Cauce: Topic-Delegable Liquid Democracy with Reputation-Weighted Delegates

Angel Zeledón Fernández
angel.zeledon.fernandez@gmail.com

October 1, 2026

Abstract

Majority rule counts a ballot cast on a belief the evidence has already ruled out — our running example is the claim that 5G infrastructure causes disease — exactly like an expert's. Liquid democracy lets citizens vote directly or delegate revocably, but delegation concentrates power and need not beat direct voting. We present *Cauce*, a topic-scoped liquid-democracy design in which every citizen keeps an equal vote that can override any delegate, and every delegate publishes a Brier-scored record of falsifiable predictions. We prove its delegation-resolution algorithm well-defined and terminating; show that, in expectation, a Brier score ranks calibrated, better-than-chance delegates exactly as the optimal (Nitzan–Paroush) weighted majority would, while a confidently wrong delegate scores worse than one who says "I don't know"; and prove a large-population limit theorem that makes the central trade-off exact: when delegates are pivotal, delegation lowers accuracy if the unaffiliated population is competent and raises it if that population is misinformed. We then make the result non-asymptotic, with exponential and Berry–Esseen error bounds and a bound in terms of the effective number of independent voters, and extend it to citizens who vote through AI agents (Section 11). Extending the theorem to correlated misinformation shocks, we derive the critical susceptibility at which direct democracy collapses. A cross-validated implementation matches every prediction to within 1.6 percentage points. With a misinformed population (competence 0.35), reputation-weighted delegation is correct on 83.5% of bills, against 0% for direct democracy and 27.7% for popularity-driven delegation; with a well-informed one (0.65), direct democracy is correct on 100% and reputation-weighted delegation on 99.0%. Cauce is therefore best understood as *insurance* against population-wide misinformation, and the reputation signal is what makes it affordable: it cuts the premium paid in normal times from 37 points to 1, so that delegation outperforms direct democracy once about 2% of bills (95% CI 1.5–2.7%) carry a shock that misleads 30% of voters — provided track records are long and accurate. We then show, by proofs and simulations, what the reputation layer and the cap cannot guarantee: records need dozens of resolved predictions to separate delegates whose competence differs by fifteen points, and hundreds when it differs by five; author-chosen claims and competition for delegations make raw scores gameable and reward overconfidence; an averaged record can rank first the delegate who is worst in a shock; competence is not alignment with a citizen's ends; one to three bribed delegates can reverse a typical bill that would take hundreds of bribes under direct democracy; and at the default cap, any two truncated delegates together hold a majority of the counted vote. An analysis of ten policy sectors — economy, health, education, defence, technology, climate, criminal justice, migration, infrastructure and pensions — with simulations of how record length changes the insurance, suggests where it is worth buying — checkable questions in economics and public health — and where it is not: long-run education effects, curriculum, the use of force and decisions about membership. A section on an automated society shows that the safe zone against AI-agent monoculture has the same threshold as the rumour shock, and that what matters is the effective number of independent AI providers, which correlation across

models reduces. A plain-language guide with illustrations opens the paper. We close with the strongest counter-evidence we found, including legal and security obstacles, and compare Cauce with mechanisms aimed at the same failure.

Cauce in Plain Words: A Guide for the General Reader

Nuestras vidas son los ríos
que van a dar en la mar...
(Our lives are the rivers
that flow into the sea...)
— Jorge Manrique, *Coplas por la muerte de su padre*, c. 1476

This paper is written for specialists, but its question belongs to everyone: *when a whole country must decide something, whose judgement should count, and how much?* This guide tells the story without equations. Each main mathematical result, or group of related results, is also followed by a short box headed "In plain words".

A river and its channels. *Cauce* is Spanish for a riverbed: the channel along which water runs to the sea. Think of each citizen's vote as water that must reach the sea, where the decision is taken. In an ordinary referendum, every drop runs straight to the sea on its own. In a parliament, the drops are poured into a few large canals, chosen once every few years, which run wherever they run until the next election. *Liquid democracy* sits in between. Each citizen may let her vote run straight to the sea, or pour it into the channel of someone she trusts — a nurse she trusts on health, a farmer she trusts on agriculture — and she may redirect it at any time. Figure 1 draws the idea.

The weakness of liquid democracy is the weakness of any river: water follows the widest channel, not the one that leads somewhere good. In practice, votes flow to the famous and the popular, and a few people end up carrying more than a hundred votes each. Cauce's central addition is this: every channel carries a public record of where it has led before. Before trusting someone with her vote, a citizen can see how well that person's past predictions, and the confidence she put on them, matched what happened.

The grammar of a vote. Grammar offers a second way to see the design. A vote can be spoken in three ways:

- In the *first person*, "I vote": a direct ballot.
- In the *third person*, "she votes for me": a delegation, given topic by topic.
- In the *second person*, as a command to the delegate, "stand aside — I'll say this one myself": the override, which lets any citizen take back her vote on a single bill without cancelling her delegation on the others.

A delegation also has a tense. It is a present that expires: unless renewed, "she votes for me" becomes "she voted for me" after twelve months, and the citizen's vote stops counting on that topic until she chooses again. One rule of grammar is never broken: every sentence begins with a citizen. Each citizen starts with exactly one vote and the platform never adds to it. A delegate carries only the votes citizens choose to send her, and never more than the cap allows.

Why crowds are wise, and when they are not. In 1785 the Marquis de Condorcet noticed something remarkable about majority voting. If each voter is right a little more often than a coin toss, a large group is almost always right: with voters who are right 55% of the time, a majority

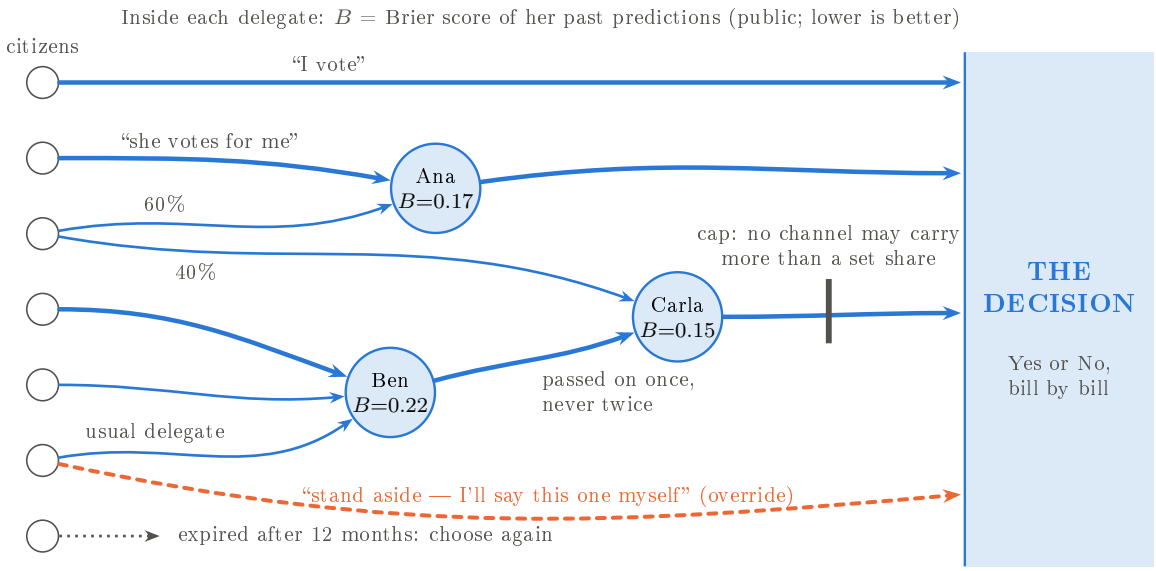


Figure 1: The river of votes. Each citizen's vote (left) either runs straight to the decision ("I vote"), or flows through a delegate she trusts on that topic ("she votes for me"), possibly split between two delegates. A delegate may pass the vote on once more, if the citizen allows it, but no further. Each delegate shows a public record of past predictions; the number is a Brier score, where lower is better. A cap stops any one channel from carrying too much (the excess is not counted), delegations dry up after a year unless renewed, and any citizen can step back into the river on any single bill ("stand aside — I'll say this one myself").

of 101 is right 84% of the time, and a majority of 301, 96%. But the same arithmetic works in reverse. If a falsehood has convinced people so that each is right only 45% of the time, the majority of 301 is right only 4% of the time: the crowd amplifies whatever tilt its members have, including a tilt toward a falsehood (Figure 2). This is the central dilemma of the paper. When people are well informed, letting everyone vote is very hard to beat. When a rumour — the paper's running example is the false claim that 5G phone masts spread disease — has captured much of the public, the crowd becomes confidently wrong. A well-informed minority can then rescue the decision, but only if it carries enough of the vote and is not fooled itself.

The weather forecaster. How do we know who is well informed? Cauce asks every delegate, each time she votes, to attach one checkable prediction with a probability: "this law will keep unemployment flat within a year, with probability 70%". When the year is up, the prediction is scored. The score, called the Brier score, was invented to grade weather forecasters [Brier, 1950]: if a forecaster says "70% chance of rain" on many days, it should rain on about seven of every ten. Figure 3 shows why the score is fair. You cannot improve it on average by exaggerating or by hedging; the best strategy is to say exactly how sure you are. Someone who always says "50%" scores 0.25. Someone who is usually wrong but always sounds certain scores worse than that — worse than admitting "I don't know". A loud voice that keeps getting things wrong stands out on the record.

Insurance, not a miracle. The simulations in this paper give an honest answer, and it is not the one we hoped for when we began. In our simulations, when the public is well informed, letting everyone vote is slightly better: delegation in which citizens choose delegates by their records gets 99 bills in 100 right, against 100 for direct democracy. But on a bill where a rumour misleads 30% of

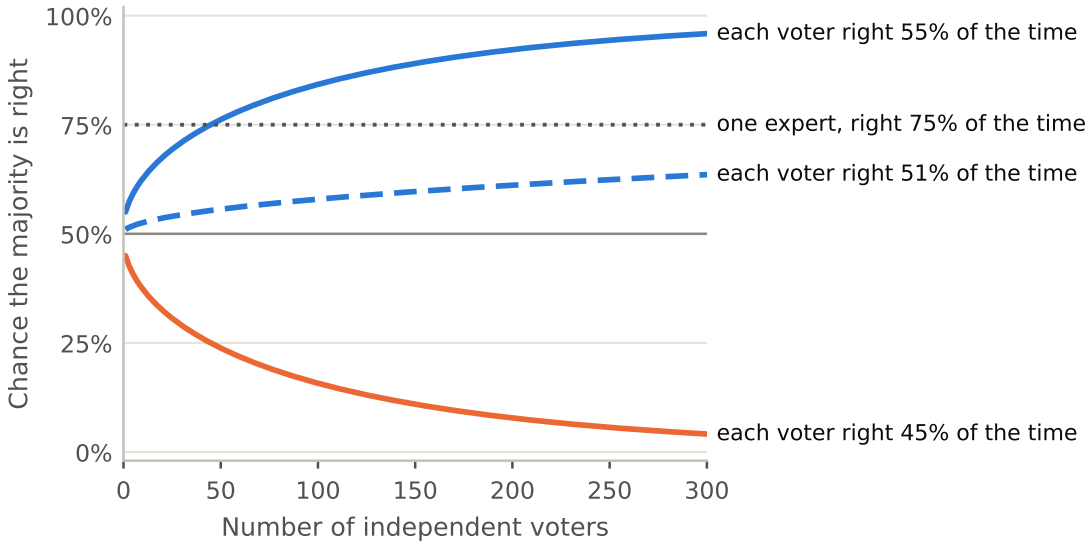


Figure 2: Condorcet's arithmetic: the chance that a majority of independent voters is right, as the crowd grows. A crowd of voters who are each right a little more than half the time becomes almost always right as it grows, and overtakes a lone expert who is right 75% of the time once it is large enough (about 45 voters at 55%, over 1,100 at 51%). A crowd of voters who are each right a little less than half the time becomes almost always wrong.

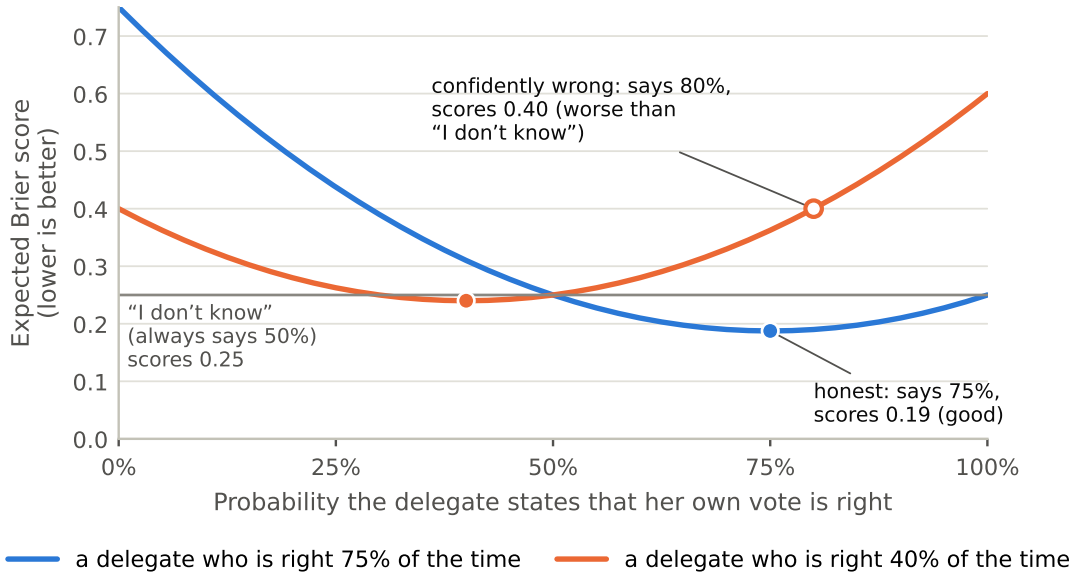


Figure 3: The Brier score rewards saying how sure you are. Each curve shows the average score a delegate earns for each probability she might state that her own vote is right. An honest delegate scores best by stating her true accuracy (filled dots). A delegate who is right only 40% of the time but claims 80% (hollow dot) scores worse than one who always says 50%.

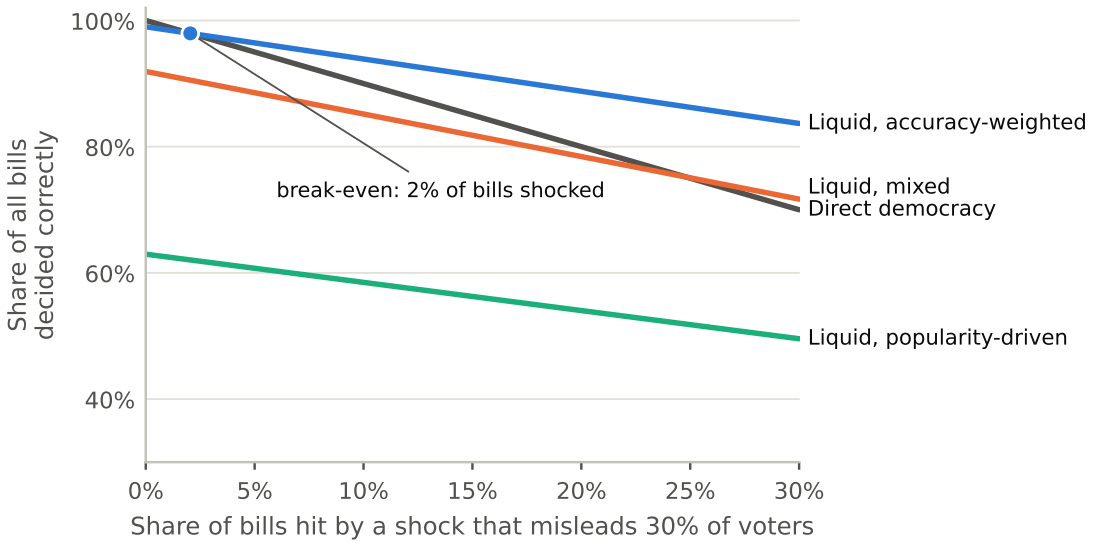


Figure 4: Delegation as insurance. Each line shows the share of all bills decided correctly as the share of bills hit by a misinformation shock grows (each shock fools each voter with probability 0.3, the same for every voter). "Accuracy-weighted" is reputation-weighted delegation; "mixed" means citizens partly follow the record and partly popularity; "popularity-driven" means they follow popularity only. Reputation-weighted delegation overtakes direct democracy once about 2% of bills are shocked. The vertical axis starts at 30%. Numbers from Table 10.

voters, direct democracy gets every one wrong, while record-based delegation still gets between 48% and 63% right. These are simulated citizens who follow the records; whether real citizens would is untested. Cauce therefore works like an insurance policy. You pay a small premium in good times, about one bill in a hundred if records are long, to be covered in bad times. As Figure 4 shows, the policy pays for itself if more than about two bills in a hundred are hit by such a rumour (never, for a rumour that fools fewer than about one voter in four; only at many more bills, for a stronger one). Delegation driven by popularity, by contrast, is an expensive policy that almost never pays out: it gets more than a third of ordinary bills wrong.

What the score cannot do. Much of the paper is about the limits of this idea, and they matter as much as its promise:

- *Records take time.* About ten predictions are enough to expose someone who is confidently wrong. Telling two good delegates apart takes dozens of predictions if one is clearly better, hundreds if they are close, and thousands if they are very close — which can take years.
- *Records can be gamed.* Anyone can look like a prophet by predicting that the sun will rise. Delegates must therefore be scored against what others predicted for the same question.
- *Competition breeds bluster.* When followers go to whoever tops the table, it pays to sound more certain than you are.
- *Knowing is not wanting.* A delegate can read the facts superbly and still want a different outcome from yours. The record shows the first, not the second.
- *Citizens may not use the record.* Everything above assumes people follow the track record and delegate sparingly. Real people may follow confidence or fame instead, and nobody has

Table 1: A glossary for the general reader.

