AIGovSpec

THE ESSAY

AI Governance From First Principles

One definition—and the questions it teaches us to ask.

BY SAMUEL ABINSINGUZA

When someone asks what AI governance is, we often answer with a list: laws, policies, standards, audits, risk frameworks, oversight bodies.

Those things matter. But a list of instruments does not explain the problem they are meant to solve. It also makes the field sound distant from anyone who does not write laws, build models, or attend policy meetings.

Consider two decisions.

When an AI company is deciding whether to release a more capable model. It needs to know what the model can do and whether its safeguards are sufficient. Its decision also depends on competitive pressure, who can challenge management, and who will bear the consequences of failure.

This pressure also applies to when a health ministry is deciding whether to buy an AI tool that helps clinicians identify patients needing urgent care. It needs to know whether the tool works for its population. It also needs to know what happens to patient data, whether clinicians can challenge a recommendation, and what alternatives remain if the supplier changes its terms.

Both decisions concern the technology. Both also concern the arrangements through which that technology affects people.

NIST's AI Risk Management Framework calls AI systems "inherently socio-technical."[1] In plain language, their effects depend on interactions between technical systems, people, and

the environments in which they operate.

The company's decision process and the ministry's working practices belong in the analysis alongside the model.

That is the starting point for the framing I propose:

AI governance is the field of inquiry and practice concerned with how societies should shape the incentives, institutions, infrastructures and safeguards around AI to share its benefits broadly, prevent and address harm, and maintain the capacity to change course as today's choices become tomorrow's conditions.

Each part of that sentence gives us somewhere to look.

It also states an ambition. Governance can exclude people, protect incumbents, or entrench power. Shared benefit and the ability to change course are standards against which to judge governance, rather than qualities every existing arrangement already possesses.

Let us unpack the sentence—and turn its parts into questions we can use.

"Inquiry and practice": learning what works, and making it count

We do not have settled answers to every governance question. What evidence should be required before deployment? Which uses should be prohibited? How should responsibility be divided between a model developer and an organization using its model?

Inquiry means investigating those questions, comparing alternatives, and testing assumptions.

Practice means someone must turn an answer into a procurement clause, an evaluation, a reporting requirement, or a decision to suspend a system.

For our health ministry, inquiry might establish whether a triage tool performs reliably across different patient groups. Practice determines who conducts that evaluation, what evidence the ministry receives, and what result would stop the purchase.

Allan Dafoe's 2018 research agenda approached AI governance through questions about the technical landscape, political dynamics, and desirable governance arrangements.[2] It offers a useful precedent: the field needs research into which arrangements might work.

Research and implementation need each other. A mechanism built on a poor understanding of the problem may enforce the wrong thing. An insight that never reaches a decision may leave the problem untouched.

> **What do we need to learn—and how will that knowledge change what someone does?**

II

"Societies": who gets to decide?

It is easy to say that we should govern AI.

Who is "we"?

Governments have authority. Companies control products and infrastructure. Researchers possess expertise. Workers understand how systems change their jobs. Affected people know consequences that may never appear in a performance report.

These roles carry different responsibilities and different claims to influence. Expertise can help determine whether a tool works. It cannot, by itself, settle which trade-offs people should accept.

Our ministry may have the authority to purchase a system. Clinicians and patients still need ways to influence how it is used. Participation requires information, accessible channels, and some route through which objections can affect a decision.

A meeting alone does not provide that.

The question crosses borders too. The UN's 2024 Governing AI for Humanity report found that 118 countries were parties to none of the seven prominent non-UN international governance initiatives it examined.[3] This concerns a defined sample, rather than every global discussion. It nevertheless exposes a substantial representation gap.

This is why I use societies in the definition. It keeps differences of authority, interest, and circumstance visible.

Humanity expresses the wider ambition. Decisions made in one place can affect people elsewhere. Their interests belong in the question even when their governments or communities have little influence over the answer.

> **Who has authority, who is affected, and how can those affected influence or challenge the decision?**

III

"Shape": which possibilities are we making easier?

Once we ask who decides, we can ask what governance enables them to do.

Governance can restrict an unsafe use. It can also fund research, train public officials, improve access to computing resources, and support locally relevant development.

For the ministry, governing the tool could mean requiring evidence before purchase. It could also mean investing in the staff needed to evaluate suppliers intelligently.

The African Union's Continental AI Strategy illustrates this enabling role.[4] Its development agenda includes skills, research and innovation, and an institutional environment supporting applications in areas such as health, agriculture, and education.

