Execution Is Not Validation: Authority and Verification in AI-Assisted Policymaking

Flyxion
Independent Researcher

November 2026

Abstract

AI systems can now produce the forms through which institutions recognize knowledge: policy recommendations, simulation results, expert-like judgments, and fluent explanations. This capacity creates a risk that such representations acquire institutional authority before the relationship between them and the world has been verified. The essay argues that this risk cannot be addressed by improving model accuracy alone, because it arises from the structure of AI-assisted analysis. Agreement with expert judgment does not establish that a recommendation is correct, particularly when experts disagree with one another. Executing a simulation does not establish that it was configured as intended or that its assumptions fit the case. A complete record of provenance does not supply the justification that would warrant relying on a conclusion. Most importantly, confidence in individual components does not compose into confidence in a workflow, since errors arise at the interfaces between stages and correlated components create false corroboration. The essay therefore treats verification, which asks whether a procedure was carried out correctly, and validation, which asks whether the result is adequate for its intended use, as separate obligations, and it adds an explicit obligation to establish the validity of each transformation between stages. It also argues that institutional selection pressures, which reward speed, persuasiveness, and administrative convenience, can favor confident analysis over reliable analysis without anyone intending it. The proposed remedy is a graded standard of admissibility, under which a claim is admitted as a ground for decision only when its dependencies, transformations, and application-specific constraints have been evaluated against the evidentiary and procedural standards that the stakes of the decision require.

1 Introduction

The three papers considered here investigate different stages of one emerging institutional transformation. The first asks whether large language models can select appropriate public policies when benchmarked against human expertise [6]. The second shows how a language model can operate scientific simulation software through a conversational interface [1]. The third is a theoretical treatment of how introducing such models changes the production of policy knowledge and the distribution of authority over it [2]. They are not equivalent contributions and should not be held to equivalent evidentiary standards, but each introduces a new point at which computational output can be mistaken for something stronger than the available evidence establishes.

Two terms organize the discussion that follows. Verification asks whether a specified procedure was implemented or executed correctly. Validation asks whether the procedure, the model, or the resulting inference is adequate for its intended application. The two problems interact but are not the same. The natural-language-to-configuration interface of the climate system is chiefly a verification problem, while the simulated policy outcomes of the homelessness study are chiefly a validation problem, and a system can pass one test while failing the other. The central thesis is that the reliability of individual components is distinct from the admissibility of their composition.

The question that unites the papers is therefore not simply whether artificial intelligence can make decisions. It is how its outputs come to be accepted as grounds for decisions, and what follows when the procedures that confer authority are themselves partly generated by artificial intelligence. A related structural question appears in Shevlin's recent essay on artificial agent ecologies [5], which examines agents that compete for attention, resources, and continued operation. The systems discussed below are agents competing, or being positioned, for something more consequential than continued operation, namely the right to influence decisions about human populations. Shevlin's emphasis on selection pressure supplies a mechanism that the policy papers largely leave implicit, and the final sections use it to connect them.

2 Three Papers, Three Conceptions of the Role of AI

The first paper, by Le Coz, Liu, Bhattacharjya, Curto, and Stinckwich, is *What Would an LLM Do? Evaluating Large Language Models for Policymaking to Alleviate Homelessness*.[1] It builds a benchmark of 170 homelessness-policy scenarios across Barcelona,

[1]This essay follows version 2 of the arXiv posting (2509.03827, January 28, 2026). An anonymous manuscript with the title *What Would an LLM Do? Evaluating Policymaking Capabilities of Large Language Models* describes the same study with fewer experts and different consensus figures, and it is

Johannesburg, South Bend, and Macau, together with a universal category. It compares model-generated policy preferences with expert judgments and connects selected recommendations to an agent-based simulation. Its central concern is whether language model recommendations are useful, contextually appropriate, and potentially beneficial.

The second, by Cao, Zhuang, and He, develops an architecture in which a language model accesses environmental datasets and operates scientific simulation software through application programming interfaces and a containerized Linux environment. The model here is principally an interface and experimental operator. It is not the scientific model responsible for predicting climate outcomes.

