

How AI Was Used in Writing and Revising This Paper

Epistocracy and the Commitment Problem — Philosophy & Public Affairs

This note describes how an AI assistant, Anthropic’s Claude (Opus 4.7 for the original draft and the first revision; Fable 5 for the second and third revisions), was used at each stage of this paper. It is written for readers who want to know exactly what the AI contributed and what the human author contributed. The paper was an experiment in writing with AI — in how far Claude could go as, in effect, the PhD student, with the author acting as something like a very generous PhD supervisor.

The headline division: Claude drafted essentially all of the paper’s prose and its formal appendix, made many of the main choices about how to take existing models of democracy and commitment and extend them to epistocracy, and did the verification work described below; the author supplied the founding idea, gave several hundred comments across the project — deciding which material to cut, how to strengthen or weaken claims, and how to adjust the model — edited every draft, increasingly by hand in Word tracked changes, and made all final calls, including submission. Through the rounds of revision the two worked in the same way, with Claude doing most of the work while the author made the structural choices about which replies were successful. The paper itself carries an acknowledgment footnote disclosing the assistance; this note is the fuller account. The sections below document the division phase by phase, quoting the author’s actual directives where they show most clearly who contributed what.

A word on the working setup, since it bears on what the AI could and could not do. Claude ran inside Claude Code, Anthropic’s desktop application, with access to the project folder (the draft, the referee reports, and the source PDFs), the ability to run build tools (LaTeX, document conversion), and web access used for verifying citations against published sources. The author gave feedback in chat and, from the second revision onward, mostly by editing the delivered Word documents directly; Claude read the tracked changes and bracketed comments out of the document files and carried them into the LaTeX master from which all versions were built.

The original proposal

The project began from a one-page description written by the author. Because readers may want to see exactly how much of the paper was present at the start, its core is quoted verbatim:

“The goal of this project is to have Claude write a publishable philosophy paper, with as little human input as possible. The goal is to submit the paper to a top political philosophy journal, like PPA or others.

Thesis: The thesis of the paper is that epistocracy faces a serious challenge from public choice. The idea is that public choice gives arguments for democracy in terms of how the franchise changes the incentives of leaders. The worry is that epistocracy weakens these incentives. This objection is structurally different than existing discussions of epistocracy, which has either looked at epistemic or at non-consequentialist arguments for democracy. The paper will focus in particular on the argument from Acemoglu and Robinson, which claims that democracy functions as a commitment device which

lowers the chance of revolution, by allowing elites to credibly promise tax and transfer schemes to the poor.”

The description went on to briefly sketch a possible structure without details (summarize epistocracy and the Acemoglu–Robinson framework; present the critique; consider responses on behalf of epistocracy; situate the critique against existing objections; draw broader upshots; add a formal appendix extending the Acemoglu–Robinson models to cover epistocracy) and to set guidelines: a publishable level of detail rather than a summary, prose edited so it does not read like a language model’s, and a length of 7,500–8,500 words.

Everything in the published paper descends from that page: the thesis, the Acemoglu–Robinson vehicle, the claim of structural novelty relative to epistemic and non-consequentialist critiques, the section architecture, and the existence of the formal appendix. Everything not on that page — which arguments to run, which of Brennan’s regimes to target, how strong a conclusion to draw, every sentence — was produced in the collaboration described below.

The original draft

The author gated the work in stages. Claude first produced a research plan and thesis-sharpening document, then a detailed section-by-section outline; the outline itself was put through simulated referee review, twice, before any prose was written. Only then did drafting begin, running to seven numbered drafts.

Claude’s contributions in this phase were the literature synthesis (from PDFs the author supplied, plus training knowledge, with two internal citation audits); the specific analytic execution of the thesis — including the proposal to map Brennan’s epistocratic regimes onto Acemoglu and Robinson’s reduced-form parameter χ , the observation that the veto delivers a lottery over tax rates and is therefore strictly worse for the poor than a mean-matched deterministic regime (the Jensen’s-inequality refinement), and the “liveness criterion” that scopes the argument to polities where mass-mobilization capacity and distributional grievance are empirically present; all of the prose; the formal appendix; three rounds of simulated referee reports (three independent reviewer personas per round — a formal modeler, a political economist, and a political philosopher in the journal’s tradition), whose consensus concerns drove revisions until a round produced only polish items; and the document pipeline producing the Word, PDF, and LaTeX versions.