A useful tool may never reach people who need it. A government may remain dependent on services it cannot understand, negotiate over, or replace. Those are governance concerns alongside harmful deployment.

> **Which possibilities are we making easier to pursue—and for whom?**

The next four parts of the definition identify practical places where those possibilities are shaped.

IV

Incentives: what does the environment reward?

Imagine two companies that both believe further safety testing would be useful. Testing takes three months. Each believes its competitor will release in six weeks.

Their intentions matter. So does the pressure to move first.

Competitive pressure appears in published policy. OpenAI's 2025 Preparedness Framework update said it "may adjust our requirements" following a rival's release of a high-risk system without comparable safeguards.[5]

The provision also imposed conditions: a changed risk landscape, public acknowledgment, an assessment that severe-harm risk would not meaningfully increase, and safeguards more protective than the rival's. It shows how competition enters governance decisions; it does not establish that a particular release was unsafe.

Governance asks what would make careful testing a decision companies can sustain. Buyers might require evidence. Disclosure requirements might make skipped testing visible. Liability might change who bears the cost of failure.

Each measure needs scrutiny. The starting point is to understand what behavior the environment encourages.

The same question applies inside our ministry. If a department is rewarded only for processing patients faster, it may give too little attention to mistakes, accessibility, or unequal performance. If reporting problems threatens a team's budget, monitoring may produce reassurance rather than useful evidence.

When you hear a promise about responsible AI, ask:

> **What makes that promise hold under pressure?**

V

Institutions: who can act?

Changing incentives requires people and organizations capable of making decisions. Institutions give authority and responsibility somewhere to live: regulators, courts, company boards, procurement offices, professional bodies, and independent evaluators.

For our ministry, who approves the purchase? Who can inspect the evidence? Who investigates a complaint? Who can suspend use?

Naming an organization is only the beginning.

It needs expertise, resources, access to information, and enough independence to exercise its authority.

A regulator may have a strong mandate and too few staff. A company's safety team may understand a problem and lack the power to delay release. A complaints office may receive objections without being able to secure a remedy.

Jennifer Wang and colleagues examined the publicly disclosed implementation of sixteen companies' voluntary White House commitments. Their 2025 workshop paper reported an average rubric score of 53%, using information available through the end of 2024.[6]

That score does not measure everything the companies did internally. It reveals a verification problem: public commitments leave outsiders uncertain unless there are ways to obtain and assess evidence of implementation.

> **Who is responsible? And what makes them capable of exercising that responsibility?**

VI

Infrastructures: what does this depend on, and who controls it?

An institution's options depend partly on the resources it can access.

AI relies on computing resources, data, networks, electricity, technical skills, and capital. Their distribution shapes who can build, who can adapt, and who must accept someone else's terms.

The World Bank's AI Foundations report organizes readiness around connectivity, compute, context, and competency: infrastructure, computing resources, relevant data, and skills.[7]

Its accompanying factsheet reports that high-income countries hosted 77% of global co-location data-center capacity in mid-2025.[8] This specific infrastructure measure helps make the uneven material basis of participation visible.

Our ministry might buy a tool that performs well but depends on a proprietary service. Patient records, staff training, and clinical workflows could gradually become tied to it.

The supplier controls prices and product changes. The ministry needs to understand what alternatives it will retain.

Foreign supply is not automatically a governance failure. Cooperation and specialization can provide valuable capabilities. The practical questions concern information access, bargaining power, data portability, and the ability to replace a service.

Representation in a discussion does not create those abilities by itself.

They require capacity.

> **What are we becoming dependent on—and what alternatives will remain?**

VII

Safeguards: what makes protection real?

Institutions organize responsibility. Infrastructures shape their options. Safeguards translate a protective objective into something that happens in practice.

"AI should be accountable" expresses a principle.

A safeguard makes some part of that principle operational.

For our tool, safeguards might include testing on relevant patient populations, protecting patient data, recording recommendations, enabling clinicians to override them, investigating incidents, and providing routes to challenge consequential decisions.

Human review deserves particular care. A person's presence does not guarantee meaningful oversight. They need time, relevant information, competence, and authority to intervene.

I use five questions to examine how fully a safeguard has been specified:

1	2	3	4	5
Who acts?	What is required?	What evidence is produced?	What decision follows?	What happens if the requirement is not met?

Suppose the tool performs poorly for a patient group. Who detects that? Does the evidence trigger further testing, a restriction, or suspension? Who verifies that the response occurred?

