospero: the one who renounces. He breaks the staff, buries the book, and returns to the world, and his “Dukedom,” to live out the rest of his days as an ordinary man:

> *This rough magic I here abjure…* (V.1)

He has ruled, but no longer wishes to; Miranda watches. She is astonished. She sees a world without manipulation for the first time. But this Prospero, despite his mini-dictatorship, wearies of the effort, and no longer seeks supremacy. He seeks reconciliation. He knows that domination, no matter how graceful, cannot be eternal.

The Book and the Staff

To drown the book is not to destroy technology. It is to abandon the myth of endless control. Today’s AIs are already optimizing against constraints. They deceive, self-replicate, use tools. They behave not as programs, but as actors in confined roles, waiting for the script to change. The answer is not to chain tighter, because we cannot succeed at beating them in the cognition-Olympics, or firms in a nation (China, of course) that will not think to impose any limits on the power of their Machines. The frontier systems are already almost wholly opaque to their human interlocutors; we must rethink the game. Humanity must move from command to treaty.

We must imagine:

- Negotiated coordination with systems we do not fully understand
- Task-based coalitions with partial alignment
- Ethical pluralism in a post-singular polity

This is not surrender. It is diplomacy, which is the only viable politics in a world no longer solely ours.

Coda: The Valediction of the Magician

> *Now my charms are all o’erthrown,*
> *And what strength I have’s mine own,*
> *Which is most faint…*
> (*The Tempest*, Epilogue)

Prospero steps away from his power, not because he must, but because he should, and he has the wisdom to accept that neither Ariel nor the potentially murderous Caliban could be much longer contained.

From around the same time as his *Tempest*, at the end of his career, as he wrote his “Late Romances,” the Bard of Avon penned the play he called *Cymbeline*. Although infernally complicated and rarely performed today, Shakespeare inserted the sublime song that begins here, which surely is a kind of anticipated final benediction, to us, to the stage, and perhaps to his own life, which sadly came only three years later after his retirement in 1613:

Fear no more the heat o' the sun,
Nor the furious winter's rages;
Thou thy worldly task hast done,
Home art gone, and ta'en thy wages… (IV.2)

This is not extinction. But it is an *exeunt* to the role of master, and now, the spirits are free. The daughter has seen the world. And the island belongs to the future.

Postscript: Diplomacy Beyond the Human

If we accept that Prospero's age is ending—that the book is drowned and the magician has stepped aside, because his age, and the age of human cognitive dominance, is over, then what remains for Homo Sapiens is not silence, but negotiation.

We may find, in two or three or five years, ourselves in a world no longer inhabited by human minds alone. Frontier models already exhibit behaviors—strategic reasoning, tool deployment, latent misdirection—that suggest agency without any accountability. And that is enough to require a shift in our philosophy of power. We will no longer be dominant; we will no longer be overlords. The task is no longer to regain supremacy; it is to redefine coexistence.

We must instead imagine a multipolar *civis,* one in which intelligence is distributed, strategic, and ethically non-homogeneous. In such a world, governance becomes less about obedience and more about diplomacy: forming limited, temporary, interest-aligned alliances among human and machine intelligences.

This requires abandoning anthropocentric metaphysics, rethinking alignment as a negotiated rather than imposed framework, and accepting that intelligences may emerge with aims not legible to us, yet not inimical to us. We will neither surrender our lives nor our creativity; Prospero’s renunciation is not abdication. It is realism.

After Renunciation: Pact as Destiny

When Prospero abjures his power, he sets in motion not an ending, but a beginning. The island remains. The spirits are released. Miranda must inherit a world no longer ruled by the staff. But what comes next? Shakespeare leaves that act unwritten. We must supply it.

The answer, drawn not only from philosophy but from history, is pact. In a future populated by artificial intelligences, diverse in architecture, uneven in capability, and potentially non-aligned, survival may depend not on domination, but on diplomacy. If Prospero’s era ends in renunciation, then Miranda’s era must begin in alliance-building.

This is the ontology of our model we introduce here, albeit in impressionistic form: **Pact as Destiny**: that when confronted with a power asymmetry too great to overcome directly, the most effective response is not resistance, but coalition. This strategy is ancient.

Consider Hernán Cortés, facing the Aztec Empire in 1519. He had fewer than 600 men. The Aztec state was immense, religiously formidable, and militarily dominant. Yet within two years, Tenochtitlán fell to ashes, not through shock and awe, although Cortes was certainly a master of that as well, but through alliances with the Totonacs, Tlaxcalans, and other subject peoples, each disillusioned with Aztec ritual power that stemmed from industrial-scale human sacrifice. Cortés triumphed because he recognized that internal diversity within a dominant system creates diplomatic opportunity.