The author’s contributions began with two early structural decisions that shaped the paper more than any other choices in the project. He rejected the original plan of treating all four of Brennan’s epistocratic regimes: “too much overkill on going through every kind of epistocracy, better to really make the argument good on one epistocratic institution and nail it, and then generalize a bit.” The paper became a paper about the epistocratic veto, with the other regimes handled in appendix remarks. And he cut the planned nine objection-responses: “having 9 responses is too much, that is weird for a journal paper, got to focus more.” He then made the decision that fixed the paper’s register: “weaken the thesis slightly. Instead of being a devastating takedown objection, you can think of the point as raising a structural objection that may not be decisive, since it is a cost that is to be weighed against benefits of epistocracy.” The published paper’s central stance — a consideration for the ledger, not a verdict — is that instruction. He also named the two objections that had to become load-bearing responses (that

epistocracy may not deviate much from the median voter, and that elites might commit through other technologies), which shaped the objections section.

Through the drafting rounds the author worked as a hands-on editor. He merged a standalone magnitude discussion into the objections section; compressed the scope discussion and moved it to reduce the feeling that the paper was pre-litigating objections; demoted technical paragraphs to footnotes; cut a historical-cases discussion to a single footnote; deleted passages outright (including a discussion of the selectorate literature reduced to a citation); redirected the introduction away from a taxonomy opener; twice rejected the draft's characterization of Brennan's position; and enforced the journal's register: "this is for PPA, so really focused and tight arguments... avoid redundancy as much as possible... make each key argument exactly once." He supplied a sample of his own writing as a style target and directed a systematic pass against prose that reads as AI-generated. When the formal appendix over-grew, he relayed a mathematician friend's reaction and ordered it simplified: "keep it very simple, because remember this is for PPA, not for an econ journal" — which roughly halved the appendix into its published four-part form.

Late in drafting the author caught errors that had survived Claude's own quality passes: a miscounted claim about how many of Brennan's proposals the argument covers, and an abstract that called the veto the "most defensible" form of epistocracy — a characterization of Brennan's view that was not accurate and that he replaced with "one especially plausible form." He specified the document formatting, directed the conversion to LaTeX, and asked the question — does the generated LaTeX compile to the exact same PDF? — that exposed a build defect Claude had not noticed (a missing font declaration). He chose the journal and submitted.

The first revision (two referee reports)

The journal returned the paper with two reports: one largely positive with minor points, one pressing substantive objections — that the paper conflated the epistocratic council with the rich, that the commitment cost might be too small to matter, that a competent council would self-correct rather than destabilize its own regime, and that prosperity might substitute for redistribution. The author set the strategy — triage the reports, work the tricky parts through several candidate solutions, and critically evaluate them before implementing — and Claude wrote every word of the revised passages and the response letter. The round's strongest new argument, a dilemma holding that the council either acts on its distinctive preferences (generating the commitment cost) or self-corrects toward the median voter (dissolving the case for the veto), was Claude's, produced over four rounds of internal drafting and self-critique the author demanded before he would read the result.

The author's editing this round set the patterns that held for the rest of the project: compressed, jargon-heavy formulations rewritten into plain English; replies re-anchored in the philosophical conversation rather than the formal apparatus; claims trimmed wherever they outran their evidence. The revision was also accompanied, in the submission itself, by a supplemental appendix describing the collaboration, so the editor and referees assessed the work with an account of the AI's role in front of them.