Australia's Robodebt Royal Commission offers a concrete lesson. Its recommendations called for clear review routes, plain-language explanations of automated decision-making, independent scrutiny of algorithms, and a body empowered to monitor and audit automation.[9]

Robodebt concerned automated administration rather than today's generative AI. Its relevance is the need to make scrutiny and remedy possible when automated processes affect people.

The NIST AI RMF Core similarly connects governance with responsibilities, policies, monitoring, and decisions.[1] A safeguard needs a process through which evidence can lead to action.

> **What prevents, detects, limits, or remedies failure—and what happens when it fails?**

VIII

Power runs through all four

These dimensions interact. An institution can change an incentive. Infrastructures can limit what an institution can inspect. A safeguard may depend on professional norms that encourage people to report problems.

Power runs through those relationships: the ability to set terms, obtain information, refuse a deployment, demand an explanation, or secure a remedy.

Governance mechanisms create power too.

Imagine a requirement for independent evaluation before certain systems can be deployed. If evaluation is expensive or available from only a few providers, established companies may find it easier to comply. A protective measure could also raise barriers for smaller organizations.

That possibility calls for better design: proportionate requirements, credible evaluators, and attention to who can access the process.

Girish Sastry and colleagues make a related warning in their research on compute governance: interventions through computing resources may help achieve safety and other objectives, while poorly scoped approaches can threaten privacy or centralize power.[10]

Every proposal therefore needs scrutiny of the authority it creates, alongside the harm it seeks to prevent.

> **What power does this mechanism create—and who will hold it accountable?**

IX

“Broadly shared benefits”: who gains?

A system can improve a headline measure while distributing its gains and costs unevenly.

A 2019 study by Ziad Obermeyer and colleagues shows how. A widely used healthcare algorithm predicted spending as a stand-in for medical need. Because less was spent on Black patients, it underestimated their needs: at the same risk score, Black patients were sicker than White patients.[11]

In the studied setting, correcting the disparity would increase the proportion of Black patients selected for additional help from 17.7% to 46.5%.

This was a care-management algorithm, rather than our triage tool. The lesson transfers: accurate prediction of a chosen target can coexist with unfair allocation of help.

Our ministry must therefore define what improvement means. Shorter waiting times? Better access? Fewer missed urgent cases? Improved working conditions? Which groups should benefit, and what evidence would show that they do?

UNESCO's Recommendation on the Ethics of AI makes an explicit commitment to human rights, dignity, fairness, inclusion, and accessible benefits.[12] It establishes a standard to pursue, rather than evidence that a particular policy achieves it.

Benefits belong inside the governance question from the beginning.

> **Who gains, who bears the risks and costs, and what would count as a fair outcome?**

---

x

"Today's choices become tomorrow's conditions": why timing matters

At first, the ministry is choosing whether to buy a tool.

Later, it may be choosing whether to renew a service around which hospitals have reorganized. Staff have been trained. Data formats have changed. Alternative procedures have disappeared. Replacing the tool now means changing an entire working environment.

The decision remains reversible in principle. In practice, reversal may have become expensive and disruptive.

David Collingridge described this dilemma in The Social Control of Technology: early in a technology's development, its consequences are difficult to understand; later, when they are clearer, changing it may be much harder.[13] The OECD's account of the dilemma quotes the first half:

“When change is easy, the need for it cannot be foreseen.”

That creates two questions:

1	2
Do we have enough evidence to act?	What happens to our ability to act if we wait?

Some decisions are easy to test and reverse. Others create dependencies quickly. A limited pilot may reveal useful evidence; replacing essential capacity before that evidence arrives may remove the ability to withdraw.

Preserving choice does not require keeping every option open forever. A justified prohibition deliberately closes an option. The purpose is to make consequential commitments through legitimate, informed decisions—and retain ways to correct arrangements that fail.

The OECD’s anticipatory-governance framework approaches this through values, foresight, stakeholder engagement, adaptive regulation, and international cooperation.

Our practical test is:

> **What are we committing ourselves to, and how difficult will it become to change course?**

XI

Governance must be able to learn

Acting while choices remain open means acting with incomplete knowledge.

A rule can be wrong. An evaluation can miss something important. A reporting requirement can impose costs while producing little useful information.

Governance mechanisms also become embedded. Agencies acquire mandates. Organizations build compliance systems. People develop interests in keeping existing arrangements.

Revision needs to be part of the original design.

Jack Stilgoe, Richard Owen, and Phil Macnaghten's responsible-innovation framework joins anticipation with reflexivity, inclusion, and responsiveness.[14] In plain language: consider consequences, examine your assumptions, involve others, and be capable of responding.