Term	Meaning
Delegation	Asking someone you trust to vote for you on one topic. You can take it back at any time.
Override	Voting yourself on a single bill, which sets your delegation aside for that bill only.
Topic	A policy area, such as health or taxes. Delegations are given topic by topic.
Cap	A limit (25% by default) on the share of the participating vote that any one delegate may carry. Weight above the limit is not counted; it is not handed to anyone else.
Hop limit	The most delegates a vote may pass through. The default is two: the delegate you chose may hand your vote on to one further delegate, if you allow it, and no further.
Expiry	A delegation lapses after twelve months unless it is renewed.
Brier score	The average squared gap between the probability someone stated and what happened. 0 is perfect and lower is better; always saying 50% scores 0.25.
Calibration	Saying 70% about things that happen about 70% of the time.
Competence (θ)	How often a person's vote lands on the side she would choose if fully informed.
Pool	The many ordinary citizens who vote for themselves, together with the share of weight others hand to small, unnamed delegates; each is treated as an independent voter.
Shock (σ, π_{shock})	A falsehood that sweeps through many minds at once. σ is the chance it fools a given voter; π_{shock} the share of bills it hits.
Premium	The accuracy given up in ordinary times by routing votes through delegates.
Break-even (π^*)	The share of shocked bills above which delegation pays for its premium.
Alignment (α)	How likely it is that, were both of you fully informed, she and you would choose the same side.
Relative score	A Brier score measured against what others predicted for the same question, so that easy questions earn no credit.
Capture	Bribing, coercing or otherwise controlling the few people who carry many votes.

tested it.

- *Few hands are easier to buy.* A briber who needs hundreds of voters under direct democracy may need only one to three delegates.

Different laws, different rivers. Laws are not all alike. Some claims about vaccines or interest rates — an infection count, a monthly inflation figure — can be checked within weeks or months, although the long-run effects of such laws take years. A law on school curricula makes claims that take decades to check, claims about deterrence cannot be checked at all, and disputes in both are more about values than about facts. The paper compares ten sectors — the economy, health, education, defence, technology, climate, criminal justice, migration, infrastructure and pensions. It finds Cauce most useful for checkable questions inside the economy and health, and least useful for curricula, the use of force, and decisions about who belongs (Figure 13).

How to read the rest. Section 3 describes the platform. Sections 4–6 give the core mathematics (Section 6.1 makes it precise for finite crowds), with plain-words boxes (more results, each with its box, appear in Sections 8, 9, 10 and 12). Sections 7–8 report the simulations, Section 9 the limits, Section 10 the sectors, Section 11 what changes if machines do much of the work and some of the voting, and Section 12 the strongest evidence against the idea. Table 1 explains the technical words.

1 Introduction

The design problem behind this paper starts from a tension any voting system must resolve. A system that hands a policy decision to elected representatives asks citizens to accept whatever bundle of positions their candidate holds, checked only every few years. A system that puts every question directly to a majority vote treats a domain expert's ballot and a ballot cast on a belief the relevant evidence has already ruled out — the example that motivates this design is the belief that 5G infrastructure causes disease, which spread widely during 2020 [Ahmed et al., 2020] — as equally informative, simply because both count as one vote. Between these extremes sits *liquid democracy* (also called delegative democracy or delegable proxy voting): a citizen may cast a direct ballot on any question, or transitively delegate that vote to someone they trust, and may change that choice at any time [Green-Armytage, 2015, Blum and Zuber, 2016, Paulin, 2020].

Liquid democracy has documented failure modes. In the largest early deployment — LiquidFeedback within Germany's Pirate Party — Kling et al. [2015] found a power-law distribution of received delegations (indegree exponent 1.38), with 38 users holding more than 100 delegations each out of 13,836 registered users and 14,964 delegations, and inequality that grew over time; token-based delegation in blockchain governance shows the same pattern at larger scale [Fritsch et al., 2024]. Kahng et al. [2021] prove that *local* delegation mechanisms cannot be guaranteed to beat direct voting even when voters only delegate to better-informed neighbours, and in laboratory experiments Campbell et al. [2022] find delegation rates "unexpectedly high", with liquid democracy underperforming both universal voting and abstention. The underlying reason is classical: delegation replaces many independent votes with a few correlated ones, and the Condorcet Jury Theorem rewards independence [Grofman et al., 1983]. Nothing about the mechanics of delegation guarantees that influence tracks competence, and even when it does, concentrating influence has a cost.

This paper's proposal, Cauce[1], keeps liquid democracy's core commitment (every citizen's vote counts equally and every citizen may act as their own representative at any time) and adds a mechanism aimed at the first failure mode: a delegate's influence over a policy *topic* is chosen alongside a public, falsifiable record of how accurate that delegate's past judgments have been. Concretely, delegation is scoped per topic; a direct ballot on any single bill overrides the standing delegation for that bill; three platform rules (a cap on any one delegate's share of the vote, a limit on delegation chains, and an expiry) bound concentration and volatility; and every delegate vote carries a probabilistic prediction that is later scored with the Brier score [Brier, 1950] and shown to citizens choosing whom to trust.

Our contributions are:

1. **The algorithm** (Section 4). We state Cauce's delegation resolution exactly and prove that direct ballots dominate delegations, that cap enforcement is order-independent and ballot-independent, and that resolution terminates in the presence of cycles (Propositions 1–3).
2. **Why a track record is the right signal** (Section 5). A Brier score decomposes into a calibration term and a competence term; for calibrated, better-than-chance delegates it is a strictly monotone function of competence, and a confidently wrong delegate scores worse than the uninformative forecaster (Proposition 4). Since the probability-maximizing decision rule weights independent voters by the log-odds of their competence (Proposition 5), displaying the score gives citizens, in expectation, exactly the ranking the optimal rule would use — while Cauce, unlike an epistocracy, never assigns anyone's weight by fiat.
3. **An exact trade-off** (Section 6). A large-population limit theorem for Cauce's tally (Propo-

[1] Spanish for "channel" or "riverbed" — the path along which a vote flows when it is delegated.

sition 6) shows that, whenever the delegates are pivotal, delegation *strictly lowers* accuracy when the unaffiliated population is competent and *strictly raises* it when that population is misinformed, and quantifies both effects through the distribution of the delegates' correct weight. Non-asymptotic versions (Section 6.1) give an exponential rate of convergence to the limit, a certified error margin for the semi-analytic predictor, and a bound $\exp(-2n_{\text{eff}}\bar{m}^2)$ in terms of the *effective number of independent voters* $n_{\text{eff}} = 1/H$, which delegation cuts from about 2,400 to about 29 in our populations.

4. **Correlated misinformation** (Sections 8.5–8.6). No configuration of the tally survives a shock that fools every voter (Proposition 10); a shock that fools each voter with some probability lowers every voter's effective competence (Proposition 11), which yields a closed-form critical susceptibility for direct democracy and a break-even shock frequency above which delegation pays (Corollaries 3–4).

5. **Validated simulations** (Sections 7–8). Two implementations of the tally, cross-validated on both the weights and the final decisions, and a semi-analytic predictor derived from Proposition 6 that tracks every simulated curve to within 1.6 points.

6. **What the score and the cap cannot guarantee** (Section 9). How long a record must be to rank delegates (Proposition 12); why author-chosen claims call for relative scores (Proposition 13); why competition for delegations rewards overconfidence (Proposition 14); when an averaged record ranks first the delegate who is worst in a shock (Corollary 5); why competence is not alignment (Proposition 15); and how cheaply a bill can be captured, with a corrected cap (Proposition 16).

7. **Policy domains** (Section 10). What is particular to ten policy sectors, from economic, health, education and defence legislation to technology, climate, criminal justice, migration, infrastructure and pensions; how record length, which depends on how fast a domain's claims resolve, changes the insurance (Table 14); why raw scores fail on rare events (Corollary 6); and a per-topic configuration for each domain.

8. **An automated society** (Section 11). A limit theorem for citizens who vote through AI agents (Proposition 17): a safe zone whose threshold equals the rumour threshold, and a bound in terms of the effective number of independent AI providers; a provider-level cap; and the evidence on AI forecasters, correlated model errors, persuasion and labour-replacing automation.

9. **Objections, counter-evidence and alternatives** (Section 12). A structured review of the strongest evidence against the premise, with what the design does about each item, and a comparison with mechanisms aimed at the same failure.

The results do not support the naive version of the premise — that routing votes toward demonstrably accurate delegates is always better than letting everyone vote. They support a narrower and, we think, more defensible claim: reputation-weighted delegation is cheap insurance against the scenario that motivated the design, a population whose effective competence on a question falls below one half, and the reputation signal is what keeps the premium low.

2 Related Work

Formal models of liquid democracy. Christoff and Grossi [2017] embed delegable proxy voting into binary aggregation, characterize the "proxy majority" rule axiomatically, and study delegation

cycles and the loss of individual consistency when delegating on logically interdependent propositions — a problem topic-scoped delegation makes sharper (Section 12.8). Kahng et al. [2021] model liquid democracy as aggregation of noisy binary judgments and prove that local mechanisms cannot guarantee an improvement over direct voting. Caragiannis and Micha [2019] strengthen the critique: natural delegation mechanisms can be much worse than direct voting, delegating to *less* informed voters can sometimes raise accuracy, and computing optimal delegations is hard to approximate. In the other direction, Halpern et al. [2023] and Berinsky et al. [2025] show that under realistic random delegation models the largest weight any delegate accumulates stays bounded with high probability, which is what the positive results require. Gölz et al. [2021] study delegation profiles that respect delegates' capacities; our cap is a simpler, platform-enforced version of that concern. Zhang and Grossi [2021] define a delegative Banzhaf index for measuring power in delegation graphs, and computing a priori power in liquid democracy is #P-hard [Colley et al., 2023]. When delegation is modelled as a game, pure equilibria exist in natural cases but can lose much of the achievable accuracy [Bloembergen et al., 2019], and iterated best responses need not converge [Escoffier et al., 2019]. With more than two alternatives, delegating separate pairwise comparisons can yield intransitive outcomes [Brill and Talmon, 2018]. Chatterjee et al. [2025] show that local delegation can beat direct voting on several graph families when the mechanism preserves enough variance in the outcome, and Gilbert et al. [2026] that when a public signal changes voters' competences after delegations are fixed, delegating on average competence can do harm (Section 9.4). Bribery and control of delegation graphs are computationally hard in general [D'Angelo et al., 2022, Alouf-Heffetz et al., 2024], and delegation weakens the secret ballot's protection against vote buying [Nejadgholi et al., 2021]. Papasotiropoulos and Schmidt-Kraepelin [2025] give an annotated reading list.

Fractional and multi-proxy delegation. Degrave [2014] proposes splitting a vote fractionally across several proxies. Bersetche [2025] formalizes fractional multi-agent delegation, introduces a continuous penalty on chain length, and proves that pure-strategy equilibria exist once that penalty is positive — naturally read as a fractional analogue of our discrete hop limit, though we prove no equilibrium result for it. Viscous democracy discounts a vote's weight at each delegation step [Boldi et al., 2011], and higher viscosity reduces super-voters' weight and can raise group accuracy [Armstrong et al., 2024]; Cauce's hop limit is its all-or-nothing version. Ranked delegations supply fallbacks when a chain breaks [Brill et al., 2022], and no non-fractional delegation rule satisfies anonymity, copy-robustness and confluence together, while a fractional rule does [Utke and Schmidt-Kraepelin, 2023]. Unlike these general graphs, a Cauce delegate forwards the *entire* weight it receives on a topic to at most one further delegate, choosing legibility over these axiomatic properties: a citizen can always see exactly who carries their vote.

Empirical and experimental evidence. Beyond Kling et al. [2015] and the practitioner account of Behrens [2015], the evidence is thin and mixed. A three-year internal deployment at Google on low-stakes questions saw delegated votes remain a small minority of votes cast [Hardt and Lopes, 2015]. In six experiments with pre-existing groups (168 participants, 62 delegation graphs and over 11,000 votes), actual delegation behaviour satisfied the conditions under which liquid democracy's positive theoretical guarantees hold [Berinsky et al., 2025]; an earlier analysis of the same kind of data found that higher-expertise participants were less likely to delegate [Revel et al., 2022]. In contrast, in a laboratory experiment in which the better-informed participants were identified and an online task in which signal precision was ambiguous, Campbell et al. [2022] found delegation rates "unexpectedly high" and liquid democracy underperforming both universal voting and abstention; a 2026 revision adds a theoretical result in delegation's favour and reaches the same experimental

verdict [Mooers et al., 2026]. In blockchain governance, a small number of addresses hold most voting power, though they rarely overturn the community's preferred outcome [Fritsch et al., 2024]. We found no liquid-democracy experiment that shows delegators a measured track record. The nearest evidence is encouraging. In MakerDAO's governance, token holders gave more delegated votes to delegates whose votes were followed by abnormal token returns, a working paper's measure of voting correctly [Fan et al., 2025]. In Uniswap's, delegation is "at least partly reputation- or merit-based", although delegates affiliated with the controlling venture-capital firm receive more delegations [Bongaerts et al., 2025]. In nine advice-taking experiments, people given feedback preferred accurate advisers, while without feedback they were "biased to select advisors who tend to agree with them" [Jaquiery and Yeung, 2024].

Weighted aggregation and select crowds. For independent voters with known competences, the decision rule that maximizes the probability of a correct decision is a weighted majority with log-odds weights [Nitzan and Paroush, 1982, Shapley and Grofman, 1984]; we restate and prove it as Proposition 5. Empirically, averaging a small "select crowd" of judges chosen on past accuracy is accurate and robust [Mannes et al., 2014], weighting forecasters by their measured contribution beats unweighted aggregation [Budescu and Chen, 2015], and forecasting skill measured by Brier scores is identifiable and persistent in geopolitical tournaments [Mellers et al., 2014, 2015].

Scoring forecasters, and alternatives. Proper scoring rules elicit beliefs from a forecaster whose utility is linear in the score; with non-linear utility, especially combined with other stakes in the event, reports can depart from beliefs [Kadane and Winkler, 1988], and competition among forecasters rewards exaggeration [Lichtendahl and Winkler, 2007, Witkowski et al., 2023]. Raw Brier scores also depend on which questions a forecaster chooses [Bo et al., 2017]. Section 9 develops what these results imply for a delegate paid in delegations. Mechanisms that need no resolved outcomes score reports against other reports [Miller et al., 2005, Prelec, 2004, Karger et al., 2021] or select the answer that is more popular than predicted [Prelec et al., 2017]; Section 12.12 compares them with Cauce.

Correlated errors. The Condorcet Jury Theorem needs independence as much as competence. Ladha [1992] extends it to correlated votes; Dietrich and Spiekermann [2013b,a] show that once voters share common causes (such as the same evidence), votes are at best independent *conditional on the problem*, and the probability of a correct majority then converges not to one but to the probability that the problem is "easy" (plus half the probability that it is on the boundary). Our shock model (Sections 8.5–8.6) is a concrete instance of this structure, and Corollary 2 is its finite-committee analogue.

Political theory and values versus beliefs. Blum and Zuber [2016] survey liquid democracy's normative case; Green-Armytage [2015] treats proxy voting as a hybrid of direct and representative democracy; Paulin [2020] surveys a decade of research. The epistocracy debate [Mill, 1861, Estlund, 2008, Brennan, 2016, Landemore, 2013] is taken up in Section 12.9. Hanson [2013] proposes futarchy: vote on values, bet on beliefs. Cauce shares that separation but intervenes differently — futarchy prices *policies* through markets with monetary stakes; Cauce scores *people* through a non-monetary proper scoring rule — and inherits a version of futarchy's hardest problem, scoring predictions about actions that were never taken (Section 12.7).

3 The Cauce Platform

Before formalizing the resolution algorithm, we describe the platform it implements, in the vocabulary the running system exposes to a citizen.

- **Topics.** Every bill belongs to one topic (economy, health, education, environment, security, transportation, ... in the prototype). Delegation is chosen per topic: a citizen may delegate economic bills to an economist, health bills to an epidemiologist, and vote directly on every education bill.
- **Delegation and splitting.** For each topic a citizen may name up to two delegates and a percentage split between them (e.g. 60%/40%), or delegate to no one and vote directly.
- **Override.** On any single bill, a citizen with a standing delegation may instead cast a direct ballot, which replaces the delegation for that bill only (Proposition 1).
- **Chained delegation.** A delegate may cede an entire topic to a further delegate ("Elena Brenes votes on economic bills herself; Tomás Herrera, who accepted economic delegations, forwards all of them to Elena"). Each citizen decides whether their own weight may follow such a forward at all (`allow_chain`); the platform also bounds how many delegates a chain may pass through (`max_hops`; the default of two allows one forward).
- **Cap.** No delegate may cast more than a fixed share (default 25%) of a bill's *participating* weight — the weight of citizens who neither abstained from the topic nor let their delegation lapse. Excess weight is excluded rather than reassigned (Proposition 2).
- **Expiry and review window.** An unrenewed delegation lapses after a fixed period (twelve months by default). Delegates vote a fixed number of days before a bill closes, so citizens can read the delegate's justification and override.
- **Reputation.** Every delegate vote carries a short justification and exactly one falsifiable prediction with a stated probability ("the delegate's support for this bill... will keep the unemployment rate flat within one point over twelve months, probability 0.6"). Once the outcome is observable, the prediction is scored; the delegate's running Brier score — over every prediction made, since a prediction is mandatory with every vote — is shown beside their name wherever a citizen chooses whom to delegate to (Section 5).