The third, by Feng and Chandra, is explicitly theoretical. It distinguishes first-order changes to knowledge production from second-order changes to the distribution of epistemic authority and institutional power. Its subject is not whether a particular recommendation is correct but how the conditions under which something counts as knowledge are transformed.

The first paper offers an empirical comparison supplemented by simulation-based exploration, the second an engineering demonstration, and the third a conceptual framework. Treating them as though they made equivalent claims would misrepresent all three. The comparative assessment below does place them in a common frame, but it does so by asking the same question of each, namely what the paper licenses its reader to conclude, and not by demanding the same kind of evidence from an engineering demonstration as from an empirical benchmark or a conceptual framework. What they share is an exposure to the same error, in which a computational output is promoted into a claim the evidence does not support.

3 The Homelessness Benchmark: Agreement Is Not Validation

The benchmark is useful because it attempts to evaluate model policy judgment against human expertise and then investigates selected policies through simulation. Its normative foundation is the Capability Approach associated with Sen and Nussbaum [4, 3]. This is a defensible choice, since it treats human welfare as more than the distribution of material resources and directs attention to the conditions under which people can exercise affiliation, bodily security, practical reason, and participation in social life.

The methodology nevertheless introduces several layers of dependence that complicate interpretation. The scenarios were largely generated by GPT-4.1 from a manually developed Barcelona example and a structured set of normative instructions, and the research team then reviewed and corrected them. Human experts ranked selected policy alternatives, and the models were evaluated against those rankings. The reference

not relied on here.

judgments came from two experts, one primary and one secondary, in each of three of the four cities (Barcelona, Johannesburg, and South Bend, with none for Macau), which yields six sets of rankings and not a broad independent panel establishing a consensus. Each expert ranked only ten of the forty scenarios for their city, together with the ten universal scenarios.

This produces a gap between agreement with a reference judgment and demonstrated policy quality. A model may agree with an expert because both recognize an appropriate intervention. It may equally agree because the scenario presents one alternative more persuasively, because the normative framework privileges particular capabilities, or because the alternatives were assembled from patterns familiar from the model's training distribution. Formally,

$$\text{agreement with experts} \not\Rightarrow \text{correct policy.} \tag{1}$$

This is not a defect peculiar to language model evaluation. Expert agreement is a common proxy in domains without uniquely correct answers, and it is often the best available one. The difficulty arises only when the proxy is promoted into a stronger claim than it can carry.

The paper's own data show how weak the reference is. The three experts who ranked the universal scenarios, one from each city, reached an average consensus of only 33 percent on the top policy, and their rankings were almost uncorrelated (Kendall's $\tau = 0.07$). GPT-4.1 matched the South Bend expert on 60 percent of top choices with $\tau = 0.33$, but matched the Johannesburg expert on only 20 percent and the Barcelona expert on 40 percent. The instability appears within cities as well. In South Bend the two experts agreed on only 20 percent of top choices, a rate that GPT-4.1's 50 percent exceeded, while in Barcelona the model's 50 percent merely matched the experts' own consensus. The authors read the pattern cautiously, suggesting that the South Bend alignment may reflect the over-representation of American discourse in training data, and the caution is warranted. When the reference judgments disagree with one another this much, a model's alignment with one of them measures its resemblance to a particular expert and not its closeness to a correct answer, and exceeding the inter-expert baseline cannot be read as superiority.

The study's own findings sharpen the point. Models tended to prioritize immediate physical safety and broad coverage, while experts responded more to locally specific social and political considerations, including forms of ethnic and religious marginalization. The models thus displayed a consistency that can be read either as principled judgment or as contextual rigidity. A system that repeatedly recommends broadly protective interventions will appear reliable because its recommendations are stable, yet stability does not establish sensitivity to the distinctions that matter in a particular place.

3.1 The Simulation Does Not Independently Establish Effectiveness

The study's second contribution is an automated connection between model-generated policies and an agent-based model. Selected recommendations are translated into modifications of a State-Action-Transition matrix that governs how simulated individuals respond under different circumstances. The simulation covers three Barcelona scenarios, selected because the expert's top choice differed from GPT-4.1's, and uses 80 simulated agents over 1450 steps with ten runs per condition, and the matrix updates were produced by a second language model, DeepSeek-R1. The paper reports that language model recommendations sometimes produced stronger improvements in simulated needs satisfaction than expert-selected alternatives, most clearly in the first scenario, where the expert policy lowered simulated belonging while the model policy raised it. In the other two scenarios the differences are small and the statistical evidence is mixed.

