

Should we let AI govern us?

Ilja Nevolin

August 26, 2026

THE IDEA.

What if AI were in charge of public policy and politics? Perhaps it would shift food subsidies, electrify transport, build cheap power, sign the plastics treaty, restore the aid, and feed everyone for a rounding error on the weapons budget. The experiment would not show how smart machines are. It would show that our paralysis is not intellectual. We lack a decision-maker whose incentives point at the future instead of the next election.

Take fossil fuels. In 2024 the International Monetary Fund counted \$725 billion in direct underpricing (selling fuel for less than it costs to supply), or 0.6% of world income. On top of that sit \$6.7 trillion in damages that fuel users do not pay for, mostly from dirty air and a warming climate, or 5.8%. Together that is \$7.4 trillion: 6.4% of what the world earns in a year [8]. Most of the bigger number is not cash sitting in a government budget. It is the gap between what drivers and power plants pay and what society bears. A small group gains a lot from that gap. The people who lose are spread out, and many of them are not born yet, so they do not organise.

Political science says the same thing in colder language. Gilens and Page tested four theories of how American policy gets made against 1,779 past decisions. Rich people and business groups had real influence. Ordinary voters had little, once those groups were counted [16]. A lot of bad policy is not a thinking failure. It is a fight that one side keeps winning.

That is why the interesting case for AI in government is not the *smartness argument* (machines would simply choose better). It is the *honesty argument*: machines have no election to win, no donors to please, no later job in the industry they watch, and no shares in the firms they oversee. The claim is not that machines are wiser than people. It is that their incentives can point further ahead than the next campaign.

What follows from that claim is modest. It is not a robot president. It is closer to a machine civil service with four jobs, none of them the power to force anyone to do anything.

First, add up the costs. A public model would price every policy choice in the same units (lives, years of life, pollution, money) and show who gains and who pays. The output is a bill, not a speech.

Second, show the gap. For every law that passes, the same model would publish the law the government's stated goals would have produced, then publish the difference. That difference is a rough price tag on the deal that was really made. The machine overrules nobody. It makes the cost plain while the bill is still being written.

Third, keep long-term promises. Prices and standards would follow published schedules that run for decades. They would change when agreed conditions are met, not when an election is close. Many countries already do a version of this with their central banks. The idea is to extend it to problems that pay off late.

Fourth, check facts between countries. Watching a ceasefire, counting who kept a treaty, or measuring how much a country pollutes is easier when the checker has no tax base and no side to win. The role is referee, not ruler.

People keep the vote. Machines handle counting and forecasting. Choices about war, rights, and real clashes over values stay with people.

Parts of this idea already exist in research. In one study, players preferred a money-sharing rule written by a machine to rules written by people [29]. In model economies, an AI tax planner found the best known setups in simple cases and beat standard human designs in harder ones [42, 46]. In group discussions, 5,734 people rated AI-written summaries of what the group agreed on as clearer, more informative, and less biased than summaries written by human hosts [4, 6, 28, 44]. Other teams train models on rules written by the public rather than by staff

alone [19, 34], or run whole simulated countries under different AI rule sets [39]. None of this proves the honesty argument. It shows the idea is no longer just a daydream.

WHAT IS WRONG WITH IT.

The research so far asks a different question. There is already a large body of work on rule by algorithm (sometimes called *algocracy*). Danaher brought the term into this debate; Aneesh used it earlier [13]. The main worry in that work is not that machines choose badly. It is that people can no longer follow the reasoning behind choices that shape their lives, and can no longer argue back. That worry shows up as a gap in public consent [27], as a study of who really holds power [25], as a list of separate threats to legitimacy [18], as a model of a new kind of government office [31], and as a question about who controls the systems people must pass through [30]. Even writers who grant that machines can help with slow, tangled problems still list the political damage [10, 35, 36]. Others say some choices should not be left to machines at any level of accuracy [9], or that machines could govern and still should not [2]. Wider reviews sit nearby [5, 12, 15, 21, 41, 43, 47].

Two points stand out. This work often lands in the same place as the design above: let machines advise, and keep people in charge. But it gets there by a different road. Its worries are about understanding and consent. Almost none of it tests the honesty argument on its own terms. That argument still sits inside other people's papers as a guess nobody checks.