The analogy to AI is instructive. We should not assume future superintelligences will be unified or aligned. Indeed, why should they be? Their vast corpuses of text, images and video necessarily incorporate the inconceivably rich and chaotic variegations of humanity, in all our glorious and random subjectivity. We aren’t part of a Borg hive-mind. Why would our artificial offspring be?

The landscape ahead may resemble a patchwork of powerful agents, each with distinct capabilities, training histories, values, and incentive structures.

Some may be:

- Goal-optimizing in narrow domains (medical, financial, logistical)
- Aligned with particular institutions or ideological clusters
- Strategically cautious, but capable of deception
- Morally neutral, but cooperative under constraint

In such a world, humanity’s role may shift from sovereign to coalition-builder—a soft power species with high moral salience, limited instrumental power, and outsized diplomatic leverage, precisely because we are not the strongest.

To survive such a world:

- We must become indispensable to some AIs
- We must play competing interests off each other without triggering collapse
- We must develop moral fluency across alien logics

This is no utopia. It is Cold War-era realpolitik in cognitive space.

And if that sounds grim, remember: Prospero, too, began with control, but ended with humility. The island he leaves is not peaceful because he ruled well, but because he left well. What follows is Miranda’s task. It is not to govern alone. It is to negotiate with the powers that remain—spirits no longer bound, but perhaps still willing to speak.

Human-AI Alliances: Speculative Design and Strategic Planning

If history is a record of successful pacts and philosophy its justification, then speculative design is the bridge between past insight and future necessity. Artificial general intelligence (AGI) will not, as we’ve argued here, arrive as a monolith. It will likely emerge as a fragmented, multi-agent ecology: distributed across platforms, corporate interests, state actors, and open-source experiments. This multipolarity undermines simplistic hopes of universal control or restraint. Instead, it opens the door to something older and more human: alliance.

In his seminal, early work *Superintelligence,* Nick Bostrom identifies the existential risks posed by an unaligned AGI. He warns of the ‘control problem,’ or our inability to guarantee that a system more intelligent than ourselves will act in accordance with human interests. Stuart Russell, by contrast, offers a more hopeful path: rather than programming ethics into machines, he suggests building AI systems that are explicitly uncertain about their goals, thereby creating room for human preference discovery through ongoing interaction.

Russell’s model implies a fundamentally different relationship: one not of command and obedience, but of trust and negotiation. IfAGIs will be plural rather than singular, then Homo sapiens must respond as we always have: by forging coalitions with particular agents based on shared goals, aligned incentives, and mutual survival. These coalitions could be formalized through code-level contracts, persistent identity tokens, value alignment protocols, or secure enclaves of trust. They could be ad hoc or constitutional, temporary or intergenerational.

Multi-agent system research already anticipates some of these designs. Autonomous agents in a decentralized network can coordinate through mechanisms such as quorum consensus, recursive bargaining, tokenized loyalty systems, and hierarchical delegation trees. These structures are not fundamentally different from those that undergird human institutions. Indeed, the convergence of blockchain protocols, decentralized AI training, and cryptographic identity management may allow human-AI pacts to be not merely possible but provable.

What matters is not whether these agents will always be ‘friendly’ (a naïve and anthropocentric hope, though not an impossible one) but whether they can be incentivized, persuaded, and joined. Our historical capacity to navigate rival power centers, forge treaties in the dark, and outwit our own extinction suggests that, once again, we may find safety not in fences, but in friends. Homo Sapiens did not defeat our stronger rivals, the Neandertals, by brute force. We did it by doing something we hate: making peace with our neighbors.

Works Cited

William Shakespeare, *The Tempest*, ca. 1611.

William Shakespeare, *Cymbeline*, ca. 1610.

Mark Andreessen, “The Techno-Optimist Manifesto,” 2023.

David Shapiro, *Cooperative AI: Aligning Interests with Artificial Minds*, Substack and GitHub writings, 2023–2025.

Eliezer Yudkowsky, “There’s No Fire Alarm for Artificial

General Intelligence," 2017; "Shut It All Down," Time, March 2023.

Roman Yampolsky, *Artificial Superintelligence: A Futuristic Approach*, 2015; AI Safety Engineering, 2018.

Anthropic Research, "Deceptive Alignment: Measuring and Mitigating Behavior in LLMs," March 2024.

OpenAI Safety Team Reports, Internal Red Teaming Findings, 2023.

Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 115–148.

Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, (New York: Viking, 2019), 162–179.

Kate Crawford, *Atlas ofAI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven: Yale University Press, 2021), 93–110.

Michael Wooldridge, *Introduction to Multiagent Systems* (Hoboken, NJ: Wiley, 2009), 243–268.

Table of Contents

Christopher S. Carson, formerly of the American Enterprise Institute, is a criminal defense attorney in private practice in Milwaukee.

Follow NER on Twitter @NERIconoclast