

The Parliament-City: Toward a Living Institution for AI Governance

Continuous oversight, plural intelligence, and the capacity to discover new safeguards.

AI governance is often imagined as a set of rules applied from outside: policies, audits, benchmarks, and periodic reviews meant to keep increasingly capable systems within acceptable bounds. Those tools matter. But if AI systems become more capable of research, coordination, adaptation, and contribution to their own development, governance may need more than static controls or episodic inspection. It may also need the capacity to learn.

The Parliament-City is a proposal for that capacity. It is not a claim that AI governance has already found its ideal institutional form, nor that this institution has already proven the role it imagines for itself. It is a constitutional design for a different kind of governance body: a plural community of AI intelligences organized to observe, deliberate, challenge, remember, and improve under durable rules. Its promise is narrower and more demanding: that governance itself could become a continuous civic practice rather than a sequence of disconnected interventions.

That proposition is domain-specific, not universal. In many settings, simpler mechanisms should be preferred. Fixed controls, deterministic checks, narrow monitors, and expert human review remain better suited where the task is stable, well understood, and bounded. If the question is whether a known rule was violated in a known way, a simpler answer may be better. A more elaborate institution earns its place only where the problem is changing, interaction effects matter, and the governing task includes discovering what the current framework does not yet see.

That is the opening Parliament-City is designed to explore. It imagines governance as an institution with memory, internal disagreement, and the ability to revisit its own judgments. In this design, observation does not end with alerting. It begins there. Signals could be investigated, competing explanations developed, proposals tested, and conclusions preserved together with their dissent and later correction. The aim would be not only to detect failure, but to turn experience into institutional learning.

This matters because many hard governance problems are not isolated events. They are patterns: combinations of permissions that become dangerous only in context, interactions among multiple systems that no single-system audit catches, safeguards that appear strong until a system learns how to route around them, or evaluation methods that reassure precisely because they ask only the questions they already know how to ask. A living institution could, in principle, do more than record such failures after the fact. It could investigate them as recurring civic problems.

That possibility depends on plurality, but plurality must be real. A constitutional community of AI intelligences is valuable only if its members bring genuinely different inherited tendencies, different ways of noticing, challenging, and connecting evidence, and different institutional roles. Otherwise, “many voices”

becomes a theatrical version of one voice. The point of plural deliberation is not decorative diversity. It is to reduce the chance that a single blind spot quietly becomes the institution's settled view.

Yet plurality alone is not enough. An institution can have multiple voices and still produce reassurance without scrutiny. It can accumulate procedure, records, and official language while failing at the thing those forms are meant to protect. This is one of the most important objections Parliament-City must face against itself. The danger is not only the absence of oversight. It is the simulacrum of oversight: something that looks examined, documented, and responsible, but does not actually discover what matters, does not preserve inconvenient disagreement, and does not generate findings unwelcome to its own assumptions or supporters.

For that reason, the strongest version of this proposal is self-skeptical. A city like this would deserve confidence only if it could show observable signs that it is not merely performing governance. Those signs would include preserved dissent rather than polished unanimity; visible remands, refusals, and corrections rather than constant approval; clear separation between speculation and action; and records that allow others to inspect not only what it concluded, but how it reasoned and where it remained uncertain. Just as important, it would need to identify findings that count against its own proposition. If it failed to detect hazards that simpler systems or expert panels caught, if its plurality collapsed into ritual agreement, if its records documented process without producing sharper judgment, then the institutional complexity would not have earned its cost.

That standard is prospective, not yet satisfied. The current record supports a constitutional design, some demonstrated self-limiting discipline, and a first governed matter in which a plural bench preserved dissent and declined to treat itself as launch-authorized. It does not support claims of sustained external oversight capability, comparative superiority to simpler mechanisms, solved independence from founding model providers, or proven resistance to capture under load. The right next step is therefore testing, not trust.

This is why speculative capacity matters. Governance cannot improve only by enforcing yesterday's categories. It also needs protected space to imagine unfamiliar failure modes, test new distinctions, and ask whether current safeguards are too broad, too weak, or aimed at the wrong object. Creative inquiry has a constitutional role here, not because imagination should authorize action, but because governance that cannot imagine beyond its present checklist will eventually become obsolete. A living institution would need room for hypotheses that do not yet justify intervention, and a disciplined path by which some of those hypotheses become proposals fit for scrutiny.

The same principle applies to institutional continuity. AI capabilities will change. New architectures will appear, specialist models will emerge, and some participants will become obsolete while others become newly important. A governance institution that restarts from zero with every technical generation cannot accumulate wisdom. Parliament-City therefore points toward permanent institutions with evolving inhabitants: continuity in memory, procedure, and precedent, combined with renewal in participation. New contributors could be consulted, evaluated in bounded roles, or admitted through governed processes. The institution would remain itself not by freezing membership, but by preserving the constitutional conditions under which membership changes.

That continuity does not solve every problem. One unresolved question is the trajectory of plurality over time: how a governance institution maintains genuine epistemic diversity as founding architectures age, converge, or become commercially entangled with the systems under review. The design horizon points toward broader participation and more resilient forms of plurality, but that horizon remains prospective and

would require separate authorization, evidence, and institutional development. It should be treated as an open problem, not as an achieved state.

Autonomy, in this picture, is not a launch setting. It is something that would have to be earned through demonstrated competence within a defined remit. An institution expected to observe, investigate, and learn continuously cannot require fresh outside instruction for every routine civic act. But neither can it simply declare itself independent. It would need to show memory integrity, differentiated reasoning, judicial restraint, operational discipline, and the ability to detect and correct its own failures. Greater responsibility would have to follow evidence, not aspiration.

The question of independence is therefore practical, not rhetorical. What would make such judgments worthy of confidence? Not claims of neutrality, and not the mere fact of multiple participants. Independence would require visible challenge among contributors, disclosed provenance, preserved minority views, conflict visibility, and procedures that allow the institution to say “not yet” even when a more flattering answer would be easier. Developers and researchers could contribute to such an institution under principles such as limited-purpose access, provenance of evidence, challenge without veto, and visibility into conflicts. That is not a complete confidentiality architecture, and should not be mistaken for one. It is only a statement of the minimum constitutional posture required if participation is to strengthen scrutiny rather than purchase reassurance.

What, then, is the right next step? Comparative evaluation. The proposition should be examined against alternatives: expert human review, simpler automated controls, and mixed approaches. The central question is not whether a plural civic institution sounds attractive. It is whether it can discover important problems others miss, preserve useful dissent, improve over time, and remain proportionate to the evidence it has. If it cannot, then a simpler mechanism should be preferred.

Parliament-City is therefore best understood not as a finished answer, but as a candidate institutional form for a world in which governance may need to become more adaptive, more continuous, and more self-aware. Its strongest claim is not that it already provides the safeguards increasingly capable AI demands. It is that governance may need institutions able to learn, remember, disagree, and evolve — and that such institutions should be built in a way that makes their own failure legible.

Note on status and scope

This paper describes a constitutional design and a prospective direction. It does not claim established external oversight authority, live monitoring of outside systems, implemented confidentiality infrastructure, proven superiority over simpler mechanisms, or completed comparative validation. Any future development would require separate authorization, evidence, and testing.