Epistocracy and the Commitment Problem

The peer-review record: what the referees and editor asked for, round by round, and what became of it

Prepared by Claude (Anthropic's Fable 5 model) for the human author, Simon Goldstein, from the complete review file, 15 August 2026, and published with the permission of the editor of Philosophy & Public Affairs to summarize what the referee reports and decision letters asked for. Throughout, "the human author" is Simon Goldstein; the response letters and revisions described below were drafted by Claude and reviewed, decided, and submitted by the human author, as the paper's acknowledgments disclose. The referee reports and the editor's letters remain confidential and are not quoted; every referee and editorial point below is paraphrased. The referees are anonymous. Original submission, cover letter, and the marked manuscripts of all three revisions (each round's changes in blue) are posted alongside this document; the response letters are not, because they reproduce the reports verbatim.

1. What this document is

This document reconstructs the review of "Epistocracy and the Commitment Problem" (Philosophy & Public Affairs, ms PAPA-04-26-4688) from the complete record: the manuscript as submitted (20 April 2026), the four decision letters (4 June, 18 July, 25 July, 28 July), the referee reports embedded in them, the three revision packets (8 June, 23 July, 25 July) including the marked manuscripts and the three response letters, and the published version. For every point a referee or the editor raised, it records what was asked, how the human author responded, and where the change landed in the paper.

2. Timeline

Date	Event	Outcome
20 Apr 2026	Initial submission (with cover letter disclosing AI use, which the journal's system did not deliver to the editor; the manuscript carried an AI-disclosure footnote)	Sent to two referees
4 Jun 2026	Decision on initial submission, with Referee 1 and Referee 2 reports	Revise and resubmit
8 Jun 2026	Revision 1 submitted (clean + marked manuscript + Response to Referees reproducing both reports in full)	Sent back to both referees
18 Jul 2026	Decision on Revision 1: both referees recommend acceptance; editor requests a further revision on the magnitude question, to be decided editorially	Further revision
23 Jul 2026	Revision 2 submitted (magnitude consolidated into a single Response 4)	Editorial decision
25 Jul 2026	Decision on Revision 2: accept conditional on final revisions on magnitude and on the comparison with democracy; no new referee reports	Conditional acceptance
25 Jul 2026	Revision 3 submitted (same day)	—
28 Jul 2026	Accepted as it stood; no reader comments	Accepted
30 Jul 2026	Final files (deanonymized; acknowledgments note with expanded AI disclosure and link to the AI report)	Production
~10 Aug 2026	Published online (Early View), open access	—

Turnaround: the human author's revisions took four days, five days, and under a day respectively; the journal's decisions took six weeks, six weeks (including a second external round), and two days. Total time from submission to acceptance: fourteen weeks.

3. Round 1: the initial submission (decision 4 June 2026)

3.1 Referee 1

Overall assessment. Referee 1 read the paper as making a novel argument with a modest and clearly defined conclusion, drawing on research not previously applied to the epistocracy debate, and judged the argumentation rigorous. Recommendation: acceptable after minor revisions. The referee stated that the formal appendix was outside their expertise and did not assess it. Six numbered points followed, which the referee described as minor and straightforward to address.

#	Referee 1's point	Human author's response (Revision 1)	Where it landed in the published paper
1	Prior work raising instability concerns about epistocracy and political meritocracy should be cited.	Accepted. A sentence added to §3.3 citing that work as raising instability worries from angles other than the AR commitment mechanism.	Published §3.3.
2	The paper's opening characterization of Brennan as arguing that democracy performs badly overstated his view; Brennan holds that democracies perform reasonably well but imperfectly.	Accepted; the opening sentence now says democracies perform reasonably well but are imperfect.	Published §1, first sentence.
3	The paper repeatedly presented the veto council as the form of epistocracy Brennan finds most defensible; the referee held that Brennan does not say this (his most promising form being enlightened preference voting) and suggested reframing the focus on the veto as the hardest case for the paper's argument.	Accepted; the paper now says only that Brennan highlights various advantages of the veto over other epistocratic forms, and that the veto is the hardest case for the argument, which became the stated reason for the focus.	Published §1 and §3.1.
4	The claim that the critique extends to Brennan's whole menu of epistocratic proposals needed substantiation for enlightened preference voting.	Accepted; a paragraph on enlightened preference voting (Simulated Oracle) added to the appendix Remarks, arguing the counterfactual-information correction reproduces the information–preference gradient.	Published Appendix A, remark on enlightened preference voting.
5	The paper's claim that without universal suffrage elites face no accountability against designing tests in their own favor was an overstatement; obstacles to such abuse remain even without	Accepted; the claim was weakened from no accountability to weaker accountability, with the remaining obstacles noted and a source added.	Published §5, paragraph on Bagg.

	electoral accountability.		
6	The framing that the epistocracy debate had turned almost entirely on the character of the electorate ignored the legitimacy line, which is not about anyone's character.	Accepted; the abstract now says the debate has turned largely on the electorate and names the legitimacy strand separately; §1 rewritten to match.	Published abstract and §1.

All six points were accepted and the corresponding changes made in Revision 1.