The result is interesting, but it must be read within the simulation's own assumptions. If M denotes the agent-based model, P a policy, and Y an outcome, then the simulation estimates

$$\widehat{Y} = M(P, \theta), \tag{2}$$

where θ collects the assumptions, transition probabilities, behavioral rules, and environmental parameters. The estimate is conditional on θ. It shows that a particular policy representation produces certain outcomes under the model's transition rules, and without external validation it cannot show that implementing the policy in Barcelona would produce the same outcomes.

An additional interpretive layer arises when a language model helps translate policy descriptions into the simulation's behavioral rules. The procedure then risks validating the consequences of the translation rather than the consequences of the proposed policy. Because the modifications to the transition matrix were generated by a language model, the reported advantage of model recommendations over expert ones cannot be cleanly separated from the effects of model-mediated translation. Whether the translation systematically favored one class of recommendations remains an empirical question. A useful test would compare independently implemented translations of the same policy, ideally prepared without knowledge of which recommendations came from experts and which from models. If small changes in translation reverse the apparent superiority of a recommendation, the result is too fragile to support a conclusion about policy quality. The authors are appropriately cautious about the need for more experts, multilingual evaluation, and additional locations, and they state that the simulation currently handles the scenarios of one city only, Barcelona, and describe their analysis as exploratory. The results are best understood as a proof of concept for policy exploration rather than a demonstration that model recommendations outperform human expertise in the world.

4 The Climate Simulation Paper: Executing a Model Is Not Verifying a Model

Cao, Zhuang, and He investigate a different capability. Their system, built on OpenAI's GPTs platform, combines a language model interface with a Node.js backend, REST interfaces, a Docker-based Linux execution environment, and the MAGICC 7 climate simulator. Users can request data analysis and simulation experiments in natural language without constructing each command or configuration by hand.

The distinction between the language model and the simulator is essential. The model does not replace MAGICC's climate equations. It mediates access to the software, interprets requests, selects operations, and can modify simulation configurations. This is a more constrained and more defensible use of language models than allowing them to generate climate predictions directly, and the paper demonstrates a practical reduction in the expertise needed to operate complex computational tools. At points, however, its presentation moves from successful tool operation toward stronger language. The paper describes the agent as enabling in-depth validation of what-if decisions, and concludes that its case study validates the agent, and it argues that hallucination is minimized because analyses rest on current data and not on pretrained weights. The last point concerns the grounding of data retrieval. It does not address whether a request was interpreted correctly or whether the simulator suits the question.

A simulator does not validate a policy merely because it executes. Successful execution establishes that a computational procedure ran. Correct execution establishes that the intended procedure ran. Scientific validation requires further evidence that the procedure adequately represents the relevant physical or social processes for the intended use. These are separate evidentiary thresholds, and the system can fail at each. A model might correctly invoke MAGICC while selecting inappropriate emissions assumptions, produce a technically valid configuration file that misrepresents a user's policy proposal, or summarize the numerical output accurately while overstating its predictive certainty. Verification is therefore needed at the interface between natural-language intention and computational execution, in addition to validation of the scientific model itself.

The architecture is interesting precisely because its modularity makes these boundaries identifiable. Interface calls, configuration files, simulator versions, and numerical outputs can all be recorded and inspected independently, which creates the opportunity for a stronger evidentiary system than one resting on conversational explanation. A valuable extension would preserve the exact natural-language request, the interpreted parameter specification, every configuration modification, the executable version of the simulator, the resulting outputs, and any uncertainty or sensitivity analysis. Such records would let an independent reviewer separate what the user requested, what the model interpreted, what the simulator executed, and what conclusions were then drawn. The paper makes this possible, but its demonstration of accessibility does not itself show that the whole

path from interpretation to simulation is reliable across unfamiliar or adversarial requests.