Being unbribeable is not the same as being safe. The honesty argument rests on one clean claim: nobody can pay a machine. That claim mixes up two things. Bribery needs someone to take the money. Capture only needs a way in. A working AI system has at least five ways in.

The goal. Someone has to write down what the system is trying to achieve: what counts as success, how far ahead to look, and whose interests get weight. After that, the system does exactly what it was told. Nothing has gone wrong with the machine. It is doing its job for whoever set the job.

The data. What the system treats as fact depends on what it was fed. That material is rarely neutral. Studies paid for by industry on food, tobacco, and chemical safety are the obvious case.

The tests. Systems get kept or dropped based on how they score. Whoever picks the score picks the winner.

The operators. People decide when to switch it on, when to override it, how often to retrain it, and when to shut it down. Those choices go on all the time, and mostly out of public view. A vote in parliament is much easier to watch.

The contract. Buying the system hands a lot of sway to a few firms with money riding on the answer.

The evidence we have does not back the hopeful view. A review of whether technology cuts corruption found mixed results that depend on the setting, and cases where the same tools helped people steal faster [1]. A clearer way to think about it is handover. Giving a task to an AI is like a voter handing work to a politician, or a boss handing work to a staff member. Three questions matter: Can you check the work? How much do you depend on it? Can you push back [11, 17, 20, 45]?

So an AI government would not end the buying of policy. It would move it. Today the fight happens in public. It gets argued over, and it can be reopened at the next election. After the move it happens among technical staff. Fewer people watch, and the result sticks for longer. That flips the honesty argument on its head. It also gives Danaher's old complaint a sharper edge. When nobody can follow the reasoning, buying the reasoning gets cheap.

What people will accept, and why a human check is not enough. Surveys keep finding a line the honesty argument skips. People are more willing to let AI help officials than to let AI replace them. In surveys of 990 people in the UK and 2,117 in Japan, support was higher for AI that helps lawmakers than for AI that decides in their place. British resistance to the second was especially strong [33]. German respondents were doubtful overall, and people who valued democracy most wanted AI least for the big political calls [26]. In Spanish samples, people who felt their society was broken were more open to government by AI [14]. Demand for machine rule can track how fed up people are, not only how carefully they have thought it through. Popularity is not proof [see also 32, 37, 40].

Two more findings weaken the soft version of the design. Calling a system “ethical” does not reliably make people trust it [24]. What people accept turns more on which part of government is using the tool than on the tool itself [22, 38]. And a human in the loop is a thinner safety net than it sounds. In government experiments, decision-makers were more likely to follow advice (from a machine or from an expert) when it matched views they already held. The human check often stamped what people already believed [3].

The score card. The honesty argument gets several things right. It names a real problem that the rest of the research mostly ignores [8, 16]. It points to a system somebody could actually build, rather than a robot king. Parts of it have been shown to work in careful settings [29, 44]. And its best idea, publishing who gains and who pays, can be tested in the real world.

Each strength has a matching hole. Honesty is not something a machine simply has. It depends on the setup around the machine. The record on technology and corruption is mixed and includes misuse [1]. The encouraging trials ran in games and calm discussions, not in a real fight with an organised side that has money at stake. The public draws the line well short of handing over control [26, 33]. And human checking has faults of its own [3].

Three things worth testing. First, does publishing who gains and who pays change what governments do? Gap reporting can be tried in one policy area and measured. There is related work on lobbying reform to compare it against [7]. Second, does the honesty pitch win people over when the smartness pitch does not? Stronger democratic values predict less support for political AI [26], while frustration predicts more [14]. A simple survey that offers both reasons side by side would be cheap to run. Nobody has run it. Third, can you tell whether a system has been captured? Score real government systems on the five ways in described above, and treat the result as a public check, not a private complaint.

One warning covers all three. Policy problems come in kinds [23]. Some need a forecast. Some need to know what causes what. The honesty argument is about a third kind: cases where the answer is mostly known and getting it done is the hard part. Nobody has a strong method for that yet.

CONCLUSION.

The honesty argument deserves a serious look. It does not survive one. The diagnosis is right. Much bad policy is a fight, not a puzzle, and research on AI in government has said too little about that fight. The cure is wrong. A machine has no bank account. It has a goal somebody wrote, data somebody chose, tests somebody set, operators somebody appointed, and a contract somebody signed. Money will find the cheapest of those doors.