For the ministry, learning requires information: performance monitoring, incident reports, complaints, independent evaluation, and records that allow investigators to understand what happened.

An institution cannot respond to consequences it cannot observe.

Adaptation also needs accountability. Who can request a review? What evidence triggers it? Who can revise the requirements? Must they explain their reasons? Can affected people challenge the change?

Some technical requirements may need frequent revision, while commitments to rights and lawful procedure provide continuity.

> **How will we discover that this arrangement is failing—and how will we change it?**

XII

What this changes in practice

Return to the ministry.

It began with a familiar purchasing question:

Is the model accurate enough?

It now has a larger set of choices to examine.

The ministry might authorize a limited pilot while retaining existing clinical capacity. It could require evidence about performance across patient groups, secure access to relevant records and rights to export data, establish independent scrutiny, and specify conditions for expansion or suspension.

These are illustrative choices, rather than a universal prescription. Their purpose is to show how the questions change a decision.

The ministry is examining the power it transfers, the dependency it creates, the benefits it expects, and the alternatives it retains.

That is what makes this framework practical. Its parts give us a way to move from a general concern to choices someone can make, inspect, and challenge.

XIII

A guide to carry into the next conversation

Begin with the real-world outcome a proposal is meant to shape. Then use these questions to examine the arrangements behind it.

LOOK AT	ASK
Inquiry and practice	What must we learn, and how will it change action?
Authority and participation	Who decides, who is affected, and who can challenge?
Incentives	What behavior is rewarded under pressure?
Institutions	Who can act, and do they have the capacity?
Infrastructures	What dependencies are created, and what alternatives remain?
Safeguards	What protects people, and what follows when it fails?
Power and distribution	Who gains, who bears costs, and who holds new authority?
Timing	What becomes harder to reverse if we wait?
Learning and revision	What evidence triggers review, and who can change course?

The questions apply across research, development, access, deployment, and use. They can help examine civilian or military applications, a company policy, a public purchase, or an international agreement.

They reveal missing evidence and make disagreements more specific. They cannot remove the need for judgment about values, rights, authority, and acceptable trade-offs.

XIV

A new way to talk about AI governance

Regulation establishes enforceable requirements. Standards make expectations repeatable. Evaluations produce evidence. Procurement influences markets. Technical controls restrict actions. Public investment builds capacity. International cooperation addresses problems that cross borders.

We can understand those instruments better once we can explain the choices they are meant to shape.

The foundations are established. Dafoe helped articulate a research agenda for AI governance. Collingridge explained the timing dilemma. Responsible innovation and anticipatory governance connected early intervention with reflection, participation, and responsiveness.

What I offer is a teachable synthesis: one definition whose parts become questions, and whose questions help people examine decisions.

AI capabilities open possibilities. Human arrangements shape which possibilities are pursued, who benefits, and who can object. Repeated choices can become patterns; embedded patterns can become conditions that others inherit.

When someone asks what AI governance is, we can begin there:

> **How do societies shape the conditions around AI—and who has the power to influence those choices?**

Then we can ask what good governance must preserve: broadly shared benefits, accountable power, protection from harm, and the ability to learn and change course.

Govern while choices are still open. And build governance that can still change its mind.

NOTES

1. NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0).
2. Dafoe, A. (2018). AI Governance: A Research Agenda. Centre for the Governance of AI.
3. UN High-level Advisory Body on AI (2024). Governing AI for Humanity: Final Report.
4. African Union (2024). Continental Artificial Intelligence Strategy.
5. OpenAI (2025). Our updated Preparedness Framework.
6. Wang, J., Huang, K., Klyman, K. & Bommasani, R. (2025). Do AI Companies Make Good on Voluntary Commitments to the White House?
7. World Bank (2025). Digital Progress and Trends Report 2025: Strengthening AI Foundations.
8. World Bank (2025). Strengthening AI Foundations: Emerging Opportunities for Developing Countries (factsheet).
9. Royal Commission into the Robodebt Scheme (2023). Report.
10. Sastry, G., Heim, L., Belfield, H., et al. (2024). Computing Power and the Governance of Artificial Intelligence.
11. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science 366.
12. UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence.
13. Collingridge, D. (1980). The Social Control of Technology. Frances Pinter.
14. Stilgoe, J., Owen, R. & Macnaghten, P. (2013). Developing a framework for responsible innovation. Research Policy 42(9).

Printed from aigovspec.com/read · Research preview, independent expert review pending · Dataset 0.7.0, dated 2026-10-27 · PDF made 2026-10-27 · CC BY 4.0