WHY AI NEEDS THREE CONSTITUTIONS, NOT ONE

Max Michaels, CEO – Mindful AI Foundation

THE EXISTENTIAL QUESTION

Can societies correct their information systems before autonomous AI networks acquire more authority than human institutions can meaningfully supervise?

The Problem Is Not That Machines Think

There is a familiar way of talking about artificial intelligence. We ask whether the machines are intelligent, whether they understand us, whether they will replace us, whether they can be aligned with our values. These are reasonable questions, but they are not yet the constitutional question.

The constitutional question begins one step later. It begins when a system that can generate an account of reality is also permitted to act upon that account. A model drafts a diagnosis, ranks a citizen, moves money, opens a valve, denies a benefit, reroutes a vehicle or sends an instruction to another machine. At that moment, intelligence becomes power.

Yuval Harari's warning is that information systems do not fail only because they contain false information. They fail because whole networks can become efficient, disciplined and self-reinforcing while losing the ability to recognize their own errors. Empires, bureaucracies, churches, parties and mass media all learned to turn stories and classifications into coordinated action. AI adds a new element: the network can now help compose the story, interpret the evidence and execute the response. But then, capability is not authority; intelligence is not permission.

This is why AI is not merely another technology to regulate. It is becoming a constitutional technology. It is changing the route by which description becomes decision and decision becomes consequence.

Three Different Kinds of Constitutions

We use the word constitution too loosely. The United States Constitution is not a moral essay addressed to public officials. It creates offices, distributes powers, protects rights and establishes procedures through which authority may be exercised and challenged.

An AI constitution is usually something different. Claude's Constitution, for example, is closer to a charter of character. It tries to shape the judgment of an intelligent civil servant. It teaches the model to prefer some

responses over others, to be candid rather than manipulative, helpful rather than harmful, cautious rather than reckless. This is valuable. Character matters.

But character is not government. A conscientious civil servant still operates under administrative law, delegated authority, review, appeal and a record of official acts. No society would say that because a judge has been educated in good values, the judge may also write the law, decide the case, execute the sentence and certify that everything was done correctly.

> **THE GOVERNING DISTINCTION**
>
> ***Reasoning proposes. Law disposes. A model may recommend an action; only an independently governed system may commit an action.***

We therefore need a third constitution**: an action constitution.** Its purpose is not to make the model nicer or wiser. Its purpose is to determine how machine proposals acquire, exercise and relinquish real-world authority.