Figure 5 shows the resulting data flow for one bill: weight flows along delegation edges (possibly split, possibly forwarded), lands on a terminal delegate who has cast one of {yes, no, abstain}, and is subject to cap truncation before the totals are computed. This is Algorithm 1.

4 A Formal Model of Delegation Resolution

Fix a topic t and a bill b under it; we suppress both from the notation where unambiguous. Let $N = \{1, \dots, n\}$ be the set of citizens. A subset $M \subseteq N$ are *delegates for* t: citizens who have received at least one delegation on t. Every citizen is in exactly one of three modes for t: DIRECT, DELEGATED, or NONPARTICIPANT.

Definition 1 (Delegation and forwarding). *For a* DELEGATED *citizen i, the* delegation function $\delta(i) = \{(d_1, w_1), \dots, (d_k, w_k)\}$ *assigns citizen i's unit weight across $k \geq 1$ targets $d_j \in M$ with $\sum_j w_j = 1$. The* forwarding function $\varphi : M \to M \cup \{\perp\}$ *maps a delegate to the single further delegate it cedes topic t to, or to $\perp$ if it is* terminal *(casts its own ballot on t). Each citizen i*

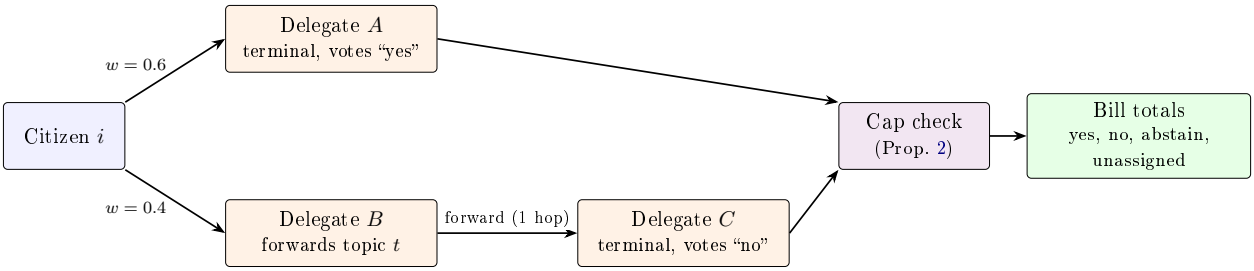


Figure 5: Resolution of one citizen's split delegation on one bill of topic t. Delegate B has ceded the topic to C; because the citizen permits chains and the path (B then C: two delegates, one forward) is within `max_hops` $= 2$, the 0.4 share reaches C's "no" ballot. Both terminal delegates' received weight is checked against the cap before the bill's totals are final.

carries a flag $\text{chain}(i) \in \{0, 1\}$ *governing whether its weight may traverse a forward beyond* i*'s own first-named delegate.*

Definition 2 (Rules)**.** *A rule triple* $\rho = (c, h, e)$ *fixes a cap* $c \in (0, 1] \cup \{\infty\}$*, a hop limit* $h \in \mathbb{N}_{\geq 1}$ *(the largest number of delegates a delegation path may contain;* $h = 1$ *forbids forwarding), and an expiry* $e \in \mathbb{R}_{>0} \cup \{\infty\}$ *(in months). A* Delegated *citizen's delegation on* t *is* lapsed *for bill* b *if its age exceeds* e*.*

Definition 3 (Path resolution)**.** *For a delegate* $d_1 \in M$ *and a citizen* i *with* $\text{chain}(i) = \chi$*, define the resolved terminus* $\pi(d_1, \chi, h)$ *by: if* $\varphi(d_1) = \perp$*, return* d_1*. Otherwise, if* $\chi = 0$*, return* ChainCut*. Otherwise follow* $d_2 = \varphi(d_1), d_3 = \varphi(d_2), \ldots$ *At each step, if the path's length in delegates would exceed* h*, return* ChainCut*; if the next delegate already occurs earlier in the path, return* Cycle*; otherwise, once some* d_m *has* $\varphi(d_m) = \perp$*, return* d_m*.*

Algorithm 1 tallies one bill under a rule triple ρ, given a ballot $\beta(d) \in \{\text{Yes}, \text{No}, \text{Abstain}\}$ for every terminal delegate d and, for citizens who cast a direct ballot on this bill (Direct-mode citizens voting as usual, or Delegated-mode citizens overriding), a ballot $\beta_0(i)$.

Proposition 1 (Override dominance)**.** *If citizen* i *casts an override ballot* $\beta_0(i)$ *on bill* b*, then* i*'s entire contribution to* b*'s tally is* $\beta_0(i)$*, independently of* i*'s standing delegation* $\delta(i)$ *or of any ballot cast by* i*'s delegates.*

Proof. Immediate from line 3 of Algorithm 1: the override branch is checked first and, when taken, is the only branch executed for i; $\delta(i)$ is never read. □

Proposition 2 (Cap well-definedness)**.** *The set of (citizen, weight) contributions moved to* Unassigned *by* Apply-Cap*, and the resulting value of every* totals$[o]$*, do not depend on the order in which delegates* $d \in M$ *are processed in the capping loop (lines 21–24), and depend on the ballots only through which bucket each delegate's excess is removed from.*

Proof. Each terminal delegate d casts one ballot $\beta(d)$ per bill, so every unit of weight in received$[d]$ maps to the single bucket $\beta(d)$. Truncating d's excess subtracts the fixed quantity $\max(0, \text{received}[d] - c \cdot \text{participation})$ from exactly one bucket, totals$[\beta(d)]$, and adds it to totals[Unassigned]. Since received$[d]$ and participation are computed before the loop and are not mutated by it, no delegate's truncation depends on any other's, and the loop's net effect is a sum of independent corrections, whatever the processing order; the amounts depend only on δ, φ, ρ and on which citizens override on b. □

Algorithm 1 RESOLVE-TOPIC-BILL($N, \delta, \varphi, \rho, \beta, \beta_0$)

totals[YES], totals[NO], totals[ABSTAIN] $\leftarrow 0$; totals[UNASSIGNED] $\leftarrow 0$; received[d] $\leftarrow 0$ $\forall d \in M$
for $i \in N$ **do**
  if i is DELEGATED and casts an override ballot $\beta_0(i)$ **then** ▷ Prop. 1
    totals[$\beta_0(i)$] += 1; **continue**
  else if i is NONPARTICIPANT, or DELEGATED and lapsed **then** ▷ tallied separately
    continue
  else if i is DIRECT **then**
    totals[$\beta_0(i)$] += 1; **continue** ▷ a standing direct ballot
  else ▷ i is DELEGATED and active
    for $(d, w) \in \delta(i)$ **do**
      $r \leftarrow \pi(d, \text{chain}(i), h)$
      if $r \in \{\text{CHAINCUT}, \text{CYCLE}\}$ **then**
        totals[UNASSIGNED] += w
      else
        totals[$\beta(r)$] += w; received[r] += w
      end if
    end for
  end if
end for
participation $\leftarrow \sum_{o \in \{\text{YES}, \text{NO}, \text{ABSTAIN}\}} \text{totals}[o] + \text{totals}[\text{UNASSIGNED}]$
for $d \in M$ with received[d] $> c \cdot$ participation **do** ▷ APPLY-CAP, Prop. 2
  excess $\leftarrow$ received[d] $- c \cdot$ participation
  totals[$\beta(d)$] −= excess; totals[UNASSIGNED] += excess
end for
return totals

A consequence used throughout Section 8 is that *who* is capped, and by how much, can be computed once per rule configuration and set of overriding citizens and reused across bills; since our accuracy experiments have no overrides, once per configuration suffices, which is what makes the accelerated implementation of Section 7 correct and fast.

Proposition 3 (Termination)**.** *For every citizen i, delegate d, and hop limit h, evaluating $\pi(d, \text{chain}(i), h)$ (Definition 3) terminates after at most $\min(h, |M|)$ steps.*

Proof. Each step either reaches a terminal delegate, exceeds the hop budget h (CHAINCUT), or revisits a delegate already on the path (CYCLE). The path length is bounded by h by construction and by $|M|$ because a path of length $|M|+1$ over $|M|$ delegates repeats a vertex (pigeonhole). Hence resolution halts within $\min(h, |M|)$ steps regardless of φ, including when φ contains cycles. □

In plain words. However the channels are dug, every vote ends somewhere visible. A vote cast directly always beats a standing delegation; the cap gives the same result whatever order it is applied in; and a vote caught in a loop of delegations (Ana trusts Ben, Ben trusts Ana) is set aside openly rather than lost in silence. Like water in a well-built irrigation system, every drop is either delivered or accounted for.

Remark 1 (Relation to proxy majority)**.** Once Algorithm 1 has resolved every citizen's weight to a terminal ballot (or to UNASSIGNED), comparing the weighted totals is exactly the *proxy majority*

aggregator of Christoff and Grossi [2017], which they characterize as the aggregator that is anonymous, independent, monotonic, responsive and neutral and that minimizes undecisiveness (their Fact 1). Cauce departs from that baseline only upstream: in *which* weighted ballots reach the final step, governed by δ, φ, ρ and, through Section 5, by the information citizens use to choose δ.

5 The Reputation Layer

5.1 What a Brier score measures

Every delegate vote carries one falsifiable prediction $p \in [0,1]$: the delegate's stated probability that a named, checkable claim about the bill's effect will turn out true. Once the claim resolves, an outcome $o \in \{0,1\}$ is recorded and the prediction contributes $(p-o)^2$ to the delegate's running *Brier score* [Brier, 1950],

$$B_d \;=\; \frac{1}{|H_d|} \sum_{(p,o)\in H_d} (p-o)^2,$$

over the delegate's history H_d of scored predictions. We use the binary form, which lies in $[0,1]$ with lower values better; forecasting tournaments often report the two-outcome form, which is exactly twice ours [Mellers et al., 2014], and Gneiting and Raftery [2007] orient scores as rewards (the negative of ours). $B_d = 0$ is a perfect record; $B_d = \frac{1}{4}$ is what the uninformative forecast $p = \frac{1}{2}$ earns on every question.

Proposition 4 (Brier decomposition)**.** *Let a forecaster state probability p for a binary event whose true probability is q. Then*

$$\mathbb{E}\big[(p-o)^2\big] \;=\; (p-q)^2 + q(1-q).$$

Consequently:

(i) (Strict propriety.) *The expected score is uniquely minimized by reporting $p = q$.*

(ii) (Calibrated competence.) *If the scored event is "the delegate's own ballot is correct" and the delegate's competence is $\theta \geq \frac{1}{2}$, a calibrated delegate who always reports $p = \theta$ has expected score $\theta(1-\theta)$, which is strictly decreasing on $[\frac{1}{2}, 1]$ and inverts as $\theta = \frac{1}{2}\big(1 + \sqrt{1-4B}\big)$.*

(iii) (Confidently wrong.) *A delegate with competence $\theta < \frac{1}{2}$ who states any $p > \frac{1}{2}$ has expected score strictly greater than $\frac{1}{4}$, the score of the uninformative forecast.*

Proof. $\mathbb{E}[(p-o)^2] = q(1-p)^2 + (1-q)p^2 = p^2 - 2pq + q = (p-q)^2 + q - q^2$. Part (i) is immediate since $q(1-q)$ does not depend on p. For (ii), set $q = p = \theta$; $\frac{d}{d\theta}\theta(1-\theta) = 1 - 2\theta < 0$ for $\theta > \frac{1}{2}$, and solving $\theta^2 - \theta + B = 0$ on $[\frac{1}{2}, 1]$ gives the inverse. For (iii), $(p-\theta)^2 + \theta(1-\theta) > \frac{1}{4} \iff (p-\theta)^2 > (\frac{1}{2} - \theta)^2 \iff p - \theta > \frac{1}{2} - \theta$ (both sides positive since $p > \frac{1}{2} > \theta$) $\iff p > \frac{1}{2}$. □

In plain words. Think of a weather forecaster. If she says "70% chance of rain" on many days, it should rain on about seven in ten of them. The Brier score charges her the squared distance between what she said and what happened. She cannot improve her expected score by exaggerating or by hedging: the best strategy is to say exactly how sure she is. Someone who always says "50%" scores 0.25; someone who is usually wrong but sounds certain scores worse than that — worse than admitting "I don't know" (Figure 3).

Part (i) is the standard reason proper scoring rules are used [Brier, 1950, Gneiting and Raftery, 2007]: a forecaster gains nothing by misstating a probability she has decided to report. Parts (ii)–(iii) are why a track record addresses the paper's motivating failure mode. The stylization in (ii) —

treating each scored prediction as a statement about the delegate's own ballot — is an idealization of Cauce's predictions, which concern a bill's effects; it is the reading under which "demonstrated accuracy" and "competence" coincide, and it is what our simulations operationalize (Section 7). Two limits of (ii) matter. Below chance the score is symmetric — a calibrated delegate with $\theta = 0.4$ scores like one with $\theta = 0.6$ — which is why the ranking is claimed only above chance; our simulations assume that below-chance influencers behave as in part (iii), confidently wrong, rather than as calibrated forecasters. And a delegate whose stated confidence varies across questions is scored on calibration and sharpness jointly, so the one-number ranking is an idealization for a delegate with constant confidence. Part (iii) matters for the 5G example: a popular voice that is confidently wrong does not merely fail to stand out, it scores measurably *worse* than someone who says "I don't know".

5.2 Why the score should steer weight, and why the platform does not

Proposition 5 (Optimal weighted majority; Nitzan and Paroush, 1982, Shapley and Grofman, 1984)**.** *Let voters $v = 1, \dots, k$ vote independently on a binary question with known competences $\theta_v \in (0, 1)$, and let both answers be a priori equally likely. Among all rules mapping ballot profiles to a decision, the probability of a correct decision is maximized by the weighted majority with weights*

$$w_v \;=\; \log \frac{\theta_v}{1 - \theta_v},$$

i.e. decide YES *iff $\sum_v w_v s_v > 0$, where $s_v = +1$ if v votes* YES *and -1 otherwise (ties broken arbitrarily).*

Proof. Let y be the true answer and $\mathbf{b}$ the ballot profile. For any rule f, $\Pr[f(\mathbf{b}) = y] = \sum_{\mathbf{b}} \Pr[\mathbf{b}, y = f(\mathbf{b})] \le \sum_{\mathbf{b}} \max_{y'} \Pr[\mathbf{b}, y']$, with equality exactly when f picks the more likely answer given $\mathbf{b}$. With equal priors this compares likelihoods, and by independence

$$\log \frac{\Pr[\mathbf{b} \mid \text{YES}]}{\Pr[\mathbf{b} \mid \text{NO}]} = \sum_{v:\, s_v = +1} \log \frac{\theta_v}{1 - \theta_v} + \sum_{v:\, s_v = -1} \log \frac{1 - \theta_v}{\theta_v} = \sum_v w_v s_v. \qquad \square$$

In plain words. If you knew how often each adviser is right, the best way to combine their votes would be to listen more to the reliable ones — and to do the opposite of those who are usually wrong. This holds if their mistakes are independent. No democracy does that, and Cauce does not either: it hands no one weight by fiat, and a delegate carries only the votes citizens choose to send her. It only puts each delegate's record on display, the way a guidebook shows ratings, and leaves the choice to each citizen.

Remark 2 (What the platform can and cannot do with Proposition 5)**.** Three observations organize much of Section 12. First, a voter with $\theta_v < \frac{1}{2}$ gets a *negative* optimal weight: the optimal rule inverts her ballot. No democratic procedure does this. Second, restricting weights to be non-negative sets those voters' weights to zero — disenfranchisement, which is the epistocratic remedy [Mill, 1861, Brennan, 2016] Cauce rejects. Third, Cauce assigns no weights at all: every citizen's unit of weight goes wherever she sends it, and the platform's only lever is information. By Proposition 4(ii), for calibrated, better-than-chance delegates the Brier score is a monotone transform of competence and hence of the optimal weight, $w = \log \frac{1+\sqrt{1-4B}}{1-\sqrt{1-4B}}$; displaying it hands citizens, in expectation, the ranking Proposition 5 would use, and leaves the decision to them; Proposition 12 says how long a record must be for a realized score to do so. Section 8.2 measures how closely each delegate-choice rule follows that ranking (Figure 8).

What the scoring rule cannot do — discipline the ballot itself, score predictions about bills that failed, or keep pace with fast-moving misinformation — is taken up in Section 12.7; how long a record must be, how it can be gamed, and what it says about a delegate's ends, in Section 9.

6 An Epistemic Model of the Tally

To ask whether delegation helps, we need ballots that can be right or wrong. Every bill has a binary ground-truth-correct side, unknown to voters. Delegate d's ballot matches it with probability θ_d (its *competence*); the *unaffiliated pool* — DIRECT-mode citizens plus whatever share of their weight DELEGATED citizens route to small, unnamed delegates — consists of many citizens, *each* voting correctly with probability θ_{pool}, independently (we model a citizen's share routed to small delegates as if it were her own independent vote). Let W be the weight cast for either side: participating weight minus weight excluded by the cap, by chain cuts or cycles, or cast as abstentions (our accuracy experiments have no chains or abstentions, so only the cap matters). It splits into the pool's weight $q = \sum_{j=1}^{m} \omega_j$, spread over m individual contributions, and the delegates' capped weights w_d with $\sum_d w_d = W - q$. Write $X_d \in \{0, 1\}$ for "delegate d votes correctly" and

$$D \;=\; \sum_d w_d X_d$$

for the delegates' correct weight. The aggregate outcome is correct iff the correct weight exceeds $W/2$.