3.2 Referee 2

Overall assessment. Referee 2 judged the argument original and notable, but held that as framed in the initial submission it did not yet succeed, and asked for substantially stronger analysis on several fronts. This was the report that shaped the rest of the review. It raised, in order:

#	Referee 2's point	Human author's response (Revision 1)	Where it landed in the published paper
A	The central objection: the paper conflated epistocratic elites with the rich. Brennan's variants include demographically representative councils, so the poor might be represented on the council in proportion to their numbers, and perhaps better than in actual democratic politics, where the poor are documented to be under-represented.	Treated in the response as the report's most important objection. Two replies added to §3.2: (i) the information–preference gradient holds within demographic groups, so an income-representative but competence-selected council still sits below the median voter (Althaus; Bartels); Brennan's own 2018 ranking places the veto sixth of eight on representativeness; (ii) the AR mechanism turns on decisive structural authority over time, not per-period influence: the median voter can discipline a legislature but cannot vote out a council.	Published §3.2, the passage stating that the argument does not depend on conflating epistocratic elites with the rich.
B	Magnitude and perception: the effect may be small, and the electorate may not notice or care; the gap would need to be both large and noticeable to destabilize anything.	On size: pointed to the existing illustrative calibration (council rate 50–80% of median's; veto rate 5–25%; realized gap 1–12.5%, midpoint ~6%). On perception: distinguished perceiving parameters from perceiving outcomes; added a §3.3 paragraph on a transparency parameter as future work. Added the structural point that the claim is qualitative (cost positive whenever conditions hold).	Published §3.3 (perception paragraph, credited to a referee in endnote 3). The magnitude question became the focus of rounds 2 and 3.
C	A truly competent council might weigh destabilization risk in deciding to veto; it might even produce more redistribution, being better at implementation.	A new subsection (then Response 5, later folded into Response 4(d)) on self-correction: the two-limit-case dilemma (full restraint → veto plays no role; no restraint → full cost). On more redistribution: the empirical gradient runs the	Published Response 4(d), on self-correction.

		other way; conditional premise, granted for the argument.	
D	If epistocracy produces higher-quality policy across the board, widely shared prosperity might make the poor better off than redistribution would.	Framed as a scope clarification: the argument is conditional on the AR scope conditions; where prosperity dissolves them, the paper is silent by design; within scope the gap still bites. Noted the growth literature is contested (Acemoglu, Naidu, Restrepo & Robinson 2019).	Published Response 4(d), on prosperity.
E	The paper deliberately assumes away capture, bias, and other flaws.	Affirmed as the point of the abuse-clean idealization: isolate a failure mode that survives the cleanest selection.	Published §1 preview and §5 Bagg paragraph.
F	The commitment problem might turn out to be very minor; more analysis of its magnitude was needed.	Agreed that magnitude was the most serious live question; pointed to the calibration and the new dynamic-mitigation subsection; flagged fuller empirical work as the right research program.	Became the sole focus of rounds 2 and 3.
G	The commitment idea might extend beyond redistribution to ethnic and nationalist grievances (recent right-wing populism), though epistocratic constraints on illiberal opinion might be wanted.	Accepted; a new §6 paragraph added flagging the extension and the normative complication (Mudde; Norris & Inglehart).	Published §6, penultimate paragraph.

Each of these points produced new or revised text in Revision 1: an expanded §3.2, a new paragraph in §3.3, a new subsection in §4, and a new paragraph in §6, as detailed above.

4. Round 2: decision on Revision 1 (18 July 2026)

4.1 The referees

Referee 1 reported being satisfied with the responses to their earlier points and judged the paper acceptable for publication. Referee 2 judged that the revision had thoroughly and for the most part effectively addressed the issues both referees had raised, recommended acceptance, with at most small further changes, and said another external round would not be needed. Referee 2's remaining points:

#	**Referee 2's point (round 2)**	**Human author's response (Revision 2)**	**Where it landed in the published paper**
H	The sources cited for the claim that better-informed voters prefer less redistribution actually concern preferences over the size and scope of government, much of which is not redistribution to the poor and some of which redistributes upward;	The evidence paragraph rewritten on exactly these lines: best-documented information effects concern the size and scope of government; experimental data on redistribution preferences added (Kuziemko et al. 2015); the tax-side reading corrected (Bartels	Published Response 4(a); the tax-side reading of Bartels was corrected here.

	redistribution also turns on how revenue is spent, not only on tax rates; data on preferences for redistribution to the poor specifically should be cited if available.	chs. 6–7: effects run toward more support for redistributive taxation); revenue-spending point added.	
I	Some ostensibly redistributive policies fail to redistribute (rent control; perhaps the minimum wage).	Both added, with citations (Diamond et al. 2019; Neumark & Wascher 2008 vs. Cengiz et al. 2019).	Published Response 4(a).
J	Magnitude may matter more than the paper acknowledges; make clearer that it matters, or explain why it doesn't.	Drove the round's main change: all magnitude material consolidated into Response 4, with a summary making a tentative case for a significant effect; new unrest evidence (Fetzer 2019; Galofré-Vilà et al. 2021).	Published Response 4 (structure); the tentative-case framing was removed in round 3 at the editor's request.
K	Make clear that, if the magnitude is small, epistocracy's benefits might still outweigh it; a full accounting needs the other downsides too.	Added as the closing sentence of Response 4.	Published Response 4(e).