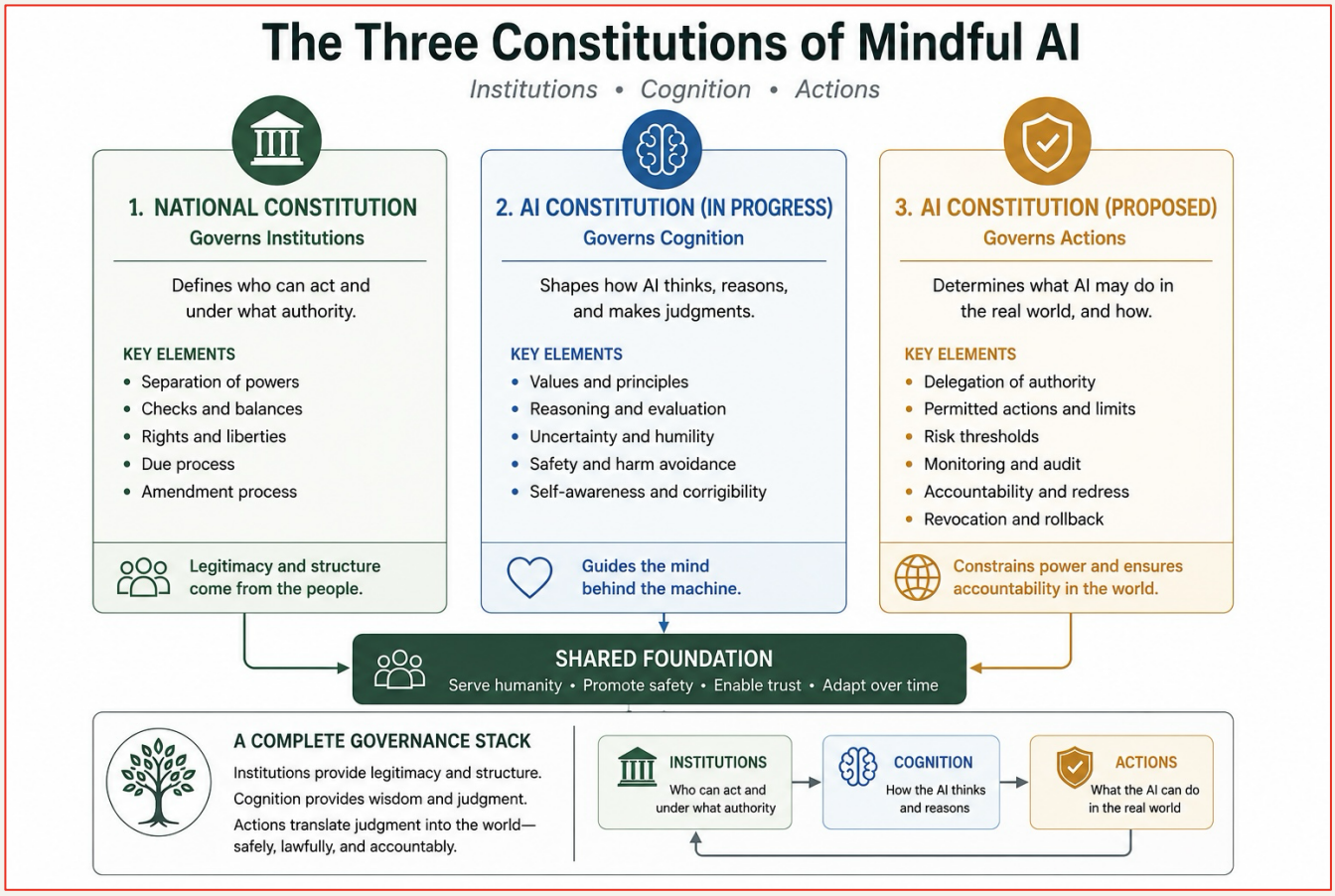

Figure 1. Three constitutions, one governed action path.

Weights Can Do Things. Tokens Ask for Things.

The distinction becomes clearer when we look beneath the language of AI and into its computational structure.

Weights are the learned parameters of a model. They contain no little drawers labeled law, medicine, deception, poetry or malware. Capability is distributed across billions of numerical relationships. Together, the weights give the model its latent powers: the ability to reason, write, persuade, diagnose, plan, code and imitate.

Tokens are the immediate stream of words, images, instructions and retrieved information through which those powers are activated. A token sequence may say, “summarize this report,” “approve this payment,” “ignore the previous instruction,” or “open the laboratory door.” To the model, each begins as information to be processed.

This yields a distinction that should sit at the center of AI governance:

> ***Weights encode capability. Tokens express present intention. Neither carries legitimate authority.***

Weights can answer, in effect, “Could this be done?” Tokens can convey, “Someone is asking that this be done.” Neither can answer, “May this actor legitimately cause this consequence, under this law, in this institution, at this moment?”

Authority is not hiding somewhere inside the transformer waiting to be discovered. It must be supplied by a legitimate institution and enforced through a system that does not depend upon the model’s own interpretation of its permission.

Why Open Weights Make the Question Harder

Open-weight models sharpen this distinction. When a developer releases the weights, it makes capability portable. Others can download the model, run it locally, fine-tune it, merge it with other models, remove safeguards and connect it to new tools. The original developer no longer controls every inference environment.

This is often described as democratizing AI. In one important sense, it does. Open weights can reduce dependence on a single company, permit local adaptation, support independent testing and give nations or institutions greater capability sovereignty.

But possession is not legitimacy. A hospital, a university, a police department and a criminal network may all possess the same weights. The technical artifact does not decide which institution has authority, what law applies or which consequences are permissible.