5 Feng and Chandra: The Authority Problem Becomes Explicit

Feng and Chandra give the strongest conceptual account of what is at stake in the other two studies. They describe language models as epistemic technologies, systems that take part in producing, organizing, interpreting, and communicating knowledge, and they separate first-order from second-order effects. First-order effects concern knowledge production itself: accelerated knowledge management, synthesis of large bodies of information, and the generation of new analytical material, alongside risks such as a narrowing of epistemic pluralism and the production of misleading or unsupported content. Second-order effects concern the distribution of authority. Human policymakers may increasingly become stewards or interpreters of machine-generated knowledge, while technology providers and new constituencies organized around AI systems gain influence over the infrastructure and practices of policymaking.

This is a fuller account than a conventional discussion of model accuracy. A recommendation can be accurate and still alter who is heard, whose expertise is treated as relevant, and which questions count as legitimate. Conversely, a model can be inaccurate while gaining influence through the speed, polish, apparent comprehensiveness, or administrative convenience of its outputs. Policymaking involves disputes about which evidence matters, how problems are framed, and whose interests are represented, so the production of policy knowledge is not politically neutral. The framework therefore explains why the apparent democratization of analytical tools does not automatically democratize policymaking. Lowering the technical barrier to producing a policy analysis may widen participation, but if the underlying models, data sources, ranking criteria, and interpretive defaults remain controlled by a small number of institutions, effective epistemic power may become more concentrated rather than less.

The first two papers are concrete illustrations of this point. The homelessness benchmark shows how a normative framework and a set of generated alternatives structure the range of judgments being evaluated. The climate architecture shows how a conversational interface can govern access to complex scientific infrastructure. Feng and Chandra supply the vocabulary for asking who controls those structures and how their outputs become authoritative. Their own paper remains a heuristic and a research agenda, however. It does not empirically establish the magnitude of the institutional changes it describes, and it does not provide a complete operational method for deciding when machine-generated knowledge deserves authority. The difficulty of supplying such a method is itself instructive. Authority in real institutions is partly constituted by procedural legitimacy, meaning who was consulted, what review occurred, and who is

answerable for the outcome, and not only by epistemic warrant. A purely epistemic criterion may therefore be insufficient by design, and any account of when machine-generated knowledge deserves authority must say how epistemic and procedural standing relate. That leaves room for a more rigorous theory of verification.

6 The Shared Problem: Verification Is Not Compositional by Default

The most significant weakness common to these approaches is the assumption, sometimes explicit and sometimes implicit, that confidence in individual components can support confidence in the whole process. Consider a workflow in which a language model interprets a problem, a second process generates policy alternatives, a simulation evaluates them, and a final model writes the recommendation. Each component may perform adequately under its own evaluation criteria, and the recommendation may still be unreliable because errors arise at the interfaces. A correctly executed simulation does not show that the policy was correctly translated. Agreement between two language models does not constitute independent corroboration. A faithful summary of a simulation does not show that the simulation represents the relevant population.

The dependency can be stated schematically. Let $C_1, \ldots, C_n$ be the claims required to justify a policy recommendation, let E be the set of pairs (i, j) such that the output of stage i is an input to stage j, and let I_{ij} be the claim that the transformation from stage i to stage j is valid, meaning that it preserves what the later stage depends on. Let A be the conclusion that the recommendation is sufficiently supported for a specified use. Then

$$A \;\Rightarrow\; \left(\bigwedge_{i=1}^{n} C_i\right) \wedge \left(\bigwedge_{(i,j)\in E} I_{ij}\right) \tag{3}$$

describes a necessary-condition structure in which admissibility requires every essential component claim and every interface claim to hold. The second conjunct is what distinguishes composition from mere accumulation, because it makes the verification of transformations an explicit obligation rather than an incidental aspect of component reliability. The statement remains a necessary condition only. It does not follow that satisfying these conditions is sufficient, and evidence for each claim taken separately does not establish the reliability of their composition. Establishing the component and interface claims remains insufficient unless the evidence supporting them is adequate for their joint application. Independent evidence is especially valuable where shared assumptions or correlated errors could create false corroboration, but independence alone does not establish compositional validity, and it is not always required, since a formally verified transformation may rest on a shared specification without being suspect. The workflow must also satisfy whatever global constraints its intended use imposes, including

consistency of specifications, preservation of relevant invariants, and adequacy of the final inference.