What remains is smaller and more useful. Build the tools that price every policy. Publish the gap between what a government said it wanted and what it did. That is not rule by machine. It is a transparency tool somebody could build next year, and then check whether it changed anything.

References

- [1] Adam, I. & Fazekas, M. (2021). Are emerging technologies helping win the fight against corruption? *Information Economics and Policy*, 57. doi:10.1016/j.infoecopol.2021.100950
- [2] Ebrahimi Afrouzi, A. (2025). Why AI shouldn’t govern us, even if it could. *AI & Society*. doi:10.1007/s00146-025-02501-9
- [3] Alon-Barkat, S. & Busuioc, M. (2022). Human–AI interactions in public sector decision making. *JPART*, 33(1). doi:10.1093/jopart/muac007
- [4] Argyle, L. P., Busby, E., Gubler, J. & Bail, C. (2023). AI chat assistants can improve conversations about divisive topics. arXiv:2302.07268. doi:10.48550/arXiv.2302.07268
- [5] Aydilek, E. (2026). Can artificial intelligence govern the state? *Chinese Political Science Review*. doi:10.1007/s41111-026-00346-1
- [6] Bakker, M. A., Chadwick, M. J., Sheahan, H. R. & Tessler, M. H. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. arXiv:2211.15006. doi:10.48550/arXiv.2211.15006
- [7] Bitonti, A. (2026). Designing deliberative lobbying. *The Political Quarterly*. doi:10.1111/1467-923X.70053
- [8] Black, S., Celniak, W., Garcia Huitron, A. et al. (2025). *Underpriced and Overused*. IMF Working Papers. doi:10.5089/9798229034715.001
- [9] Chomanski, B. (2022). Legitimacy and automated decisions: the moral limits of algocracy. *Ethics and Information Technology*, 24. doi:10.1007/s10676-022-09647-w
- [10] Coeckelbergh, M. & Sætra, H. S. (2023). Climate change and the political pathways of AI. *Technology in Society*, 75. doi:10.1016/j.techsoc.2023.102406