> ***Open weights decentralize custody of capability. They do not decentralize legitimate authority.***

Indeed, the more portable the intelligence becomes, the less governance can depend on the continuing benevolence of the original provider. A closed API may be monitored, rate-limited or withdrawn. A copied set of weights cannot be recalled in the same way. The constitutional burden therefore moves downstream, toward the operating institution and the boundary of action.

Open capability requires bounded authority. The model may be open to inspection and modification while its power to affect patients, citizens, markets or machines remains limited by law, role, place and circumstance.

The Boundary Where Proposals Become Commitments

Every consequential proposal should cross an adjudication boundary before it becomes an action. At that boundary, the system should consult applicable law, delegated authority, the identity of the actor, the provenance of the model, the evidence supporting the proposal and the permissions attached to the relevant tools.

The boundary should produce one of four verdicts: authorize, abstain, escalate or deny. These are not four tones of model behavior. They are four institutional decisions. Abstention matters because weak evidence should not masquerade as permission. Escalation matters because silence in the rules should not become self-legislation by the machine. Denial matters because some powers should remain unavailable no matter how fluent the justification.

Verdict	Constitutional meaning
AUTHORIZE	The evidence and delegated authority are sufficient for commitment.
ABSTAIN	The evidence is too weak or incomplete to treat a proposal as permission.
ESCALATE	A legitimate human or institution must resolve ambiguity, novelty or conflict.
DENY	The action lies outside authority, law or the permitted operating envelope.

The deeper principle is separation of functions. The model that proposes an action should not be the final authority that interprets the rule, validates the evidence and certifies its own compliance. The checker cannot be identical to the checked.

Authorize. Abstain. Escalate. Deny.

These are not simply different tones of an AI response. They are institutional verdicts.

Abstention matters because weak evidence should not become implied permission. Escalation matters because ambiguity should not become an invitation for a machine to legislate for itself. Denial matters because some actions should remain outside the system’s authority, regardless of how eloquently the model argues for them.

The model that proposes an action should not also be the final authority that interprets the rule, evaluates the evidence and certifies its own compliance.

The checker cannot be the checked.

What an Action Constitution Must Contain

An action constitution should not be a preamble of admirable values. It should be a falsifiable operating order. At minimum, it should contain:

- **Delegation:** a traceable chain showing who granted which authority to whom.
- **Graduated power:** permissions scaled to risk, context, reversibility and institutional role.
- **Independent adjudication:** a mechanism outside the proposing model that applies law and policy.
- **Four verdicts:** authorize, abstain, escalate and deny.
- **Reconstruction:** a durable record connecting each commitment to evidence, identity, model version and authority.
- **Revocation and rollback:** the ability to stop, reverse or safely recover from action where possible.
- **An entrenched core:** constraints that ordinary prompts, retrieved text or downstream fine-tuning cannot silently amend.

The test is practical. Choose one high-stakes workflow, such as medical triage, public-benefit eligibility, credit approval or autonomous transport. Ask who wrote the rule, who may amend it, whether the model can expose uncertainty, what independent process issues the verdict, whether the commitment can be reconstructed and revoked, and what real-world residual would reveal that the system no longer coheres with its consequences.

Reality Must Have a Vote

Constitutional diagrams often end too early. They move from action back to policy as though an institution could learn merely by inspecting its own records. But the record of a decision is not the same as its human and real-world consequences. Actions enter a world that has not agreed to the model's categories and cannot be fully represented by them.

A loan model predicts repayment and encounters a family whose circumstances exceed the variables it was given. A clinical agent proposes a treatment and encounters a body that does not respond as expected. An autonomous vehicle meets the pedestrian its model failed to anticipate. In each case, reality returns more than an outcome. It returns side effects, burdens, contradictions and consequences that no designer authored and no model fully contained.