Proposition 6 (Large-pool limit: delegation as a trade-off)**.** *Let the pool grow ($m \to \infty$) with q fixed and $\max_j \omega_j \to 0$, the delegates and their weights fixed, and ballots independent. Let $T = W/2 - q\theta_{\text{pool}}$ and suppose $\Pr[D = T] = 0$. Then*

$$\Pr[\textit{aggregate outcome correct}] \;\longrightarrow\; \Pr[D > T].$$

Consequently:

- *(i)* (Direct democracy.) *With no delegates ($q = W$), the limit is 1 if $\theta_{\text{pool}} > \frac{1}{2}$ and 0 if $\theta_{\text{pool}} < \frac{1}{2}$ — the Condorcet Jury Theorem.*
- *(ii)* (Cost.) *If $\theta_{\text{pool}} > \frac{1}{2}$ but $q\theta_{\text{pool}} < W/2$, the limit is at most $1 - \prod_d(1 - \theta_d) < 1$: delegation strictly lowers accuracy relative to direct democracy.*
- *(iii)* (Benefit.) *If $\theta_{\text{pool}} < \frac{1}{2}$ and $q(1 - \theta_{\text{pool}}) < W/2$, the limit is at least $\prod_d \theta_d > 0$: delegation strictly raises accuracy relative to direct democracy.*
- *(iv)* (No single knife-edge.) *As a function of θ_{pool}, the limit is a non-decreasing step function whose jumps sit at $\theta_{\text{pool}} = (W/2 - D(x))/q$ for the $2^{|M|}$ delegate outcomes x, each of height $\Pr[x]$.*

Proof. Let $P_m = \sum_j \omega_j Y_j$ be the pool's correct weight, with $Y_j \sim \text{Bernoulli}(\theta_{\text{pool}})$ independent. Then $\mathbb{E}P_m = q\theta_{\text{pool}}$ and $\operatorname{Var} P_m = \theta_{\text{pool}}(1 - \theta_{\text{pool}}) \sum_j \omega_j^2 \le \theta_{\text{pool}}(1 - \theta_{\text{pool}})\, q \max_j \omega_j \to 0$, so $\Delta_m = P_m - q\theta_{\text{pool}} \to 0$ in probability (Chebyshev). The outcome is correct iff $D + P_m > W/2$, i.e. iff $D > T - \Delta_m$. D takes finitely many values and T is not one of them, so $\eta = \min_x |D(x) - T| > 0$; on the event $|\Delta_m| < \eta$, $D > T - \Delta_m \iff D > T$. Hence $|\Pr[\text{correct}] - \Pr[D > T]| \le 2\Pr[|\Delta_m| \ge \eta] \to 0$. (i) With no delegates $D \equiv 0$ and $T = q(\frac{1}{2} - \theta_{\text{pool}})$, so the event $\{0 > T\}$ holds iff $\theta_{\text{pool}} > \frac{1}{2}$. (ii) $q\theta_{\text{pool}} < W/2$ means $T > 0$, so the outcome "every delegate wrong" ($D = 0$), of probability $\prod_d(1 - \theta_d)$, lies in $\{D \le T\}$. (iii) $q(1 - \theta_{\text{pool}}) < W/2$ is equivalent to $W - q > T$, so the outcome "every delegate right" ($D = W - q$), of probability $\prod_d \theta_d$, lies in $\{D > T\}$. (iv)

$\Pr[D > T] = \sum_x \Pr[x]\,\mathbf{1}[D(x) > W/2 - q\theta_{\text{pool}}]$, and each indicator switches from 0 to 1 once, at the stated value of θ_{pool}. □

In plain words. Picture a tug of war decided by many independent hands. If each hand pulls the right way a little more often than not, adding hands makes victory nearly certain; if each pulls the wrong way a little more often, more hands make defeat nearly certain (Figure 2). Handing a large block of votes to a few delegates replaces many independent hands with a few strong ones. When delegates hold enough of the vote, that costs accuracy if the crowd leans right, and it is what makes a rescue possible if the crowd leans wrong.

Remark 3 (Reading Proposition 6)**.** The pool obeys its own law of large numbers, so it is nearly deterministic: it contributes almost exactly $q\theta_{\text{pool}}$ of correct weight. The delegates are a *finite committee* that never gets a law of large numbers of its own, so the outcome is decided by whether their correct weight clears the gap T the pool leaves. When the pool is competent this costs accuracy (part ii) — the formal content of the objection that delegation "reduces the variety of independent information sources" [Campbell et al., 2022, Kahng et al., 2021]. When the pool is misinformed it is the only way to be right at all (part iii). The delegate-choice rule matters through the distribution of D: moving weight toward high-θ_d delegates raises the mean of D and so tends to shrink the cost and enlarge the benefit — a tendency our simulations confirm, not a theorem, since it also changes D's spread. In our synthetic populations $q \approx 0.44$ of cast weight (Section 7), so conditions (ii) and (iii) hold for $\theta_{\text{pool}} = 0.65$ and $\theta_{\text{pool}} = 0.35$ respectively. Two cautions. Parts (ii) and (iii) compare Cauce with a direct democracy in which the citizens who delegate would instead vote themselves, independently and with the pool's competence θ_{pool}; our simulations build the direct-democracy baseline the same way. And the bounds in (ii) and (iii) are loose — for our populations the first is within 10^{-6} of one and the second is about 0.003 — so the *size* of the trade-off comes from evaluating $\Pr[D > T]$ itself, which is what the predictions below do.

For finite populations we use the natural refinement of Proposition 6: enumerate all $2^{|M|}$ delegate outcomes exactly and replace the pool by its normal approximation,

$$\Pr[\text{correct}] \;\approx\; \sum_x \Pr[x]\;\Phi\left(\frac{D(x) + q\theta_{\text{pool}} - W/2}{\sqrt{\theta_{\text{pool}}(1-\theta_{\text{pool}})\sum_j \omega_j^2}}\right),$$

evaluated separately for every simulated population. These semi-analytic predictions are drawn as dashed lines in every accuracy figure and reported beside the simulated values. They use the same simulated populations but none of the random ballots, so they check the simulation's Monte Carlo and the normal approximation; they do not test the model itself.

6.1 Finite-sample guarantees: how good is the limit, and how many voters are left?

Proposition 6 is asymptotic in the pool, and the semi-analytic predictor above is a normal approximation. Neither says how far a population of a few thousand citizens is from the limit, and neither says what delegation does to the *number of independent voters*, the quantity on which every Condorcet-type guarantee rests. This subsection closes both gaps with non-asymptotic statements. Throughout, $\Omega_2 = \sum_j \omega_j^2$ is the pool's sum of squared weights, $D(x)$ and $T = W/2 - q\theta_{\text{pool}}$ are as in Proposition 6, and $g(x) = |D(x) - T|$ is the distance of delegate outcome x from the pool's tipping point.

Proposition 7 (Exponential convergence to the limit)**.** *For every finite pool,*

$$\Big|\Pr[\textit{correct}] - \Pr[D > T]\Big| \;\le\; \sum_x \Pr[x]\, e^{-2g(x)^2/\Omega_2} \;\le\; e^{-2\eta^2/\Omega_2}, \qquad \eta = \min_x g(x).$$

Since $\Omega_2 \le q \max_j \omega_j$, the error vanishes exponentially fast in $1/\max_j \omega_j$ whenever $\eta > 0$.

Proof. Fix a delegate outcome x and write $\Delta = \sum_j \omega_j(Y_j - \theta_{\text{pool}})$, so that the decision is correct iff $\Delta > T - D(x)$. If $D(x) > T$ the decision is wrong iff $\Delta \le -g(x)$; if $D(x) < T$ it is right iff $\Delta > g(x)$. In both cases the deviation has probability at most e^{-2g^2/Ω_2} by Hoeffding's inequality, because $\omega_j Y_j \in [0, \omega_j]$ and $\sum_j \omega_j^2 = \Omega_2$. Hence $|\Pr[\text{correct} \mid x] - \mathbf{1}[D(x) > T]| \le e^{-2g(x)^2/\Omega_2}$ (trivially true if $g(x) = 0$); averaging over x gives the first bound, and $g \ge \eta$ the second. □

Proposition 8 (A certified predictor)**.** *Let N be the normal-pool predictor of this section,*

$$N = \sum_x \Pr[x]\, \Phi\left(\frac{D(x) + q\theta_{\text{pool}} - W/2}{\sqrt{\theta_{\text{pool}}(1-\theta_{\text{pool}})\Omega_2}}\right).$$

Then

$$\big|\Pr[\textit{correct}] - N\big| \;\le\; \varepsilon_{\text{BE}} \;=\; C_0\, \frac{\sum_j \omega_j^3 \left(\theta_{\text{pool}}^2 + (1-\theta_{\text{pool}})^2\right)}{\sqrt{\theta_{\text{pool}}(1-\theta_{\text{pool}})}\; \Omega_2^{3/2}}, \qquad C_0 = 0.56.$$

In particular $\varepsilon_{\text{BE}} \le C_0\, \dfrac{\theta_{\text{pool}}^2 + (1-\theta_{\text{pool}})^2}{\sqrt{\theta_{\text{pool}}(1-\theta_{\text{pool}})}}\, \dfrac{\max_j \omega_j}{\sqrt{\Omega_2}} \to 0$ as the pool spreads out.

Proof. Conditional on x, $\Pr[\text{correct} \mid x] = \Pr[S > t_x]$ with $S = \sum_j \xi_j$, $\xi_j = \omega_j(Y_j - \theta_{\text{pool}})$, and Φ is the normal approximation of $S/\sqrt{\theta_{\text{pool}}(1-\theta_{\text{pool}})\Omega_2}$. The Berry–Esseen theorem for independent, non-identically distributed summands, with the constant 0.56 of Shevtsova [2010], bounds the Kolmogorov distance, *uniformly in the threshold*, by $C_0 \sum_j \mathbb{E}|\xi_j|^3 / (\sum_j \operatorname{Var}\xi_j)^{3/2}$. Here $\operatorname{Var}\xi_j = \omega_j^2 \theta_{\text{pool}}(1-\theta_{\text{pool}})$ and $\mathbb{E}|\xi_j|^3 = \omega_j^3\, \theta_{\text{pool}}(1-\theta_{\text{pool}})\big(\theta_{\text{pool}}^2 + (1-\theta_{\text{pool}})^2\big)$, which gives the stated ratio. Uniformity in the threshold means the same bound holds for every x, hence for the average. The second form follows from $\sum_j \omega_j^3 \le \max_j \omega_j\, \Omega_2$. □

Proposition 9 (Effective number of independent voters)**.** *Let counted voters $i = 1, \dots, k$ (named delegates and individual pool citizens) vote independently and correctly with probabilities θ_i, carry weights $w_i > 0$, and let $W = \sum_i w_i$. Put*

$$H = \sum_i \Big(\frac{w_i}{W}\Big)^2, \qquad n_{\text{eff}} = \frac{1}{H}, \qquad \bar{m} = \sum_i \frac{w_i}{W}\Big(\theta_i - \frac{1}{2}\Big) > 0.$$

Then

$$\Pr[\textit{incorrect}] \;\le\; \exp\big(-2\, n_{\text{eff}}\, \bar{m}^2\big) \;\le\; \exp\Big(-\frac{2\bar{m}^2}{c_{\text{eff}}}\Big), \qquad c_{\text{eff}} = \max_i \frac{w_i}{W}.$$

Proof. The decision is wrong iff $\sum_i w_i X_i \le W/2$, i.e. iff $\sum_i w_i(X_i - \theta_i) \le -W\bar{m}$. Each $w_i X_i$ ranges over an interval of length w_i, so Hoeffding's inequality gives $\Pr \le \exp(-2W^2\bar{m}^2/\sum_i w_i^2) = \exp(-2\bar{m}^2/H)$. Finally $H = \sum_i (w_i/W)^2 \le \max_i(w_i/W) \sum_i (w_i/W) = c_{\text{eff}}$. □

In plain words. Imagine a jury in which some jurors speak with a megaphone. A megaphone does not make a juror cleverer, only louder, so the jury behaves as if it had fewer members. Proposition 9 counts how many: its *effective* size, which can be tiny even among thousands of participants. The chance that

the jury is wrong shrinks exponentially in that effective size and in how much better than a coin toss its members are on average. The other two results say that once the crowd is a few thousand strong, the tidy limit formulas of this paper are accurate to a fraction of a percentage point, and give a guaranteed error margin.

Corollary 1 (Sizing the guardrail)**.** *Under the conditions of Proposition 9, the error is at most ϵ whenever $H \le 2\bar{m}^2/\ln(1/\epsilon)$, and in particular whenever the largest counted share satisfies*

$$c_{\text{eff}} \;\le\; \frac{2\bar{m}^2}{\ln(1/\epsilon)}.$$

Delegation lowers n_{eff} from $\approx k$ (equal weights) to $1/H$, so it costs reliability exactly through H: for the same $\bar{m}$ the guaranteed error exponent shrinks by the factor kH.

Proof. Set the bound of Proposition 9 equal to ϵ and solve for H; the second statement uses $H \le c_{\text{eff}}$. For equal weights $H = 1/k$ and the exponent is $2\bar{m}^2 k$; in general it is $2\bar{m}^2/H = 2\bar{m}^2 k/(kH)$, and $kH \ge 1$ by the Cauchy–Schwarz inequality. □

What the numbers say. Table 2 checks all three results on the 20 simulated populations behind each row. Every bound held in every one of the 180 populations. The gap between simulation and the limit of Proposition 6 is at most 0.4 points on average, and the Hoeffding bound on it is between two and twenty times larger, depending on the row. The normal predictor is accurate to under 0.4 points while its certified Berry–Esseen bound is about 2.3 points: the bound is conservative but rigorous, so the dashed curves in the accuracy figures are certified to within that margin. The most informative column is n_{eff}. Among about 1,280 counted voters, reputation-weighted delegation leaves $n_{\text{eff}} \approx 29$ effective independent voters, popularity-driven delegation about 15, and direct democracy about 2,400. For the same edge over chance, the guaranteed exponent $2\bar{m}^2 n_{\text{eff}}$ is therefore about eighty times smaller under reputation-weighted delegation than under direct democracy (2,405/28.7), which is the finite-sample form of the price of delegation in Proposition 6(ii). Corollary 1 turns this into a design rule. Take the accuracy-weighted row with $\theta_{\text{pool}} = 0.65$, where $\bar{m} = 0.208$. To certify $\epsilon = 0.01$ the largest counted share would have to be below $2\bar{m}^2/\ln 100 \approx 0.019$, and the realised $c_{\text{eff}} = 0.106$ is far above that: the certificate through c_{eff} is only $\exp(-2\bar{m}^2/c_{\text{eff}}) = 0.44$. The certificate through H is much stronger, 0.088, because $H = 0.035$ is three times smaller than c_{eff}: it is the sum of *all* squared shares, not the largest one, that governs reliability. Hoeffding is conservative, and the realised error is 0.9%. A cap bounds c_{eff}, but a cap alone cannot certify accuracy.

7 Experimental Setup

Correctness. We implement Algorithm 1 literally, one citizen at a time (the *reference* implementation), and separately a vectorized version that uses Proposition 2 to precompute capped weights once per configuration (the *fast* implementation), used for the large accuracy sweeps. Nine hand-constructed unit tests check one rule at a time: a single delegation, a fractional split, override dominance, an honoured one-hop forward, a forward blocked by chain(i) = 0, a forward cut for exceeding h, a detected two-cycle, a lapsed delegation, and an exact cap truncation. All pass. We then cross-validate the fast implementation against the reference one, feeding both *identical* random ballots (each delegate's ballot and each individual citizen's ballot): the weighted totals agree to within 10^{-6} on 300 bills, and the final *decisions* agree on 1,000 of 1,000 bills both without a cap

Table 2: Non-asymptotic guarantees checked on 20 simulated populations per row (n = 3,000, M = 14, cap 0.25, 20,000 ballots each; all entries are means over the 20 populations, and the simulated side carries Monte-Carlo noise of about 0.003). $n_{\text{eff}} = 1/H$ is the effective number of independent voters and $\bar{m}$ the weighted mean edge over chance. "Err" is the simulated error probability and "H-bound" is $\exp(-2\bar{m}^2/H)$ from Proposition 9, defined only when $\bar{m} > 0$. "Gap to limit" is |simulated − Pr[$D > T$]| and "bound" the Hoeffding bound on it from Proposition 7; "gap to normal" is |simulated − N| and "BE" its Berry–Esseen bound from Proposition 8. In every one of the 180 populations each bound held (up to the Monte-Carlo noise).

Rule	θ_{pool}	n_{eff}	$\bar{m}$	Err	H-bound	Gap to limit	Bound	Gap to normal	BE
Accuracy-weighted	0.65	28.7	0.208	0.009	0.088	0.0005	0.0047	0.0006	0.023
Accuracy-weighted	0.35	28.7	0.075	0.168	0.721	0.0021	0.0481	0.0021	0.023
Accuracy-weighted	0.53	28.7	0.155	0.036	0.257	0.0010	0.0145	0.0009	0.020
Popularity-driven	0.65	14.8	0.073	0.360	0.851	0.0040	0.0077	0.0036	0.023
Popularity-driven	0.35	14.8	-0.059	0.715	–	0.0031	0.0828	0.0024	0.023
Popularity-driven	0.53	14.8	0.021	0.387	0.985	0.0036	0.0244	0.0039	0.020
Direct democracy	0.65	2,405	0.150	0.000	0.000	0.0000	0.0000	0.0000	0.013
Direct democracy	0.35	2,405	-0.150	1.000	–	0.0000	0.0000	0.0000	0.013
Direct democracy	0.53	2,405	0.030	0.002	0.013	0.0016	0.0132	0.0002	0.011

and with a binding 10% cap. Finally, every accuracy figure is checked against the semi-analytic predictions of Section 6.