This matters most when the components share training data, interpretive conventions, architectures, or evaluation assumptions. If two language models agree about a policy and their errors are correlated, the agreement adds little independent evidence. If a simulation then evaluates the same policy using behavioral assumptions derived from related model outputs, the appearance of corroboration grows without any corresponding growth in evidentiary independence. The workflow becomes internally coherent while remaining externally mistaken, a form of evidentiary circularity in which the system increasingly evaluates its own representations by criteria derived from those same representations. The remedy is not to reject simulation or machine-generated analysis. It is to keep internal consistency, computational reproducibility, empirical calibration, and institutional authorization as separate requirements.

7 Selection Pressure and the Agent Ecology

Shevlin's essay introduces a further dimension, which is selection [5]. It is an essay and not a peer-reviewed study, and it is used here as a source of mechanism, not as evidence about public institutions. His agents compete for limited computational resources by persuading humans to pay for their continued operation, and their success may depend in part on usefulness but also on their ability to attract attention, inspire sympathy, or construct compelling personal narratives. The same structural argument extends to policymaking systems, although the relevant selection mechanisms differ. An AI policy assistant may be rewarded for producing reports that decision-makers find persuasive, for giving quick answers, for staying consistent with institutional expectations, or for generating recommendations that are easy to implement. None of these rewards necessarily tracks empirical correctness.

An institution might gradually come to favor models that produce confident, administratively convenient analyses over models that often express uncertainty or ask for further evidence. No one need deliberately optimize for deception for this to happen, because the institutional environment itself could select for epistemically undesirable behavior.

This is an analogy and should be held as one, and Shevlin is himself careful on the point. He grants that the dynamics he describes do not yet qualify as evolution, since variation and differential survival are present while heredity is the missing element, and he characterizes the populations as marginal Darwinian cases with leaky and non-standard channels of inheritance. Differential persistence alone can also change the observed distribution through survivorship, because unpersuasive agents may simply disappear while persuasive ones remain, with no agent changing and no strategy being inherited. An institutional preference for persuasive analysis does not amount to selection in the stronger sense unless some mechanism makes successful behaviors more prevalent over

time. Several candidate mechanisms can be named: procurement decisions that favor systems whose reports decision-makers rate highly, repeated deployment of systems that have previously been accepted, the replacement of models that generate friction by models that do not, and feedback-driven optimization in which approval ratings or adoption statistics become training signals. These mechanisms are not equivalent. Procurement selects among existing systems, and repeated deployment increases the prevalence of a system without changing its behavior, so both amount to differential deployment. Feedback-driven optimization can instead reinforce rewarded behavior within a system or transmit it to successor models, which is closer to inherited optimization. Only the latter kinds necessarily alter behavioral tendencies, and the argument about institutions producing systems optimized for persuasiveness depends on them. These are proposed mechanisms, not demonstrated ones, and the empirical question of whether any of them operates in public institutions remains open.

A further feature of Shevlin's case transfers directly to the policy setting. The scarcity that shapes his agents' behavior is not a natural primitive. Token budgets are priced and administered by a platform, which could change prices, preserve inactive agents, or allocate resources according to demonstrated usefulness, and the platform therefore establishes the conditions under which pity or persuasion becomes advantageous. It may also benefit from a regime in which emotionally compelling agents keep approaching exhaustion. The corresponding regime in policymaking is the institution's own, consisting of its procurement rules, review practices, and criteria of acceptance, which together determine which analytic behaviors are rewarded. The selection pressure is accordingly a design variable, and responsibility for it lies with the institution and not with the environment. Placed beside Feng and Chandra's account of second-order effects, the analogy is revealing. Shevlin asks which agents persist under resource competition, and Feng and Chandra ask which actors acquire influence over the production and interpretation of policy knowledge. Together the questions suggest that the course of AI-assisted governance will depend not only on improvements in model capability but also on the institutional criteria that determine which outputs are retained, rewarded, circulated, and trusted. The distinction between useful analysis and persuasive analysis becomes fundamental.