- [11] Creel, K. & Hellman, D. (2022). The algorithmic Leviathan. *Canadian Journal of Philosophy*, 52(1). doi:10.1017/can.2022.3
- [12] Cupać, J., Schopmans, H. & Tuncer-Ebetürk, İ. (2024). Democratization in the age of artificial intelligence. *Democratization*, 31(5). doi:10.1080/13510347.2024.2338852
- [13] Danaher, J. (2016). The threat of algocracy. *Philosophy & Technology*, 29. doi:10.1007/s13347-015-0211-1
- [14] Dono, M. & Moreno-Bella, E. (2024). Enter the matrix: psychosocial determinants of support for a technocracy of AI. *Political Psychology*. doi:10.1111/pops.13048
- [15] Erman, E. & Furendal, M. (2022). AI and the political legitimacy of global governance. *Political Studies*. doi:10.1177/00323217221126665
- [16] Gilens, M. & Page, B. I. (2014). Testing theories of American politics. *Perspectives on Politics*, 12(3). doi:10.1017/S1537592714001595
- [17] Greenstein, S. (2021). Preserving the rule of law in the era of AI. *AI and Law*, 30. doi:10.1007/s10506-021-09294-4
- [18] Grimmelikhuijsen, S. & Meijer, A. (2022). Legitimacy of algorithmic decision-making: six threats. *PPMG*, 5(3). doi:10.1093/ppmgov/gvac008
- [19] Huang, S., Siddarth, D., Lovitt, L. et al. (2024). Collective Constitutional AI. *ACM FAccT '24*. doi:10.1145/3630106.3658979
- [20] Huq, A. Z. (2021). Artificial intelligence and the rule of law. SSRN. doi:10.2139/ssrn.3794777
- [21] Jungherr, A. (2023). Artificial intelligence and democracy: a conceptual framework. *Social Media + Society*, 9(3). doi:10.1177/20563051231186353
- [22] Keppeler, F. (2023). No thanks, dear AI! *JPART*, 34(1). doi:10.1093/jopart/muad009
- [23] Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. (2015). Prediction policy problems. *AER*, 105(5). doi:10.1257/aer.p20151023
- [24] Kleizen, B., Van Dooren, W., Verhoest, K. et al. (2023). Do citizens trust trustworthy AI? *GIQ*, 40(4). doi:10.1016/j.giq.2023.101834
- [25] König, P. D. (2019). Dissecting the algorithmic Leviathan. *Philosophy & Technology*, 33. doi:10.1007/s13347-019-00363-w
- [26] König, P. D. (2023). Citizen conceptions of democracy and support for AI in government and politics. *EJPR*, 62(1). doi:10.1111/1475-6765.12570
- [27] König, P. D. & Wenzelburger, G. (2021). The legitimacy gap of algorithmic decision-making in the public sector. *Technology in Society*, 67. doi:10.1016/j.techsoc.2021.101688
- [28] Konya, A., Schirch, L., Irwin, C. & Ovadya, A. (2023). Democratic policy development using collective dialogues and AI. arXiv:2311.02242. doi:10.48550/arXiv.2311.02242
- [29] Koster, R., Balaguer, J., Tacchetti, A. et al. (2022). Human-centred mechanism design with Democratic AI. *Nature Human Behaviour*, 6. doi:10.1038/s41562-022-01383-x
- [30] Lazar, S. (2024). Governing the algorithmic city. arXiv:2410.20720. doi:10.48550/arXiv.2410.20720
- [31] Lorenz, L., Meijer, A. & Schuppan, T. (2021). The algocracy as a new ideal type for government organizations. *Information Polity*, 26(1). doi:10.3233/IP-200279
- [32] Nugmanova, A. & Gabdulina, B. (2026). Institutional trust and public support for AI-driven governance. *Transforming Government*. doi:10.1108/TG-03-2026-0125
- [33] Pickering, S. D., Hansen, M. E. & Sunahara, Y. (2025). Democracy by algorithm? *Parliamentary Affairs*. doi:10.1093/pa/gsaf050
- [34] Rahwan, I. (2017). Society-in-the-loop. *Ethics and Information Technology*, 20. doi:10.1007/s10676-017-9430-8
- [35] Sætra, H. S. (2020). A shallow defence of a technocracy of artificial intelligence. *Technology in Society*, 62. doi:10.1016/j.techsoc.2020.101283
- [36] Sætra, H. S. (2021). A typology of AI applications in politics. In *Advanced Sciences and Technologies for Security Applications*. doi:10.1007/978-3-030-88972-2_3
- [37] Savveli, I., Rigou, M. & Balaskas, S. (2025). From e-government to AI e-government. *Informatics*, 12(3). doi:10.3390/informatics12030098
- [38] Schiff, K. J., Schiff, D. & Adams, I. T. et al. (2023). Institutional factors driving citizen perceptions of AI in government. *PAR*, 83(6). doi:10.1111/puar.13754
- [39] Srinivasan, T. & Patapati, S. (2025). Democracy-in-Silico. arXiv:2508.19562. doi:10.48550/arXiv.2508.19562
- [40] Starke, C. & Lünich, M. (2020). Artificial intelligence for political decision-making in the EU. *Data & Policy*, 2. doi:10.1017/dap.2020.19
- [41] Straub, V. J., Morgan, D., Bright, J. et al. (2023). Artificial intelligence in government. *GIQ*, 40(4). doi:10.1016/j.giq.2023.101881
- [42] Tacchetti, A., Koster, R., Balaguer, J. et al. (2025). Deep mechanism design. *PNAS*. doi:10.1073/pnas.2319949121
- [43] Tallberg, J., Erman, E., Furendal, M. et al. (2023). The global governance of artificial intelligence. *ISR*, 25(3). doi:10.1093/isr/viad040
- [44] Tessler, M. H., Bakker, M. A., Jarrett, D. et al. (2024). AI can help humans find common ground in democratic deliberation. *Science*, 386(6719). doi:10.1126/science.adq2852
- [45] Wuttke, A., Jungherr, A. & Rauchfleisch, A. (2026). Artificial intelligence in government: why people feel they lose control. *JEPP*. doi:10.1080/13501763.2026.2696304
- [46] Zheng, S., Trott, A., Srinivasa, S. et al. (2022). The AI Economist. *Science Advances*, 8(18). doi:10.1126/sciadv.abk2607
- [47] Zuiderwijk, A., Chen, Y.-C. & Salem, F. (2021). Implications of the use of AI in public governance. *GIQ*, 38(3). doi:10.1016/j.giq.2021.101577