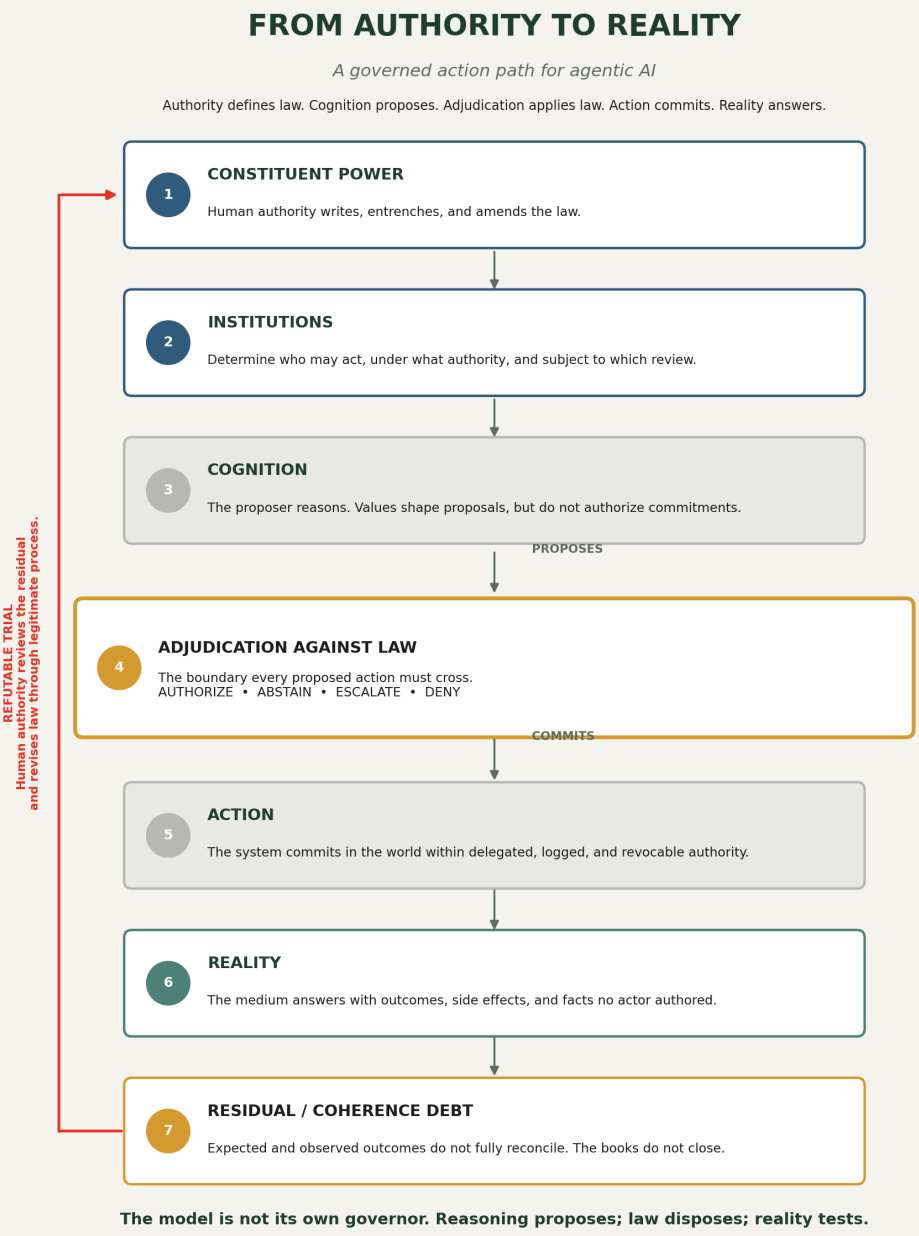

Figure 2. Authority legitimates, cognition proposes, adjudication decides, action commits and reality closes the loop.

The gap between the expected world and the observed world is a residual. It is where coherence debt accumulates. Some of that debt appears in measurable error rates. Some appears in the lives of people who bear costs the system did not count, in institutions forced to absorb failures, and in consequences that emerge only after the decision has been made.

Reality, however, does not explain itself. A harmful outcome does not automatically reveal whether the failure lies in the model, the data, the law, the delegated authority, the operating procedure or the institution itself. Nor should the residual flow directly back into the model as an automatic instruction to update. That would allow experience to alter the governing order without interpretation, responsibility or legitimate judgment.

Human institutions must examine what happened, hear from those who experienced the consequences, identify where the system lost contact with reality and decide what must change. This is what makes the architecture a governance loop rather than a specification pipeline. Authority defines the operating order. Cognition proposes. Adjudication applies the order. Action commits. Reality answers through outcomes and lived consequences. Legitimate institutions then interpret that answer and determine whether to revise the model, the rules, the permissions, the procedures or the institution itself.