These checks caught two bugs, which we report because both changed results. An early fast implementation drew a single random outcome for the whole unaffiliated pool on each bill instead of one per citizen, which destroys the Condorcet cancellation the pool's accuracy depends on. A later version compared the correct weight against a majority threshold from which lapsed weight had been subtracted a second time; because direct democracy has no delegations to lapse, this silently lowered the bar *only* for the delegation rules. On the cross-validation instance the old threshold disagreed with the exact resolver on 43% of decisions. A previous draft of this paper, built on that version, reported that reputation-weighted delegation matched or beat direct democracy at every level of population competence; with the corrected threshold, that is false (Section 8.3). The decision-level cross-validation above was added so that this class of error cannot recur silently.

Synthetic population. Each experiment instantiates n citizens and M named delegates for one topic. Delegates draw a competence $\theta_d \sim \text{Unif}(0.58, 0.85)$ and a popularity $\gamma_d \sim \text{Unif}(0.2, 0.6)$, except two *influencers* with $\theta_d \sim \text{Unif}(0.32, 0.46)$ — below chance — and $\gamma_d \sim \text{Unif}(3.5, 5.5)$: popular in spite of, not because of, being right. The popularity gap (roughly 6–28×, about 11× for typical draws) is chosen to sit in the qualitative regime of the super-voter gap Kling et al. [2015] report, not to reproduce their numbers. Each citizen is independently Nonparticipant (20%), Direct (20%) or Delegated (60%: 48% to one named delegate, 12% split between two). About 55% of delegated citizens route between 15% and 45% of their weight to the unnamed pool, so the pool carries about 16.5% of delegated weight and, together with direct voters, $q \approx 0.44$ of cast weight. Named targets are drawn with probability proportional to γ_d (rule Popularity), $\hat{\theta}_d^6$ (rule Accuracy) or $\gamma_d \hat{\theta}_d^3$ (rule Mixed), where $\hat{\theta}_d = \min(0.99, \max(0.5, \theta_d + \varepsilon_d))$, $\varepsilon_d \sim \mathcal{N}(0, 0.04^2)$, is a noisy public estimate of competence — the quantity Proposition 4(ii) says a calibrated track record reveals, clipped at 0.5 so that below-chance delegates simply look uninformative. The noise level is roughly what a record of about 125 resolved predictions per delegate would give ($\sqrt{\theta(1-\theta)/n} \approx 0.04$); shorter records mean noisier estimates and weaker reputation effects than we report. Mixed models a citizen drawn by visibility but partly sensitive to the record. Twelve percent of citizens

Table 3: Vote-share cost of each guardrail, all else at the defaults (cap 0.25, h=2, expiry 12 months), as shares of the whole electorate. "Unassigned" sums the cap, chain-cut and cycle categories.

Rule swept	Value	Yes	No	Unassigned	Lapsed
Cap	none	0.284	0.302	0.017	0.206
	0.30	0.290	0.296	0.017	0.206
	0.25 (default)	0.286	0.299	0.017	0.206
	0.20	0.289	0.297	0.017	0.206
	0.15	0.295	0.290	0.017	0.206
	0.10	0.296	0.286	0.021	0.206
	0.05	0.267	0.270	0.067	0.206
Max. hops h	1 (no forwarding)	0.281	0.297	0.025	0.206
	2 (default)	0.287	0.299	0.017	0.206
	3	0.299	0.301	0.003	0.206
Expiry (months)	none	0.378	0.404	0.027	0.000
	24	0.389	0.392	0.027	0.000
	12 (default)	0.286	0.300	0.017	0.206
	6	0.189	0.198	0.009	0.412

set chain(i) = 0. Delegation ages are uniform on $[0, 18)$ months, so a twelve-month expiry lapses about a third of delegations — deliberately aggressive, to make its cost visible.

Experiments. (1) Guardrail cost: n = 4,000, M = 16, 150 bills, exact resolver, with two of the sixteen delegates ceding their topic onward (forming one chain through three delegates). (2) Concentration: n = 6,000, M = 16. (3)–(6) Accuracy sweeps: n = 3,000, M = 14, cap 0.25, 250 bills per population and 10 independent populations per cell; we report means with 95% confidence intervals from the across-population standard error. Section 9 adds two computations: the record length needed to rank delegates, checked by Monte Carlo, and the cost of capturing a bill, measured on the populations of (3)–(6) at $\theta_{\text{pool}} = 0.65$. All code, seeds and raw outputs are released with the paper, and every table of results is generated directly from those outputs.

8 Results

8.1 The vote-share cost of each guardrail

We first ask what each platform rule costs in votes excluded from a bill's tally, holding the population and delegation graph fixed. Table 3 and Figure 6 report the three sweeps. Nonparticipants (19.2%, unaffected by any rule) are omitted.

The cap barely binds at realistic thresholds: no delegate accumulates enough weight for the default 25% cap to exclude any votes, and even a 10% cap excludes only 0.3% of the electorate (0.6% of participating weight) through the cap itself — the 2.1% "Unassigned" in that row is mostly the constant 1.7% chain-cut term; only at 5% does it exclude 4.9% of the electorate (8.1% of participating weight). Expiry is by far the most expensive rule: tightening it from twelve to six months doubles the lapsed share of the electorate (from 20.6% to 41.2%), directly at the expense of cast votes. Its calibration is a judgment about how often citizens should revisit a delegation, not something the model can derive.

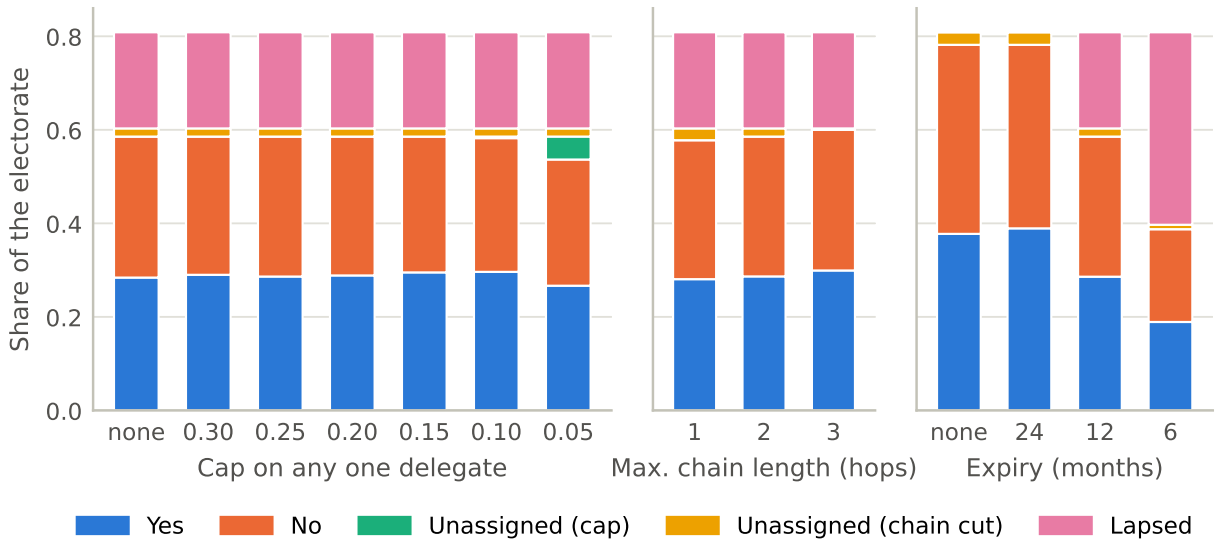


Figure 6: Where a citizen's weight ends up under each rule setting, averaged over 150 bills (nonparticipants omitted). "Unassigned" categories are exclusions the platform, not the citizen, is responsible for. Numbers in Table 3.

Table 4: Concentration of delegate weight. Shares are of participating weight; "influencer share" is the combined share of the two low-competence, high-popularity delegates.

Delegate-choice rule	Cap	Gini	Top-1 share	Top-3 share	Influencer share
Popularity	none	0.550	0.173	0.366	0.343
	0.25 (default)	0.550	0.173	0.366	0.343
Mixed	none	0.382	0.109	0.257	0.218
	0.25 (default)	0.382	0.109	0.257	0.218
Accuracy	none	0.405	0.112	0.248	0.011
	0.25 (default)	0.405	0.112	0.248	0.011

8.2 Concentration of delegate power, and whether weight tracks competence

Table 4 and Figure 7 report how concentrated delegate weight becomes under each delegate-choice rule ($n = 6{,}000$, $M = 16$, the same 16 delegates under every rule).

Under Popularity, the two influencers alone hold 34.3% of participating weight, and the default cap does not bind because it is checked per delegate, not per coalition. Mixed choice cuts the influencers' share by about a third, to 21.8%. Accuracy choice is *not* the most egalitarian (Gini 0.405 against Mixed's 0.382): rewarding demonstrated competence concentrates weight on whoever has the best record, just on someone reliable rather than someone popular.

Figure 8 and Table 5 ask the question Remark 2 poses: does delegated weight follow the optimal weights $\log \frac{\theta_d}{1-\theta_d}$ of Proposition 5? Under Popularity the rank correlation between competence and received weight is *negative* ($\rho = -0.44$) and 61.4% of the weight given to named delegates goes to delegates whose optimal weight is negative. Under Accuracy the correlation is +0.94 and only 2.0% goes to them; a unit of delegated weight lands on a delegate of average competence 0.728 rather than 0.511. Mixed sits in between, and much closer to Popularity on the measure that matters most here (39.1% of weight on below-chance delegates).

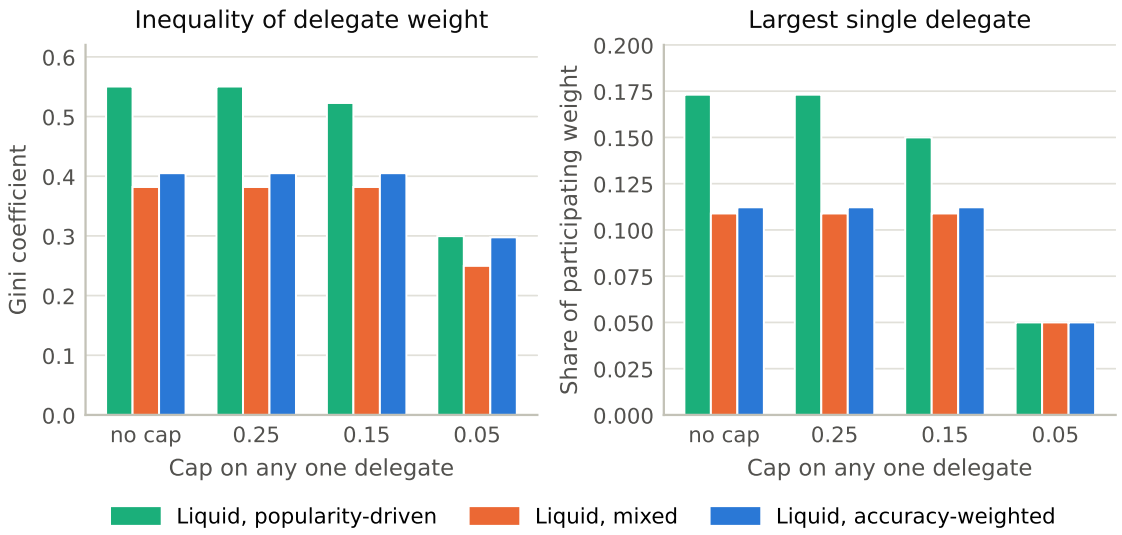


Figure 7: Gini coefficient of delegate weight (left) and the largest single delegate's share of participating weight (right), by delegate-choice rule and cap.

Table 5: How closely delegated weight s_d (share of named-delegate weight) tracks competence θ_d, for the populations of Figure 8.

Delegate-choice rule	Spearman $\rho(\theta_d, s_d)$	$\sum_d s_d\theta_d$	weight on $\theta_d < \frac{1}{2}$
Popularity	−0.44	0.511	61.4%
Mixed	+0.10	0.590	39.1%
Accuracy	+0.94	0.728	2.0%

8.3 Epistemic accuracy and the choice of whom to trust

For $\theta_{\text{pool}} \in \{0.35, 0.47, 0.50, 0.53, 0.65\}$, Table 6 and Figure 9(left) report the probability that the aggregate outcome is correct under direct democracy and the three delegate-choice rules, together with the prediction of Section 6. The predictions match the simulation to within 0.4 points on average and 1.6 points at most, against confidence half-widths of up to 6.3 points.

Four results stand out.

Direct democracy is a knife-edge. At $\theta_{\text{pool}} = 0.35$ and 0.47 it is correct on essentially none of 2,500 bills; at 0.53 it is correct on 99.8%. With roughly 2,400 independent voters, the law of large numbers concentrates the vote so tightly around θ_{pool} that three points of average competence flip the outcome (Proposition 6(i)).

When the population is misinformed, delegation rescues the vote. At $\theta_{\text{pool}} = 0.35$, every delegation rule beats direct democracy's 0.0%: 27.7% (Popularity), 47.8% (Mixed) and 83.5% (Accuracy) — Proposition 6(iii). But only Accuracy is better than a coin flip. Under Popularity and Mixed, the below-chance influencers plus a misinformed pool outweigh the competent delegates.

When the population is informed, delegation costs accuracy. At $\theta_{\text{pool}} = 0.65$, direct democracy is correct on 100% of bills; Accuracy on 99.0%, Mixed on 89.7%, Popularity on 61.4% — Proposition 6(ii). A committee of 14 fallible delegates holding 56% of the cast weight can be wrong together in a way 2,400 independent competent voters practically cannot. For every rule the

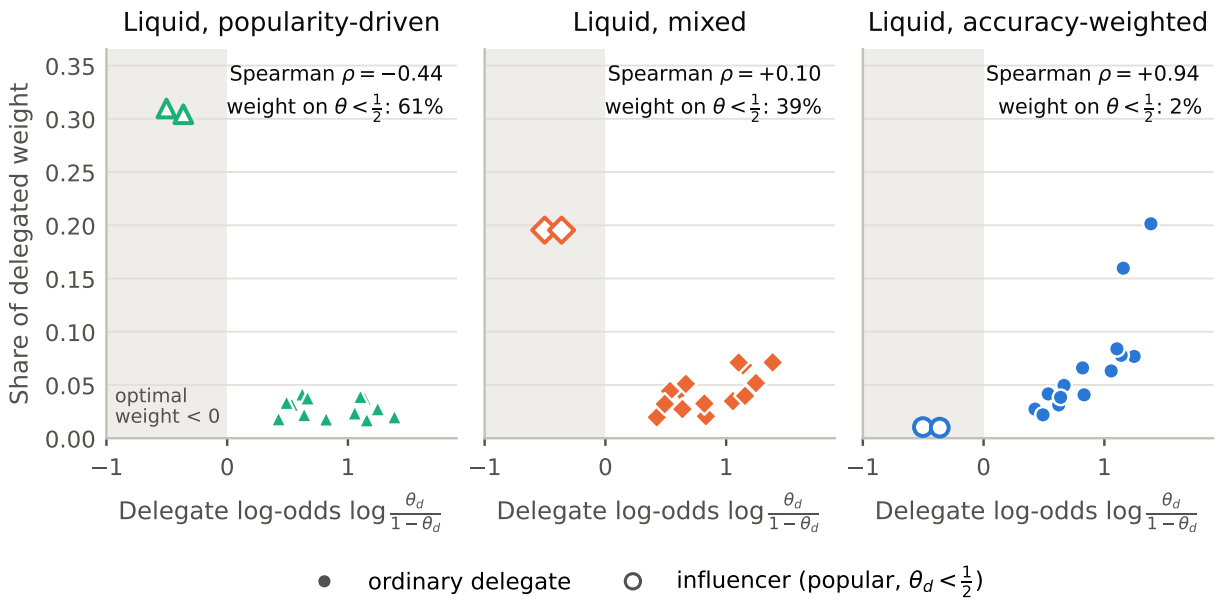


Figure 8: Share of named-delegate weight received by each of the same 16 delegates under each delegate-choice rule, against the delegate's log-odds of competence — the optimal weight of Proposition 5. Hollow markers are the two influencers; the shaded region is where the optimal weight is negative.

Table 6: P(aggregate outcome correct): mean over 10 populations × 250 bills, ±1.96× standard error; in parentheses, the semi-analytic prediction from Proposition 6.

θ_{pool}	Direct democracy	Popularity	Mixed	Accuracy
0.35	0.000 ± 0.000 (0.000)	0.277 ± 0.033 (0.268)	0.478 ± 0.057 (0.485)	0.835 ± 0.022 (0.832)
0.47	0.001 ± 0.001 (0.002)	0.554 ± 0.029 (0.543)	0.689 ± 0.063 (0.683)	0.923 ± 0.013 (0.914)
0.50	0.502 ± 0.019 (0.500)	0.588 ± 0.023 (0.573)	0.752 ± 0.050 (0.755)	0.951 ± 0.016 (0.952)
0.53	0.998 ± 0.001 (0.998)	0.630 ± 0.028 (0.629)	0.778 ± 0.057 (0.779)	0.948 ± 0.016 (0.953)
0.65	1.000 ± 0.000 (1.000)	0.614 ± 0.037 (0.617)	0.897 ± 0.046 (0.902)	0.990 ± 0.006 (0.989)

crossover lies between $\theta_{\text{pool}} = 0.50$ and 0.53.

The delegate-choice rule sets the size of both effects. Reputation weighting shrinks the cost (in Table 6, from 38.6 points under POPULARITY to 1.0 at $\theta_{\text{pool}} = 0.65$) and enlarges the benefit (from 27.7 to 83.5 points at $\theta_{\text{pool}} = 0.35$). It also turns direct democracy's knife-edge into a gradual slope (Proposition 6(iv)): ACCURACY moves only from 83.5% to 99.0% across the whole range.