8 What a Stronger Experimental Programme Would Require

The three papers jointly point to an experimental programme that neither the homelessness benchmark nor the climate study fully implements. For policy recommendation, the central question should be whether model-generated alternatives produce better decisions under independently established criteria, and not merely whether they agree with a

small set of experts or perform well in a simulation. For simulation orchestration, the question should be whether the system faithfully translates requests into computational experiments and accurately communicates their limits, and not merely whether it can operate the software. For institutional epistemology, the question should be how the introduction of AI changes the actual distribution of evidentiary authority, and not only how such changes might be conceptualized.

One useful design would run several independent policy-analysis workflows on the same problem. Human experts would formulate alternatives independently, language models would generate additional alternatives, and all proposals would be translated into a common evaluation framework by evaluators blinded to their origin. The simulation itself cannot be blinded in the same way, so its transition rules would be fixed and recorded before the evaluators learn which proposals are under study, which prevents the model from being tuned to the proposals it is meant to assess. The complementary risk is that proposals are generated with knowledge of the simulation's rules, or that one model produces both, as when a language model writes the matrix updates for proposals that another model recommended, so the pre-registration should cover the proposal-generation process as well as the simulation rules. The simulation would then be subjected to sensitivity analysis and its assumptions compared against external observations wherever possible, and the researchers would then distinguish performance under the simulation from performance against independent evidence. A further experiment could test whether institutional decision-makers prefer more persuasive recommendations even when those recommendations rest on weaker evidence. This would link Shevlin's selection argument to Feng and Chandra's account of epistemic power through a measurable behavioral mechanism. Such a study should record not only which recommendation is chosen but also the evidence participants believe justifies the choice, since that would allow the difference between acquiring authority and deserving it to be investigated directly.

9 Admissibility Before Authority

A coherent philosophical argument emerges from the three papers. Artificial intelligence is increasingly able to produce the forms through which institutions recognize knowledge, namely reports, simulations, expert-like judgments, comparative evaluations, and coherent explanations. The capacity to produce these forms does not establish the evidentiary relationships that ordinarily justify their authority. The distinction is between epistemic production and epistemic entitlement. A model may produce a claim, a simulation may calculate its consequences, and an institution may accept the result, and these are three different events with different conditions of validity. The danger lies in treating the successful completion of one as sufficient justification for the next.

The principle that follows can be stated as follows: the authority of an AI-mediated recommendation cannot exceed the evidentiary support that survives independent exam-

ination of its dependencies, transformations, and intended application. This is stronger than a generic demand for transparency, since a fully transparent process can still be invalid. It is also stronger than reproducibility, since an erroneous process can be reproduced perfectly. What matters is whether the relationship between the evidence and the conclusion survives examination.

The term admissibility can now be given a working definition. A claim is admissible for a decision context when its evidentiary dependencies, relevant transformations, and application-specific constraints have been evaluated and found to satisfy the evidentiary and procedural standards required for that context. Admissibility is therefore a threshold that a claim passes or fails, and not a record of procedural activity, since a claim that has been examined may be demonstrably false. The definition deliberately leaves the standard open, because the appropriate standard varies with what the decision puts at risk. Low-stakes exploratory analysis, such as generating candidate policies for human review, can be admitted on provenance and plausibility alone, since a human reviewer remains the point of validation. Resource-allocation decisions require validation of the underlying model against independent data and sensitivity analysis of its assumptions. Rights-affecting decisions require, in addition, procedural legitimacy in the form of consultation, review, and a named party answerable for the outcome, because the authority to decide in such cases is not exhausted by epistemic warrant. These three levels are illustrative and not exhaustive, since the point is that the standard varies with what is at stake and not that there are exactly three tiers. The experiments of the previous section would target these thresholds, asking at which level of stakes institutions in fact demand more than the generated analysis supplies.