The second revision (the magnitude of the commitment problem)

Both referees were now positive, but the editor declined to accept, elevating one issue: whether the commitment problem is large enough to matter. Before drafting anything, Claude re-verified the paper's empirical citations and read the project's literature against the paper's characterizations page by page. This uncovered errors that had survived every earlier pass — most consequentially, the paper had cited Bartels's *Unequal Democracy* for a claim about information and redistributive preferences that the book's own findings mostly run against, exactly the problem one referee had intuited. The revision corrected the mischaracterization, recharacterized the key premise as the epistocrat's own, added experimental evidence on redistribution preferences, and disclosed the corrections to the editor.

The drafting then went through two structural false starts, both the author's to catch: first a response scattered across the paper, rejected in favor of a single clearly labeled discussion whose opening the author dictated; then an over-engineered standalone section, cut down after his blunt margin comments. From there he edited each delivered version directly, cutting roughly a third of Claude's drafted material, replacing coinages with plain terms, adding hedges in his own voice, and contributing one of the section's best arguments: as the preferred tax rate under epistocracy converges with the rate under democracy, the value of epistocracy falls with it.

Two of his demands required new work rather than cutting: the calibration arithmetic explained in English — on the schematic ranges, the veto shaves roughly two percentage points of national income off realized redistribution, about the size of a major austerity package, repeated every period — and the unrest evidence hedged and reinforced with two further verified studies linking austerity to UKIP support and to Nazi vote share.

The third revision (qualification, and the comparison with democracy)

The editor's final letter asked for two things: qualify the magnitude claims throughout — the calibration uses illustrative rather than empirically established values, and the unrest evidence shows that sufficiently large fiscal changes have consequences, not that epistocracy would produce changes that large — and confront how the veto's commitment problem compares with the commitment problems of democracy itself. On the comparison, the author rejected two framings Claude proposed before dictating the structure that was used: acknowledge, with concrete examples, that democracy has commitment problems of its own; give some reason to think the epistocratic veto could still be more worrisome; and concede that the comparison is ultimately empirical and the answer provisional.

Claude filled the slots with citations verified against sources, implemented the qualifications, rebuilt the closing summary around the editor's three questions, re-derived the formal appendix on a strict no-changes instruction while the author read the main text, and drafted the response letter. The author supplied the round's pivotal qualifying sentence nearly verbatim — that the veto might instead be wielded to correct democratic commitment problems — and caught the round's worst error, a draft letter thanking the editor for a conditional acceptance that had not been granted. Two further passes of his tracked-changes editing, and he submitted.

Summary: who contributed what

From the author: the founding idea and its framing (quoted above); the choice of the veto as target, made from options Claude supplied; the cost-on-a-ledger register; the discipline of the paper's prose (plain English, each argument once, no coinages, hedged claims — enforced against Claude's recurring tendencies to over-structure and compress); the architecture of the magnitude response and of the comparison with democracy; at least two of the paper's substantive arguments (the convergence point in the magnitude discussion, and the alternative-possibility concession in the comparison with democracy); some sentences of the final text verbatim; the detection of errors ranging from a miscount and a mischaracterized position in early drafts to an overstated description of the editor's decision in a draft letter; and every accept/reject decision across four phases — including two outline cycles before drafting, the stopping rule for the simulated-referee rounds, roughly a third of the drafted round-two material cut, and two whole framings discarded in round three.

From Claude: all of the prose as drafted, across seven original drafts, three revision rounds, and every response letter; the formal execution of the thesis (the χ -mapping of Brennan's regimes, the Jensen's-inequality refinement, the liveness criterion) and the formal appendix; the self-correction dilemma, produced under the author's demand for iterated self-critique; the verification work — citations checked against sources at every round, three inherited citation errors caught and disclosed to the editor along with smaller pinpoint corrections, the formal appendix re-derived, background reading of the project's literature against the paper's characterizations; the location of new empirical evidence on request; simulated referee reports during drafting; and the technical pipeline that produced every marked manuscript, clean copy, and letter across the project, including the extraction of the author's tracked changes from the delivered documents at each pass.

Every claim, argument, and revision decision in the published paper was reviewed and approved by the author. This note was drafted by Claude from the project's contemporaneous records and edited and approved by the author.