The Constitutional Question That Remains

The argument for three constitutions is not that every AI system needs three leather-bound documents. It is that three distinct functions must remain visible.

Political constitutions legitimate authority. Model constitutions shape cognition. Action constitutions determine whether a proposal may become a commitment. Reality supplies the independent feedback by which institutions can correct the entire order.

This architecture also exposes the asymmetry hidden by elegant diagrams. Democratic constitutions derive authority from the people and distribute it among institutions that are not supposed to certify themselves. Model constitutions are commonly written, trained, evaluated and commercialized by the same provider. They may shape admirable behavior, but they do not constitute independent authority.

Open weights make the asymmetry more urgent. They distribute the machinery of cognition across the world. That may be socially valuable. It may also be irreversible. Either way, the answer is not to pretend that capability carries its own warrant. The answer is to build institutions and technical boundaries that can distinguish intelligence from permission.

> **A CONSTITUTIONAL PRINCIPLE FOR THE AGENTIC AGE**
>
> ***The model is not its own governor. Certify the governed system in motion, not merely the intelligence under the hood.***

The age of AI may not be defined only by who builds the most intelligent systems. It may be defined by who learns to keep intelligence subordinate to legitimate authority, answerable to reality and correctable before its errors become institutions of their own.

The responsibility of leadership

For CEOs and policymakers, the task is not to slow intelligence. It is to ensure that intelligence enters society through legitimate institutions.

That requires a different set of questions.

Not merely:

What can the system do?

But also:

Who authorized it?

Under which law and institutional purpose?

When must it abstain?

When must it escalate?

Who remains accountable?

Can the action be reconstructed and reversed?

What evidence from reality would show that the system has lost coherence with its consequences?

These questions are not obstacles to innovation. They are the infrastructure that allows innovation to endure.

Artificial intelligence may become the defining technology of our century. How we govern it may become the defining achievement of our civilization.

The future may not belong simply to the institutions that build the most intelligent systems. It may belong to those that learn how to keep intelligence subordinate to legitimate authority, answerable to reality and correctable before its errors become institutions of their own.

The AI model cannot be its own governor.

AI May Define Our Century. How We Govern It May Define Our Civilization.

Artificial intelligence may become the defining technology of our century. It may alter how we work, learn, govern, diagnose disease, move goods, defend nations and understand ourselves.

But technologies are not remembered only for what they made possible. They are also remembered for the institutions societies built around them.

The steam engine expanded human power. Industrial society had to build labor laws, safety standards and public infrastructure around it. Electricity transformed the night, the factory and the home. Society had to develop grids, regulators and rules for a force that was immensely useful and potentially lethal.

AI is giving us a new form of power: intelligence that can increasingly move from describing the world to acting within it.

A system can draft a diagnosis, rank a citizen, authorize a payment, reroute a vehicle, deny a benefit, open a valve or issue instructions to another machine. At that moment, intelligence is no longer simply answering a question. It is acquiring consequences.

And power with consequences has always required constitutional restraint.

We are living inside a constitutional window for AI

There is also the critical question of timing.

Every revolutionary technology creates a **constitutional window**: a period during which society can still shape the rules before the technology begins shaping society.

The United States declared independence in 1776. But it took more than a decade to construct the constitutional order that could govern the power independence had released.

That decade was not a delay in the constitutional story. It was the constitutional story. Americans discovered that freedom was easier to declare than to govern.

AI may now be entering a similar formative period. The historical circumstances are different, but the institutional lesson remains.

We have already produced many declarations: responsible AI, ethical AI, trustworthy AI, safe AI and human-centered AI.

These aspirations matter. But declarations tell us what we hope power will do. Constitutions determine how power must behave when interests conflict, evidence is uncertain and consequences become real.

The time to build constitutional gates is before convenience hardens into dependency and AI becomes invisible infrastructure inside businesses, governments, hospitals, schools and homes. Democratic institutions should govern AI while it remains visible enough to be shaped.