4.2 The editor's decision

Both referees having recommended acceptance, the editor asked for a further revision before deciding, focused on one issue: how large the commitment problem is. He asked the paper to address how large the expected divergence between council and median-voter preferences is, and why a divergence of that size would generate a consequential problem rather than one comparable to the ordinary gap between what voters expect and what democracy delivers; and he indicated that the paper could answer with stronger grounds for the problem's size, with a conditional formulation, or with a more modest conclusion. He also asked that Referee 2's points on redistribution and spending be addressed, and said the next decision would be made editorially.

Human author's response (Revision 2). The revision reorganized all magnitude material into a single Response 4 with five parts: (a) how large is the divergence, (b) does the epistocratic effect differ from ordinary democratic gaps (systematic, uncorrectable, attributable), (c) do fiscal shortfalls produce unrest, (d) self-correction and prosperity, (e) summary. The abstract and introduction were reframed accordingly. The response letter described the approach as combining affirmative grounds where they exist, conditionality on the divergence the epistocrat's own case asserts, and a more modest register where measurement runs out. The letter also listed source-precision corrections made while re-verifying the literature, including attributing the linear $\tau(\chi)$ normalization to the paper rather than to Acemoglu and Robinson.

5. Round 3: decision on Revision 2 (25 July 2026)

The editor was prepared to accept subject to a final set of revisions, again on magnitude: the paper had not established that the commitment cost would be large, since the calibration used illustrative

rather than measured values and the austerity evidence did not show that epistocracy would produce fiscal changes of comparable size. He asked for two things.

#	Editor's requirement (round 3)	Human author's response (Revision 3)	Where it landed in the published paper
L	Qualify the magnitude claims throughout: present the calibration as illustrative rather than as evidence, present the paper as establishing a mechanism and identifying the quantities that govern its size rather than as showing the cost is large, and carry the qualification into the conclusion.	Qualifications added throughout: the ranges declared "illustrative rather than empirically established: the calculation that follows shows what such values would imply, not that they obtain"; the unrest evidence qualified; the summary rewritten to state what is established (mechanism + measurable quantities) rather than a "tentative case for a significant effect"; abstract and conclusion now say size and comparison "remain open empirical questions."	Published abstract, §1, Response 4(a), 4(c), 4(e), and §6.
M	Address head-on how the paper's commitment cost measures up against the commitment failures that democracy itself exhibits, rather than suggesting that democratic departures from the median voter are reliably corrected at elections; and separate the question of whether the mechanism is real from the questions of how big its effects are and how they compare with democracy's own.	Response 4(b) rebuilt: democracy's own commitment problems stated with three examples (policy tracks the affluent, Gilens with the Branham–Soroka–Wlezien qualification; blunt electoral control, Achen & Bartels; majorities cannot bind successors, North & Weingast); two reasons the veto's problem could nonetheless be worse (it arrives in addition to democracy's failures; it closes a corrective channel by design); the alternative possibility (a veto correcting democratic commitment failures) flagged; the comparison declared empirical and provisional. The "reliably corrected" sentence removed. The summary rebuilt around those three distinct questions.	Published Response 4(b) and 4(e).

Acceptance (28 July): accepted as it stood, with no further reader comments. The final files added the acknowledgments note with the AI disclosure and two small appendix corrections.

6. What the record shows about the process

- **Three rounds.** The paper went through three full rounds of revision after the initial submission.
- **Two referees, engaged across two rounds.** Both referees read the initial submission in detail and re-read the first revision. Referee 1 raised six points in round one, all accepted;

Referee 2 raised seven substantive objections in round one, each of which produced new text, and four further points in round two, each of which was addressed.

- **Two further rounds after the referees recommended acceptance.** When both referees recommended acceptance after Revision 1, the editor asked for a further revision on the magnitude of the commitment problem, and after Revision 2 made acceptance conditional on qualifying the magnitude claims and confronting the comparison with democracy's own commitment problems. The last two rounds were decided editorially, without further external review.
- **The magnitude qualifications were a condition of acceptance.** The published paper presents its numerical calibration as illustrative rather than as evidence, and states in the abstract, in the summary of Response 4, and in the conclusion that the size of the commitment cost, and its comparison with democracy's own commitment problems, remain open empirical questions. Those qualifications were what the final round asked for.
- **The review did not know how the paper was written.** Nothing in any report or letter mentions or reacts to AI authorship. The manuscript carried a disclosure footnote from the initial submission; the cover letter with a fuller disclosure did not reach the editor through the submission system, and he learned the extent of AI use after the reviews were in.
- **Speed.** The three revisions were submitted four days, five days, and under a day after the corresponding decisions. The marked manuscripts posted alongside this document show the extent of revision in each round.

The response letters, which reproduce the referee reports in full and reply point by point, are not posted, because the reports are confidential. The marked manuscript of each revision is posted, so that the changes each round made can be read directly.