8.4 Does the cap also protect accuracy?

Fixing $\theta_{\text{pool}} = 0.35$ — where the choice of rule matters most — Table 7 and Figure 9(right) sweep the cap.

The cap does no epistemic work. Under MIXED, accuracy stays between 0.450 and 0.489 at every setting. Under POPULARITY, it slightly *worsens* as the cap tightens, from 0.286 to 0.213. Capping an influencer removes its excess weight from the tally without giving it to anyone. That shrinks the committee's share and raises the relative weight of a pool that, at $\theta_{\text{pool}} = 0.35$, is almost deterministically wrong; the prediction from Proposition 6 reproduces the decline. The cap is a

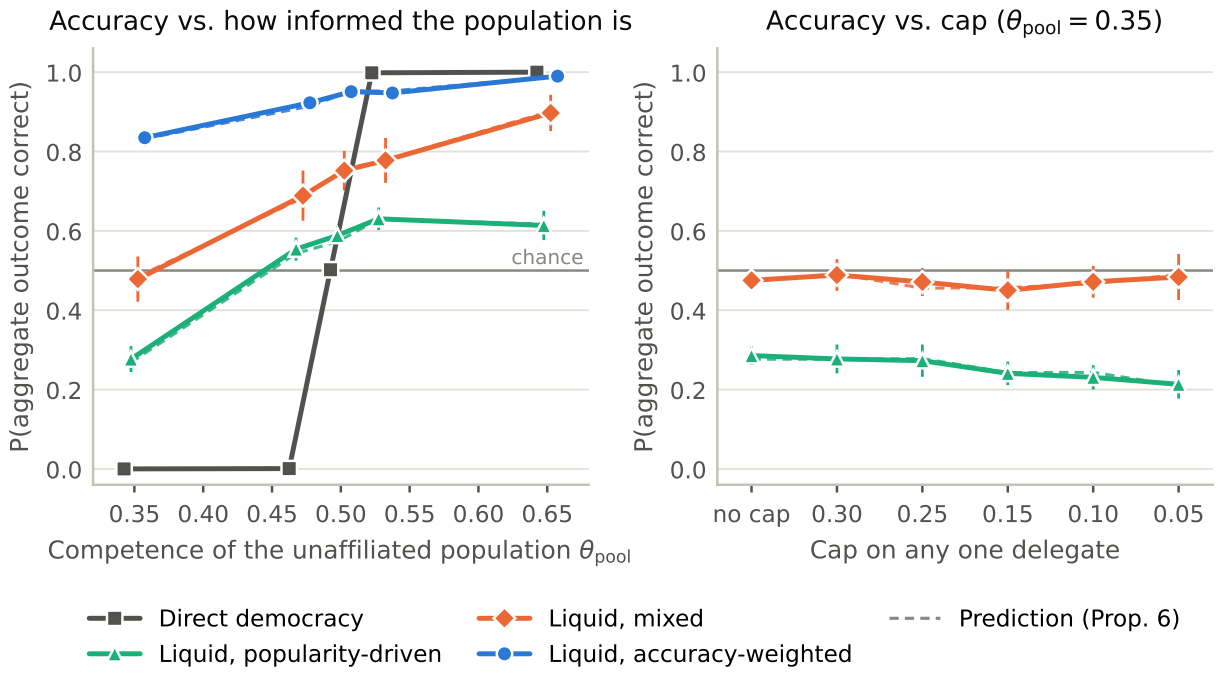


Figure 9: Left: accuracy against the unaffiliated population's competence. Right: accuracy against the cap at $\theta_{pool} = 0.35$ (Section 8.4). Solid: simulation, with 95% intervals; dashed: prediction from Proposition 6. Points at the same θ_{pool} are offset slightly for legibility. Numbers in Tables 6 and 7.

Table 7: P(aggregate outcome correct) against the cap, $\theta_{pool} = 0.35$, mean over 10 populations × 250 bills, with the semi-analytic prediction beneath each row.

Delegate-choice rule	no cap	0.30	0.25	0.15	0.10	0.05
Mixed	0.475	0.489	0.471	0.450	0.472	0.484
Prop. 6	0.473	0.492	0.456	0.457	0.467	0.490
Popularity	0.286	0.277	0.273	0.241	0.231	0.213
Prop. 6	0.277	0.276	0.278	0.243	0.243	0.209

defense against concentration of power (Section 8.2), not a substitute for a competence signal.

8.5 Correlated errors: shocks that fool everyone

Everything so far assumes independent errors, the premise that lets the pool self-correct and lets reputation weighting pay off. The 5G case is the opposite: many otherwise reasonable voters converging on the same wrong belief at once — and false stories spread farther and faster than true ones [Vosoughi et al., 2018]. We model it directly. Each bill carries an i.i.d. indicator $X_b \sim$ Bernoulli(π_{shock}). On an *ordinary* bill ($X_b = 0$) ballots are drawn as in Section 6; on a *shock* bill ($X_b = 1$) every participating voter — named delegates included, whatever their competence — casts the wrong ballot (Figure 10).

Proposition 10 (No configuration of the tally survives a full shock)**.** *On any bill on which every participating voter's ballot is the incorrect side and* participation > 0*, the outcome of Algorithm 1 is incorrect, for every rule triple ρ and every delegation profile δ.*

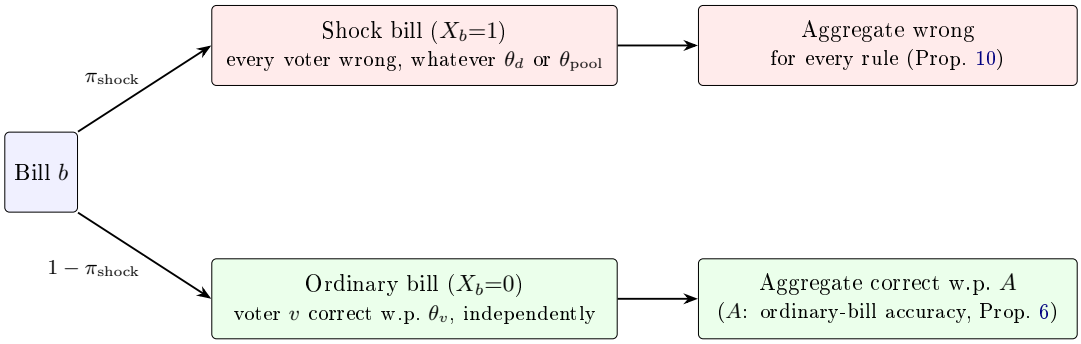


Figure 10: The two branches of a bill under the full-shock model. Every rule loses the shock branch (Proposition 10) and keeps its ordinary-bill accuracy A on the other; Corollary 2 combines them.

Proof. Let y be the correct side and $\bar{y}$ the incorrect one. Every unit of weight reaching totals[·] does so through some terminal delegate's ballot $\beta(d)$ or some citizen's own ballot $\beta_0(i)$ (lines 3–15), both equal to $\bar{y}$; the other destinations, UNASSIGNED and LAPSED, never add to totals[y]. So totals[y] = 0 before APPLY-CAP, which only moves weight from totals[$\beta(d)$] = totals[$\bar{y}$] to totals[UNASSIGNED] (lines 21–24) and therefore leaves totals[y] = 0. Since participation > 0, some other bucket is positive and y does not win. □

The argument uses only that the final comparison respects unanimity, so it holds for any aggregation rule with that property. The only rules that "survive" a full shock are ones that invert ballots, which Remark 2 already excluded.

Corollary 2 (Accuracy under i.i.d. full shocks)**.** *If a rule's accuracy on ordinary bills is A, its accuracy under the full-shock model is exactly* $(1-\pi_{\text{shock}})\,A$.

Proof. By total probability, $\Pr[\text{correct}] = (1-\pi_{\text{shock}})\Pr[\text{correct} \mid X_b{=}0] + \pi_{\text{shock}}\Pr[\text{correct} \mid X_b{=}1] = (1-\pi_{\text{shock}})A + \pi_{\text{shock}} \cdot 0$, using Proposition 10 for the second term and, for the first, that ordinary-bill ballots are generated exactly as when $\pi_{\text{shock}} = 0$. □

In plain words. When a rumour fools every single voter, no sensible counting rule can save the decision: if everyone is wrong in the same way, every way of adding up their votes is wrong too. The best any system can do is to be right on the bills the rumour does not touch.

Remark 4 (Relation to Dietrich and Spiekermann)**.** Dietrich and Spiekermann [2013a] show that when votes are independent only conditional on the "problem" (the state together with the voters' common causes), the probability of a correct majority converges, as the group grows, to $\Pr(\text{problem easy}) + \frac{1}{2}\Pr(\text{problem on the boundary})$ rather than to one. Shock bills are "hard" problems in their sense. Corollary 2 is the finite-committee analogue for Cauce's tally: the large-group limit 1 of the Condorcet Jury Theorem is replaced by the ordinary-bill accuracy A given by Proposition 6.

We sweep $\pi_{\text{shock}} \in \{0, 0.05, 0.10, 0.20, 0.40\}$ at $\theta_{\text{pool}} = 0.65$ and $\theta_{\text{pool}} = 0.35$ (Table 8, Figure 11). Across all 40 cells the simulation differs from Corollary 2 — with A computed from Proposition 6 for each population — by 0.4 points on average and 1.5 at most. The message is the one the corollary states. A shock costs every rule the same *fraction* of its accuracy, so no rule is protected on shocked bills. Each rule keeps its ordinary-bill ranking, scaled down: at $\theta_{\text{pool}} = 0.65$ and $\pi_{\text{shock}} = 0.40$, direct democracy and ACCURACY both fall to 0.600, and POPULARITY to 0.372. Reputation weighting cannot see a shock coming, because a record built from independent errors carries no information about a claim nobody has been tested against before.

Table 8: P(aggregate outcome correct) under i.i.d. full shocks, mean over 10 populations × 250 bills, ±1.96× standard error.

θ_{pool}	π_{shock}	Direct democracy	Popularity	Mixed	Accuracy
0.65	0.00	1.000 ± 0.000	0.649 ± 0.043	0.935 ± 0.026	0.990 ± 0.006
	0.05	0.955 ± 0.006	0.592 ± 0.034	0.866 ± 0.039	0.946 ± 0.009
	0.10	0.904 ± 0.013	0.562 ± 0.030	0.857 ± 0.017	0.890 ± 0.010
	0.20	0.795 ± 0.022	0.498 ± 0.021	0.740 ± 0.025	0.798 ± 0.016
	0.40	0.600 ± 0.015	0.372 ± 0.021	0.544 ± 0.020	0.600 ± 0.023
0.35	0.00	0.000 ± 0.000	0.268 ± 0.014	0.477 ± 0.020	0.802 ± 0.024
	0.05	0.000 ± 0.000	0.256 ± 0.029	0.452 ± 0.052	0.768 ± 0.028
	0.10	0.000 ± 0.000	0.248 ± 0.020	0.442 ± 0.036	0.748 ± 0.025
	0.20	0.000 ± 0.000	0.203 ± 0.025	0.372 ± 0.037	0.670 ± 0.029
	0.40	0.000 ± 0.000	0.162 ± 0.028	0.262 ± 0.031	0.478 ± 0.022

8.6 Partial shocks: when does delegation pay?

A full shock is a caricature: real misinformation fools some voters and not others, and plausibly fools the less informed more often — analytic thinking predicts better discernment of false headlines [Pennycook and Rand, 2019] and lower conspiracy belief across 145 samples [Yelbuz et al., 2022] — though on identity-laden questions greater skill can *increase* polarization [Kahan et al., 2017]. We therefore let a shock fool each voter v independently with probability σ_v, its *susceptibility*; a voter who is not fooled votes with its ordinary competence.

Proposition 11 (Effective competence under a partial shock)**.** *On a shock bill with susceptibilities (σ_v), voter v is correct with probability $\theta'_v = (1-\sigma_v)\,\theta_v$, independently across voters. Hence Proposition 6, and its finite-pool refinement, apply to shock bills with every θ_v replaced by θ'_v.*

Proof. v is correct iff it is not fooled (probability $1-\sigma_v$) and, not being fooled, votes correctly (probability θ_v); being fooled and voting are drawn independently of each other and of other voters. □

Corollary 3 (Critical susceptibility)**.** *If every pool member has susceptibility σ and $\theta_{\text{pool}} > \frac{1}{2}$, direct democracy's accuracy on shock bills tends to 1 if $\sigma < \sigma^* = 1 - \frac{1}{2\theta_{\text{pool}}}$ and to 0 if $\sigma > \sigma^*$. Each individual voter falls below chance once $\sigma_v > 1 - \frac{1}{2\theta_v}$.*

Proof. By Proposition 11 and Proposition 6(i), the limit is 1 iff $(1-\sigma)\theta_{\text{pool}} > \frac{1}{2}$, i.e. iff $\sigma < 1 - \frac{1}{2\theta_{\text{pool}}}$; likewise for a single voter. □

Corollary 4 (Insurance break-even)**.** *Suppose a fraction π_{shock} of bills are shock bills with a given susceptibility profile. Let A_R, A_D be the ordinary-bill accuracies of a delegation rule R and of direct democracy, and A^s_R, A^s_D their shock-bill accuracies, with $A_R < A_D$ and $A^s_R > A^s_D$. Then R is more accurate than direct democracy overall iff*

$$\pi_{\text{shock}} \;>\; \pi^* \;=\; \frac{A_D - A_R}{(A_D - A_R) + (A^s_R - A^s_D)}.$$

Proof. By total probability, R's accuracy is $(1-\pi_{\text{shock}})A_R + \pi_{\text{shock}}A^s_R$, and similarly for direct democracy; rearranging $(1-\pi_{\text{shock}})A_R + \pi_{\text{shock}}A^s_R > (1-\pi_{\text{shock}})A_D + \pi_{\text{shock}}A^s_D$ gives the condition. □

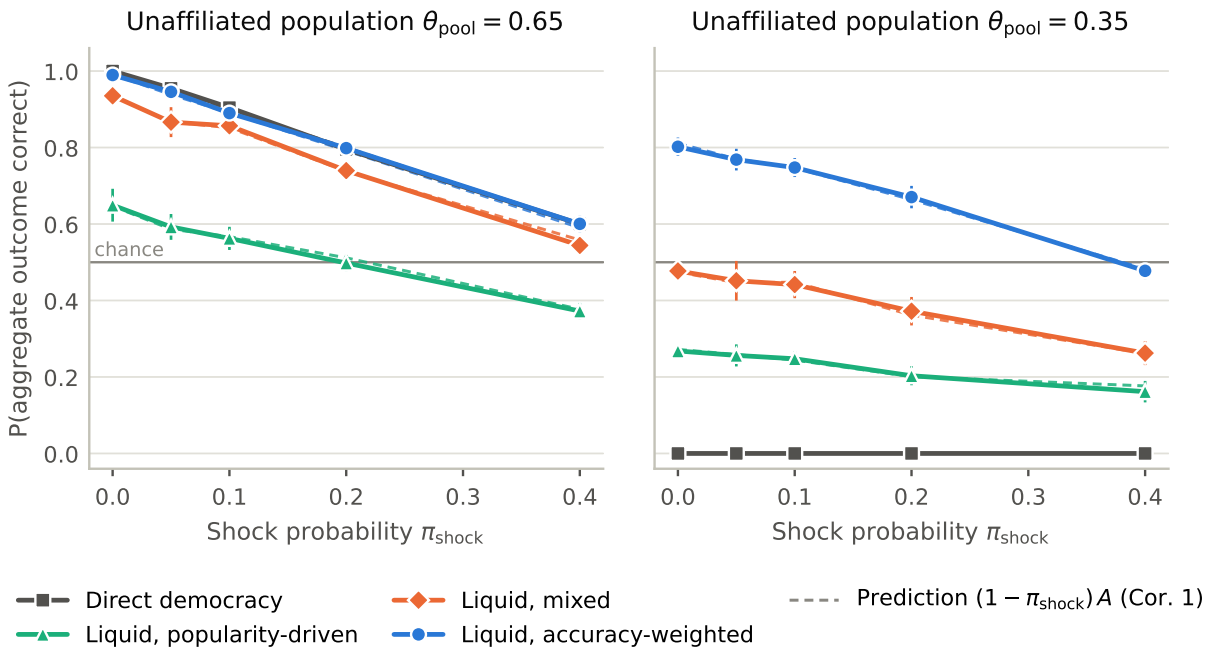


Figure 11: Accuracy against the full-shock probability π_{shock} at two population competences. Solid: simulation with 95% intervals; dashed: Corollary 2, $(1 - \pi_{\text{shock}})A$, with A from Proposition 6 for each simulated population.

In plain words. A rumour that fools only some voters behaves like a tipping point. With voters who are right about 65% of the time, the crowd survives as long as fewer than about one in four (23%) is fooled; past that point it is wrong almost surely. Delegates slide down more gently. That makes delegation work like insurance: a small premium paid in ordinary times, in accuracy given up, buys cover when a shock strikes. Whether the policy is worth buying depends on how often shocks come (Figure 4).

We call $A_D - A_R$ the *premium* (accuracy given up in normal times) and $A_R^s - A_D^s$ the *gain* (accuracy bought when a shock hits). We make every bill a shock bill, fix $\theta_{\text{pool}} = 0.65$, and sweep σ under two susceptibility profiles: *uniform* ($\sigma_v = \sigma$ for everyone), and *competence-linked*, $\sigma_v = \min\left(1, \sigma \frac{1-\theta_v}{1-\theta_{\text{pool}}}\right)$, in which the pool keeps susceptibility σ — so direct democracy is identical under both — while delegates better informed than the pool are fooled less often, and influencers more often. Table 9 and Figure 12 report the results. They match the predictions of Propositions 6 and 11 to within 0.3 points on average and 1.4 at most.