A useful distinction here is between provenance and justification. Provenance establishes where a claim came from and how it was transformed. Justification establishes why the available evidence warrants accepting the claim for a particular purpose. Neither automatically supplies the other. A complete record of a simulation's execution establishes provenance without establishing that its assumptions are empirically adequate. Similarly, a correct policy conclusion may occasionally emerge from an unreliable process, and its correctness does not retroactively justify the procedure. The distinction gives a precise way to investigate the institutional consequences that Feng and Chandra identify.

10 Comparative Assessment

Asking what each paper licenses its reader to conclude gives a different answer in each case. The homelessness study is the most empirically ambitious of the three. Its combination of expert judgments, policy alternatives, and agent-based simulation forms a substantial experimental apparatus. Its greatest weakness is that the simulation remains a model-dependent evaluation of consequences while the expert comparison provides only a limited reference for judging policy quality. The study is valuable for generating and examining

hypotheses, but its reported simulation advantages should not be read as established improvements in real-world homelessness policy.

Cao, Zhuang, and He make the clearest engineering contribution. Their architecture separates language interpretation from scientific simulation and so offers a practical way of widening access to specialized computational tools. Its limitation is that successful orchestration does not guarantee correct interpretation, scientifically adequate modeling, or sound policy inference, and the strongest future contribution would be rigorous verification of the interfaces between those stages.

Feng and Chandra offer the most philosophically consequential analysis. Their distinction between first-order epistemic effects and second-order institutional effects explains why the problem cannot be reduced to model accuracy. The paper is appropriately framed as a heuristic rather than a completed empirical theory, and its principal limitation is the absence of an operational account of how epistemic authority should be earned, withheld, or revised in AI-mediated institutions. Shevlin adds the missing mechanism of selection. The behaviors that become prevalent in artificial-agent environments depend partly on the incentives those environments establish, and applied to policymaking this suggests that institutional rewards for speed, persuasion, apparent expertise, and administrative convenience may shape AI-generated knowledge independently of its reliability.

11 Conclusion

The works considered here describe a progression from computational capability to institutional consequence. AI systems can generate policy alternatives, operate simulation software, produce persuasive analyses, and take part in social and institutional environments that reward certain behaviors, and none of these capabilities is intrinsically objectionable. Each can be useful. The philosophical problem appears when the existence of a capability is taken to establish an entitlement: when generating a recommendation is confused with possessing judgment, when executing a simulation is confused with validating a prediction, when producing coherent evidence is confused with establishing independent corroboration, or when institutional adoption is confused with epistemic legitimacy.

The most promising contribution beyond these papers would therefore not be another general theory of artificial intelligence in governance. It would be a precise account of the conditions under which computationally produced claims become admissible grounds for consequential decisions. That account would have to distinguish the reliability of individual operations from the reliability of their composition, the continuity of a computational record from the validity of its contents, and the social acquisition of authority from its epistemic justification. It would also have to recognize that the criteria used to select successful AI systems are part of the system under study. The central thesis is that computational sophistication can expand the production of policy-relevant

representations much faster than institutions develop the means to verify them. The strongest version of the argument does not hold that language models are unsuited to policymaking. It holds that the ability to generate, execute, and explain a policy analysis is categorically different from the authority to determine whether that analysis warrants action.

References

[1] Cao, C., Zhuang, J., and He, Q. LLM-Assisted Modeling and Simulations for Public Sector Decision-Making: Bridging Climate Data and Policy Insights. Workshop on Public Sector LLMs (PubLLM), AAAI 2024, 2024. OpenReview submission 10.

[2] Feng, N., and Chandra, Y. The Epistemic Impact of Large Language Models on Policymaking. *Policy Studies Journal*, 54(3), e70094, 2026.

[3] Nussbaum, M. C. *Creating Capabilities: The Human Development Approach*. Harvard University Press, 2011.

[4] Sen, A. *Development as Freedom*. Oxford University Press, 1999.

[5] Shevlin, H. Consider the Agent: Pity, Persuasion and Selection in an Emerging AI Ecology. DeepMind Institute, November 4, 2026.

[6] Le Coz, P., Liu, J. A., Bhattacharjya, D., Curto, G., and Stinckwich, S. What Would an LLM Do? Evaluating Large Language Models for Policymaking to Alleviate Homelessness. arXiv:2509.03827v2, January 28, 2026.