Direct democracy falls off a cliff exactly where Corollary 3 puts it: correct on 97.3% of shock bills at $\sigma = 0.20$, 11.2% at 0.25, and none at 0.30. Below σ^* its self-correction still wins (at $\sigma = 0.20$: 97.3% against 76.6% for ACCURACY). Delegation degrades gradually, but it does not stay above chance for long. Under uniform susceptibility ACCURACY keeps 62.7% at $\sigma = 0.25$ but only 47.9% — just below a coin flip — at 0.30; under competence-linked susceptibility it keeps 73.3% and 63.2%. By $\sigma = 0.40$ every rule is below chance (the best, competence-linked ACCURACY, at 39.3%): the pool is almost deterministically wrong, and more and more delegates' effective competence has fallen below one half — under uniform susceptibility even the best delegate, with $\theta_d = 0.85$, crosses at $\sigma \approx 0.41$ (Corollary 3 applied to each delegate).

Table 10 converts this into the question a citizen or a designer would ask. Reputation-weighted delegation pays a premium of 1.0 point on ordinary bills; against a shock that fools a quarter to

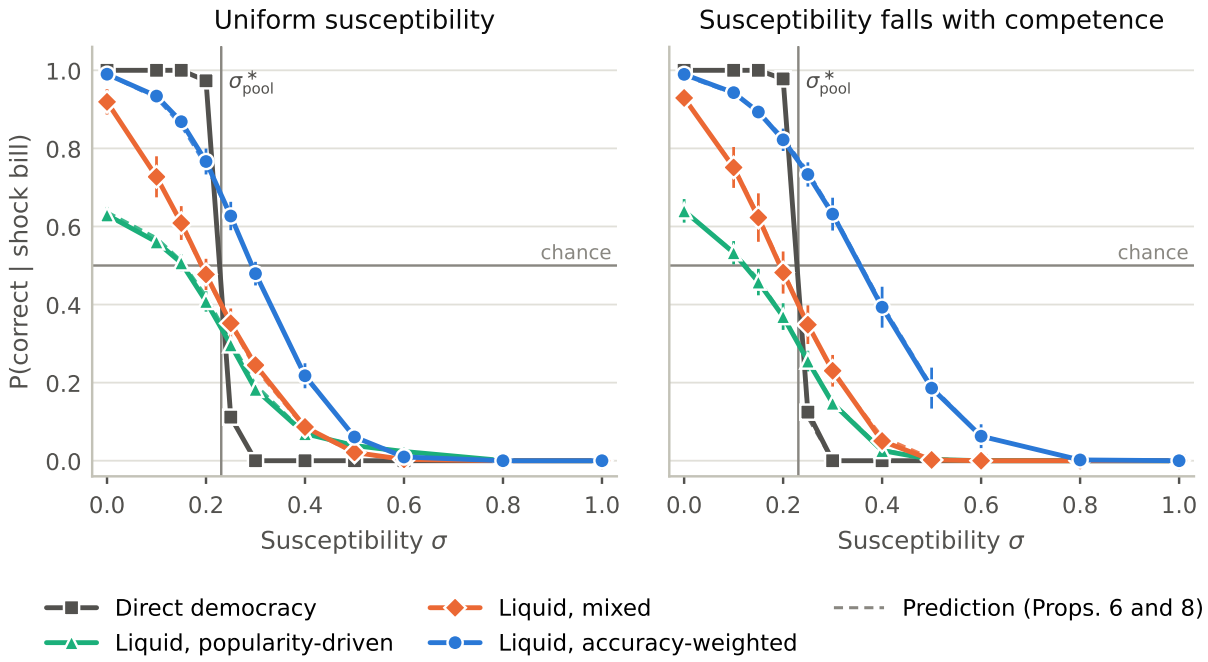


Figure 12: Accuracy on shock bills against susceptibility σ ($\theta_{\text{pool}} = 0.65$). The vertical line is the critical susceptibility $\sigma^* = 1 - \frac{1}{2\theta_{\text{pool}}} \approx 0.231$ of Corollary 3. Solid: simulation with 95% intervals; dashed: prediction from Propositions 6 and 11. Numbers in Table 9.

a third of the pool it outperforms direct democracy overall once about 2% of bills carry such a shock ($\pi^* = 0.020$, 95% CI 0.015–0.027, under uniform susceptibility at $\sigma = 0.3$; 0.016, CI 0.011–0.022, competence-linked). MIXED needs about a quarter of bills to be shocked, and POPULARITY about two-thirds. The break-even rises with shock strength, because strong enough shocks defeat delegation too: at $\sigma = 0.5$ it is 0.141 (uniform) and 0.053 (competence-linked) for ACCURACY, and close to one for the other rules. This is the paper's central quantitative claim, and it comes with three qualifications. On shocked bills even the best rule is only near a coin flip at $\sigma = 0.3$ (47.9% uniform, 63.2% competence-linked): the insurance makes shocked bills less catastrophic than direct democracy's 0%, not safe. The one-point premium holds for a well-informed population and long, accurate records (Section 7); at $\theta_{\text{pool}} = 0.53$ ACCURACY's premium in Table 6 is already 5.0 points. And against milder misinformation ($\sigma < \sigma^*$) the insurance never pays. Within those limits, delegation is insurance against misinformation that crosses the critical susceptibility, not an improvement in normal times, and reputation weighting is what makes that insurance cheap enough to be worth buying.

9 What the Score and the Cap Cannot Guarantee

Sections 5–8 took two things at face value. A delegate's record was treated as a noisy but honest estimate of her competence on the questions citizens care about, and the cap was taken to bound concentration. This section tests both assumptions formally. A record ranks competent delegates only slowly (Section 9.1). A delegate who writes her own claims can make her record look excellent (Section 9.2). Competing for delegations rewards overstated confidence (Section 9.3). A record averages over ordinary and shocked conditions, which can invert the ranking that matters (Sec-

Table 9: $P(\text{correct} \mid \text{shock bill})$ against susceptibility σ, $\theta_{\text{pool}} = 0.65$, mean over 10 populations $\times$ 250 bills (largest 95% half-width among the rows shown: 5.4 points).

Susceptibility	σ	Direct democracy	Popularity	Mixed	Accuracy
Uniform	0.00	1.000	0.630	0.919	0.990
	0.10	1.000	0.560	0.727	0.934
	0.20	0.973	0.408	0.477	0.766
	0.25	0.112	0.297	0.352	0.627
	0.30	0.000	0.183	0.245	0.479
	0.40	0.000	0.069	0.086	0.218
	0.50	0.000	0.039	0.021	0.061
	0.60	0.000	0.024	0.003	0.010
Competence-linked	0.00	1.000	0.640	0.929	0.990
	0.10	1.000	0.533	0.751	0.943
	0.20	0.978	0.369	0.482	0.822
	0.25	0.124	0.255	0.348	0.733
	0.30	0.000	0.147	0.230	0.632
	0.40	0.000	0.026	0.050	0.393
	0.50	0.000	0.003	0.002	0.186
	0.60	0.000	0.000	0.000	0.063

tion 9.4). Competence is not the same as sharing a citizen's ends (Section 9.5). And delegation makes a bill cheap to capture, in a way the default cap does not prevent (Section 9.6). For each, we state what can be proved and what the design can do about it. In Strathern's (1997, p. 308) formulation of Goodhart's law, "when a measure becomes a target, it ceases to be a good measure"; the results below are concrete mechanisms by which a score that confers power could do so.

9.1 How long must a record be?

Section 7 set the noise in the public estimate of competence to what a record of about 125 resolved predictions would give. The next result makes the dependence on record length explicit.

Proposition 12 (Record length)**.** *Let a delegate with competence θ attach the same report p to each of n predictions about her own ballot, with independent outcomes. Her Brier score is the mean of n i.i.d. scores with mean and variance*

$$\mu(\theta, p) = \theta(1-p)^2 + (1-\theta)p^2, \qquad v(\theta, p) = \theta(1-\theta)(2p-1)^2.$$

Let two such delegates a and b have independent records of length n, with $\Delta = \mu_b - \mu_a > 0$ and $v_a + v_b > 0$. Then:

(i) $\sqrt{n}\,\big((B_b - B_a) - \Delta\big)/\sqrt{v_a + v_b}$ *converges in distribution to a standard normal, so that* $\Pr[B_a < B_b] \approx \Phi\big(\sqrt{n}\,\Delta/\sqrt{v_a + v_b}\big)$*, and a record of length*

$$n_{95} = \left\lceil \left(z_{0.95}\sqrt{v_a + v_b}\,/\,\Delta \right)^2 \right\rceil$$

ranks the two correctly with probability about 0.95*;*

(ii) without approximation, $\Pr[B_a \geq B_b] \leq \exp(-n\Delta^2/2)$.

Proof. Each score equals $(1-p)^2$ with probability θ and p^2 otherwise, which gives μ and $v = \theta(1-\theta)\big(p^2 - (1-p)^2\big)^2 = \theta(1-\theta)(2p-1)^2$. $B_b - B_a$ is the mean of n i.i.d. differences with mean

Table 10: Break-even shock frequency π^* (Corollary 4): the fraction of bills that must be shock bills of susceptibility σ for each delegation rule to match direct democracy overall, from the populations of Table 9 ($\theta_{\text{pool}} = 0.65$). A_D^s is direct democracy's accuracy on those shock bills. Premiums paid on ordinary bills: POPULARITY 37.0 (uniform) and 36.0 (competence-linked) points, MIXED 8.1 and 7.1, ACCURACY 1.0 and 1.0. For ACCURACY, 95% bootstrap intervals over populations. For $\sigma < \sigma^* \approx 0.231$ no rule ever breaks even.

Susceptibility	σ	A_D^s	Popularity	Mixed	Accuracy [95% CI]
Uniform	0.25	11.2%	0.667	0.252	0.019 [0.014, 0.025]
	0.30	0.0%	0.670	0.248	0.020 [0.015, 0.027]
	0.40	0.0%	0.843	0.483	0.044 [0.031, 0.059]
	0.50	0.0%	0.905	0.792	0.141 [0.100, 0.195]
Competence-linked	0.25	12.4%	0.733	0.240	0.017 [0.012, 0.023]
	0.30	0.0%	0.710	0.235	0.016 [0.011, 0.022]
	0.40	0.0%	0.933	0.584	0.026 [0.018, 0.035]
	0.50	0.0%	0.991	0.973	0.053 [0.035, 0.076]

Table 11: Record length needed for a Brier score to rank two delegates correctly with probability 0.95 (Proposition 12(i)). Each delegate's constant report is in parentheses; $\mathbb{E}B$ is the expected Brier score. The last column is a Monte Carlo estimate of the probability of a correct ranking at n_{95} (20,000 simulated pairs of records).

Comparison	θ_a (report)	θ_b (report)	$\mathbb{E}B_a$	$\mathbb{E}B_b$	n_{95}	MC check
Competent vs. confidently wrong	0.65 (0.65)	0.40 (0.80)	0.2275	0.4000	10	0.958
Competent vs. competent	0.85 (0.85)	0.70 (0.70)	0.1275	0.2100	39	0.950
Competent vs. competent	0.75 (0.75)	0.60 (0.60)	0.1875	0.2400	56	0.948
Competent vs. competent	0.70 (0.70)	0.65 (0.65)	0.2100	0.2275	478	0.950
Competent vs. competent	0.62 (0.62)	0.60 (0.60)	0.2356	0.2400	3,239	0.950

Δ and variance $v_a + v_b$, so (i) is the central limit theorem. Each difference lies in $[-1, 1]$, so (ii) is Hoeffding's inequality, $\Pr[B_b - B_a \le 0] \le \exp\big(-2n\Delta^2/2^2\big)$. □

In plain words. A record is like a batting average. A few games are enough to expose a batter who almost never hits. Telling two good batters apart takes a few weeks if one is clearly better, and several seasons if they are close — and a season of policy predictions can last years.

Table 11 evaluates it. A confidently wrong delegate is exposed quickly. About ten predictions separate a delegate of competence 0.40 who always reports 0.80 from a calibrated delegate of competence 0.65, because Proposition 4(iii) puts the former's expected score far above $\frac{1}{4}$. Ranking competent delegates is slow. It takes 39 predictions to separate competences 0.85 and 0.70, and 56 for 0.75 and 0.60, but 478 for 0.70 and 0.65 and more than 3,000 for 0.62 and 0.60. The distribution-free bound (ii) is far more conservative — about 200 predictions for the first comparison — but scales the same way in the gap, as $1/\Delta^2$ for fixed variance.

Three consequences follow. First, the reputation layer works mainly as a filter against confidently wrong delegates, and only slowly as a fine ranking. Selecting the single best record makes this worse: in a simulated tournament of 300 forecasters, Aldous [2021] found, in a somewhat plausible model, that the winner was most likely to rank around 100th in true accuracy and that the most accurate contestants never won — a trade-off between a score's mean and its variance, in which a somewhat less accurate but lucky forecaster usually tops a large field. That is what the insurance argument

needs: what separates ACCURACY from POPULARITY in Table 5 is above all the weight moved off below-chance delegates (61.4% against 2.0%). Second, per-topic records would be short. The platform's score pools every prediction a delegate makes (Section 3); a per-topic record, which topic scoping invites, accrues T times more slowly for a delegate active on T topics. Pooling records across topics trades this variance for bias wherever competence is topic-specific. The standard compromise is a hierarchical estimate; in the geopolitical forecasting tournaments, an item-response model that also estimates each question's difficulty predicted later accuracy better than raw Brier scores did [Bo et al., 2017]. Third, records entrench incumbents. A newcomer is indistinguishable from noise for dozens to hundreds of resolved predictions, which may take years when policy claims take that long to resolve — one mechanism for the oligarchic tendency Michels [1915] found in mass parties. One design response is to let every citizen, not only delegates, attach scored predictions to her own direct ballots, so that a record exists before anyone delegates to her, as open forecasting tournaments allow.

9.2 Who writes the question? Relative scores

A Cauce delegate writes the claim she predicts (Section 3); Proposition 4 took the claim as given.

Proposition 13 (Author-chosen claims and relative scores)**.** *(i) Let r be a claim's probability given public information. A forecaster who knows nothing more and reports r has expected Brier score $r(1-r)$. By predicting only claims with $\min(r, 1-r) \le \varepsilon \le \frac{1}{2}$, she obtains an expected score of at most $\varepsilon(1-\varepsilon)$, which the inversion of Proposition 4(ii) reads as competence of at least $1-\varepsilon$.*

(ii) Let r instead be a reference forecast of the same claim, fixed independently of the delegate's report p (for instance, the median of other forecasters' reports), and define the relative score *$S = (p-o)^2 - (r-o)^2$. If the claim has probability q,*

$$\mathbb{E}[S] = (p-q)^2 - (r-q)^2.$$

Hence S is strictly proper; a forecaster who copies the reference scores exactly 0 on every claim, however easy or hard; and truthful reporting earns $-(r-q)^2$, which measures how far her information moves the forecast away from the reference.

Proof. (i) By Proposition 4 with $p = q = r$, the expected score is $r(1-r)$, which increases in $\min(r, 1-r)$ on $[0, \frac{1}{2}]$. The inverse map gives $\frac{1}{2}\big(1 + \sqrt{1 - 4\varepsilon(1-\varepsilon)}\big) = \frac{1}{2}\big(1 + (1-2\varepsilon)\big) = 1-\varepsilon$. (ii) By Proposition 4, $\mathbb{E}[(p-o)^2] = (p-q)^2 + q(1-q)$ and $\mathbb{E}[(r-o)^2] = (r-q)^2 + q(1-q)$; subtract. Since r does not depend on p, the unique minimizer is $p = q$. □

In plain words. Anyone can look like a prophet by predicting that the sun will rise tomorrow. If delegates choose their own questions, they can choose easy ones. The cure is to score each prediction against what others predicted for the same question, so that only knowing more than the crowd earns credit. A neutral body must word and resolve the questions.

Part (i) formalizes the observation that raw Brier scores "are content dependent, so people may be assigned higher or lower 'expertise' scores as a function of the events they choose to forecast" [Bo et al., 2017]. Cauce's rule of one prediction per vote stops a delegate from choosing *which bills* to forecast. It does not stop her from choosing *which claim* to forecast about each bill, and that is enough. Forecasting tournaments handle this by scoring forecasters relative to others on the same questions; the Good Judgment Project "standardized Brier scores within each question" to minimize the effects of question difficulty [Mellers et al., 2015]. Part (ii) shows that the difference

form keeps the score proper. It changes what is ranked, however: a relative score does not map to competence through Proposition 4(ii), but measures information beyond the reference, so the ranking result of Section 5 holds only for the absolute score. Ratio-form skill scores are generally improper [Gneiting and Raftery, 2007, Section 2.3], although the benefit of hedging the Brier skill score becomes negligible as the number of independent forecasts grows [Murphy, 1973].

Two design consequences follow. A claim must be forecast by more than its author: the median of the forecasts that other delegates and citizens attach to the same claim, frozen when voting opens, could serve as the reference, provided the forecasters who form it are verified individuals, since an attacker who controls many identities could otherwise move the reference and inflate a confederate's relative score. A study of a deployed system points the same way: in simulations of the algorithm behind Community Notes, "a small minority (5-20%) of bad raters can strategically suppress targeted helpful notes" [Truong et al., 2025]. And a neutral body must word and resolve claims, as tournaments do. Relative scoring also has a property that matters for the insurance argument. When a shock fools most forecasters, the reference is wrong, and a delegate who resists the shock earns a strongly negative — that is, good — relative score. For a true claim with reference $r = 0.2$, reporting $p = 0.8$ earns $0.04 - 0.64 = -0.60$. The record therefore rewards exactly the independence the insurance needs. Scores that reward agreement with other forecasters, such as peer prediction, do the opposite (Section 12.12).

9.3 Rank brings power: competition and overconfidence

Proposition 4(i) is about a forecaster who minimizes her expected score. A Cauce delegate is not paid her score. She is paid in delegations, which citizens allocate by comparing scores, and a non-linear reward breaks propriety.

Proposition 14 (Non-linear rewards distort reports)**.** *Let a delegate's payoff be $g(B)$, where B is her Brier score over n predictions. Fix one prediction whose claim has probability $q \in (\frac{1}{2}, 1)$, and let $R \in [0, \frac{n-1}{n}]$ be the other predictions' contribution to B, with any distribution independent of that claim's outcome. Suppose g is continuously differentiable on an open interval I that contains $R + x/n$ for every $x \in [0,1]$ and every R in the support. At the truthful report $p = q$,*

$$\frac{\partial}{\partial p}\,\mathbb{E}\,g(B)\Big|_{p=q} \;=\; \frac{2q(1-q)}{n}\Big(\mathbb{E}\,g'\big(R + \tfrac{q^2}{n}\big) - \mathbb{E}\,g'\big(R + \tfrac{(1-q)^2}{n}\big)\Big).$$

If g is strictly convex on I, the derivative is positive, and the delegate gains by reporting a probability more extreme than her belief. If g is strictly concave on I, she gains by reporting one less extreme. If g is affine, the derivative is zero, as Proposition 4(i) requires. By symmetry, the same holds for $q < \frac{1}{2}$, with "more extreme" meaning closer to 0.

Proof. $B = R + (p-o)^2/n$, so $\frac{\partial}{\partial p} g(B) = g'(B)\,2(p-o)/n$; differentiation under the expectation is justified because g' is continuous, hence bounded, on the compact range of B, which lies in I. At $p = q$, with $o = 1$ (probability q) or $o = 0$ (probability $1-q$) independent of R, the expectation is $\frac{2}{n}\big[q(q-1)\,\mathbb{E}g'(R + \frac{(1-q)^2}{n}) + (1-q)q\,\mathbb{E}g'(R + \frac{q^2}{n})\big]$, which is the stated expression. For $q > \frac{1}{2}$, $q^2 > (1-q)^2$. A strictly convex g has strictly increasing g', so the bracket is positive and some report $p > q$ earns a strictly higher expected payoff. Concavity reverses the sign, and an affine g gives zero. □

In plain words. When rewards climb ever faster with rank — as they do when followers flock to the best-looking record — it pays to sound bolder than you are. A student graded on a curve does better by answering boldly than by admitting doubt. The pull is small for any one prediction and fades as the

record grows, but nothing in the score itself removes it.

For a smooth g, the bracket is of order $1/n$, so the incentive to distort any one prediction fades as the record grows — the analogue of Murphy's result for skill scores. Competition among close rivals is not smooth, however. A forecaster who wants to beat another "typically should report more extreme probabilities, exaggerating toward zero or one" [Lichtendahl and Winkler, 2007]. When only the best-scoring forecaster wins a prize, "forecasters can often increase their probability of having the highest score by reporting more extreme beliefs" [Witkowski et al., 2023], who give lottery-based selection mechanisms that restore incentive compatibility. In models of professional forecasting, contests make forecasts excessively differentiated, while reputational concerns make forecasts less informative, and shade them toward the prior when the market expects honesty [Ottaviani and Sørensen, 2006]. And Kadane and Winkler [1988] show that scoring-rule reports equal beliefs when utility is linear, but that with non-linear utility, and especially when the forecaster has other stakes in the event, elicited probabilities can depart from beliefs; a Cauce delegate always has such a stake — the bill she voted on (Section 12.7).

Our own Accuracy rule is an instance. Suppose citizens apply it to the competence a record reveals through Proposition 4(ii). A delegate's share of new delegations is then increasing in $g(B) = \hat{\theta}(B)^6$, with $\hat{\theta}(B) = \frac{1}{2}(1 + \sqrt{1-4B})$, and approximately proportional to it when she is one of many. Writing $s = \sqrt{1-4B}$, one finds $g''(B) = 6\hat{\theta}^4(4s-1)/s^3$ for $B < \frac{1}{4}$, so g is strictly convex wherever $\hat{\theta} > \frac{5}{8}$, that is, for $B < \frac{15}{64} \approx 0.234$; this covers most of our competent delegates ($\theta_d \sim \mathrm{Unif}(0.58, 0.85)$). For a delegate whose score is sure to stay in that range — as a long record makes overwhelmingly likely for one with competence well above $\frac{5}{8}$ — Proposition 14 applies, and it predicts overstated confidence from exactly the delegates the rule favours. Overconfidence is not free — by Proposition 4(i) it costs expected score, and the cost is paid on every prediction — but nothing in the score alone removes the incentive. Lottery mechanisms are hard to justify for allocating political weight. The realistic levers are long records, relative scores (Proposition 13), and displaying calibration (stated probability against observed frequency) rather than a single number, so that systematic overstatement is visible to the citizens who reward it (Section 12.3).

9.4 A record averages over states

A Brier score averages over the conditions under which a delegate's predictions were made, and most of them are ordinary. Proposition 11 gives each voter's competence on a shock bill separately, which is enough to see when the average misleads.

Corollary 5 (When the better record is worse in a shock)**.** *Let delegates a and b have competences $\theta_a > \theta_b$ and susceptibilities σ_a, σ_b. On a shock bill, a is less likely than b to vote correctly iff*

$$\sigma_a \;>\; 1 - (1-\sigma_b)\,\frac{\theta_b}{\theta_a}.$$

If susceptibility is a non-increasing function of competence, $\sigma_d = s(\theta_d)$, no delegate is more likely to be correct on a shock bill than a delegate of higher competence. Both susceptibility profiles of Section 8.6 are of this kind.

Proof. By Proposition 11, a is correct with probability $(1-\sigma_a)\theta_a$ and b with $(1-\sigma_b)\theta_b$; rearranging $(1-\sigma_a)\theta_a < (1-\sigma_b)\theta_b$ gives the condition. If s is non-increasing and $\theta_a > \theta_b$, then $1 - s(\theta_a) \geq 1 - s(\theta_b) \geq 0$, so $(1-\sigma_a)\theta_a \geq (1-\sigma_b)\theta_b$. Uniform susceptibility is constant, and the competence-linked profile $\min\!\big(1, \sigma\frac{1-\theta}{1-\theta_{\text{pool}}}\big)$ is non-increasing in θ. □

In plain words. A sailor with a good record in calm seas may be the worst one in a storm. If the better record belongs to the more easily fooled sailor, a score that averages over fair and foul weather can rank first the delegate you least want at the helm when a rumour hits.

The threshold is not demanding. A delegate of competence 0.85 who is fooled by a shock more than about 39% of the time ($1 - 0.8 \times 0.65/0.85 \approx 0.388$) is worse on shock bills than one of competence 0.65 fooled 20% of the time, yet her record, built mostly on ordinary bills, ranks her first. Gilbert et al. [2026] prove the general form of this warning in a model where a public signal changes voters' competences after delegations are fixed: "delegation based on average competence is not a safe scalarisation", since it "can violate do-no-harm", and a beneficial rule may even send votes to voters with lower average competence. Their conservative mechanism delegates only to neighbours who are better "in every signal state". Corollary 5 shows that our simulations sit on the safe side by construction, because both susceptibility profiles make susceptibility fall, or stay flat, as competence rises; Section 12.6 reviews evidence that on identity-laden questions the opposite can hold. A design response in their spirit is to keep a second record for claims made under contested conditions — claims later found to have been the target of a misinformation campaign, for example — and to rank delegates by the worse of the two. That approximates "better in every state", at the price of even shorter records (Proposition 12).

9.5 Competence is not alignment

Section 6 assumed that every bill has one correct side, the one every voter would choose if fully informed — the common-interest assumption behind the Condorcet Jury Theorem. Much political disagreement is about ends as well as facts.

Proposition 15 (Competence versus alignment)**.** *On a bill, let y_i be the side citizen i would choose if fully informed and y_d the side delegate d would choose. Suppose* $\Pr[y_d = y_i] = \alpha$ *(the delegate's* alignment *with i) and, independently, d's ballot matches y_d with probability θ.*

(i) d's ballot matches y_i with probability $\theta_{\text{eff}} = \alpha\theta + (1-\alpha)(1-\theta)$, and $\theta_{\text{eff}} - \frac{1}{2} = (2\alpha - 1)(\theta - \frac{1}{2})$.

(ii) If $\alpha < \frac{1}{2}$, then $\theta_{\text{eff}} < \frac{1}{2}$ for every $\theta > \frac{1}{2}$, and θ_{eff} strictly decreases in θ. Among misaligned delegates, the more competent one is worse for i, and its optimal weight for i in the sense of Proposition 5 is negative.

(iii) Suppose i votes directly with competence θ_i relative to y_i, with errors independent of d's. Then the probability a that their ballots agree satisfies $a - \frac{1}{2} = 2(\theta_i - \frac{1}{2})(\theta_{\text{eff}} - \frac{1}{2})$. For a citizen who is better than chance, an agreement rate above one half certifies $\theta_{\text{eff}} > \frac{1}{2}$.

Proof. (i) By total probability over whether $y_d = y_i$; expanding, $\alpha\theta + (1-\alpha)(1-\theta) - \frac{1}{2} = 2\alpha\theta - \alpha - \theta + \frac{1}{2} = (2\alpha - 1)(\theta - \frac{1}{2})$. (ii) If $2\alpha - 1 < 0$, the right-hand side of (i) is negative for $\theta > \frac{1}{2}$ and strictly decreasing in θ; the optimal weight $\log \frac{\theta_{\text{eff}}}{1-\theta_{\text{eff}}}$ is negative when $\theta_{\text{eff}} < \frac{1}{2}$. (iii) The ballots agree iff both match y_i or neither does, so $a = \theta_i\theta_{\text{eff}} + (1-\theta_i)(1-\theta_{\text{eff}})$; expanding as in (i) gives the identity. □

In plain words. A brilliant navigator who wants to sail to a different port will get you there efficiently — to the wrong port. A Brier record shows how well a delegate reads the sea, not where she wants to go. Comparing her votes with your own, on the bills where you voted yourself, is one rough check on the destination, if you yourself read the sea reasonably well.

The scored claims concern facts, not ends, so under the stylization of Proposition 4(ii) the Brier score is a function of θ alone. It cannot tell apart two delegates who differ only in α. When $\alpha = 1$,

$\theta_{\text{eff}} = \theta$, so the common-interest setting of Sections 6–8 is the best case for the informativeness of the score. We do not model heterogeneous ends — with them, "the correct outcome" is no longer defined for the electorate as a whole — so the results there should not be read as bounds for value-laden topics. In representation theory, substantive representation means "acting in the interest of the represented, in a manner responsive to them" [Pitkin, 1967, p. 209], which competence alone does not secure. In Mansbridge's selection model, a citizen can rely on a representative whose goals already align with hers *and* who has a verifiable reputation for competence and honesty [Mansbridge, 2009]; Cauce verifies only competence. Citizens delegate in the first place because informing themselves costs more than it is worth to them [Downs, 1957], and that same cost makes alignment hard for them to judge. Evidence from blockchain governance is a warning for a design whose interface is a ranking: in an analysis of 14 DAO forums, delegations "are frequently misaligned with token holders' expressed priorities", and "current ranking-based interfaces exacerbate power concentration" [Messias and Ide, 2025].

Part (iii) suggests a remedy that needs no new data. Next to each delegate's Brier score, the platform can show each citizen a *personal agreement rate*: on the bills where she voted directly (overrides included), the share on which the delegate cast the same ballot. This is the logic of voting advice applications, which match "users' policy-preferences with the positions of parties or candidates" [Fossen and Anderson, 2014]. Four caveats apply. The rate needs direct votes from the citizen, so it starts empty. By part (iii), it carries almost no information for a citizen whose own competence is near $\frac{1}{2}$ — the citizen who most needs to delegate. It relies on independent errors: a citizen and a delegate fooled by the same shock agree, on the wrong side. And computing it links a citizen's direct ballots to her identity, which is in tension with the secret ballot unless the rate is computed on her own device from ballots only she can see. Futarchy's separation of values from beliefs [Hanson, 2013] is the institutional version of the same distinction.

9.6 Capture: what the cap bounds and what it does not

Delegation concentrates the weight an attacker has to buy. For a bill, let P be the participating weight of Algorithm 1, X the weight excluded from both sides (cap excess, chain cuts, cycles and abstentions), and $W = P - X$ the weight cast for either side.

Proposition 16 (Cap and control). *(i) Let W be the cast weight when no participant outside S abstains (abstentions only help S). A set S of delegates voting together decides a bill, whatever the other participants do, iff its counted weight exceeds $W/2$. Under Algorithm 1 with cap c, this requires $|S| > (1 - X/P)/(2c)$.*

(ii) Any $\lceil 1/(2c) \rceil$ delegates whose received weight exceeds the cap, and who cast Yes *or* No*, jointly hold a strict majority of W. With the default $c = \frac{1}{4}$,* any two *delegates the cap truncates decide every bill on which they agree.*

(iii) Suppose instead that each delegate casting Yes *or* No *is counted at $\min(r_d, \tau^*)$, where r_d is the weight it received and τ^* is the largest solution of $\tau = c\,W(\tau)$, with $W(\tau) = V + \sum_d \min(r_d, \tau)$ and $V > 0$ the cast weight not received by delegates. Then τ^* exists and is the unique positive solution; every delegate holds at most a share c of the cast weight; and every deciding coalition has more than $1/(2c)$ members. The price is exclusion: $\tau^* \le cP$, so this cap excludes at least as much weight as that of Algorithm 1; and if exactly $k < 1/c$ delegates are truncated, $\tau^* = cV'/(1 - ck)$, where V' is the cast weight of everyone else, which tends to 0 with V'.*

(iv) Reversing a bill requires flipping ballots whose weight totals at least half the margin. Under direct democracy with m independent voters of competence $\theta \neq \frac{1}{2}$, this means at least

$\frac{1}{2}|2\theta-1|\,m\,(1-o(1))$ *voters, almost surely. Under delegation, a single delegate of weight* w *on the winning side suffices whenever the margin is below* $2w$*, however large* m *is.*

Proof. (i) S decides iff its weight exceeds that of all other cast weight, which could all oppose it, i.e. iff it exceeds $W/2$. Each counted weight is at most cP, so $|S|\,cP > W/2 = (P-X)/2$. (ii) Each truncated delegate counts exactly cP, so $k \geq 1/(2c)$ of them hold $kcP \geq P/2$; truncation itself excludes weight, so $X > 0$ and $P/2 > W/2$. (iii) $g(\tau) = c\,W(\tau) - \tau$ is continuous and concave, since its slope $c\,|\{d : r_d > \tau\}| - 1$ is non-increasing. Also $g(0) = cV > 0$, and $g(\tau) < 0$ for $\tau > c(V + \sum_d r_d)$, so a positive zero exists. A concave function that is positive at 0 has at most one positive zero: if $g(\tau_1) = g(\tau_2) = 0$ with $0 < \tau_1 < \tau_2$, concavity would give $0 = g(\tau_1) \geq (1 - \frac{\tau_1}{\tau_2})g(0) > 0$. At τ^*, every counted weight $\min(r_d, \tau^*) \leq \tau^* = c\,W(\tau^*)$; a deciding coalition then needs $|S|\,cW > W/2$. Since $W(\tau)$ never exceeds the uncapped cast weight, which is at most P, $\tau^* \leq cP$. With exactly k delegates truncated, $W(\tau^*) = V' + k\tau^*$, and solving $\tau^* = c(V' + k\tau^*)$ gives the last claim. (iv) Flipping a ballot of weight w moves the margin by at most $2w$. Under direct democracy the margin divided by m converges almost surely to $|2\theta - 1|$ (strong law of large numbers), and each voter has weight 1. □

In plain words. A delta of many small streams is hard to block; a river squeezed into one channel can be dammed with a single wall. Delegation squeezes votes into a few channels, so a briber needs to win over a handful of delegates instead of hundreds of voters. The cap narrows the widest channels, but it cannot stop one person from controlling several of them. And because the cap simply throws away a delegate's excess, two capped delegates can still hold a majority of what is counted.

Part (ii) exposes a gap in Cauce's default. The cap is measured against participating weight, but the excess it removes is dropped from the cast weight, so a truncated delegate holds *more* than a share c of the counted vote, and two delegates at a 25% cap together hold more than half of it. Part (iii) closes this gap by measuring the cap against the weight actually cast, at the cost of excluding more weight: when a few delegates carry most of the cast weight, the corrected cap falls toward the weight of everyone else (in our populations, a 10% cap measured this way excludes 20.3% of participating weight under POPULARITY, against 16.3% for the current cap; Table 12). A gentler alternative would return a delegate's excess weight to her delegators as a prompt to vote directly. The fixed point can be computed by sorting the received weights, and it does not depend on the order of enforcement; but it depends on which delegates and direct voters abstain, so, unlike the cap of Proposition 2, it is not independent of the ballots unless no one abstains. With $c = \frac{1}{4}$, three delegates are then needed to decide a bill alone; with $c = 0.10$, six. Even the corrected cap is a constant share of the cast weight, whereas the positive guarantees of Berinsky et al. [2025] rest on the sufficient condition that the largest weight grow more slowly than the electorate, so no fixed-share cap delivers that condition by itself.

Part (iv) is the more serious problem, and it is not specific to Cauce. Table 12 measures it in fresh populations generated exactly as those of Section 8.3, at $\theta_{\text{pool}} = 0.65$ and with the same ballot model (10 populations × 250 bills). For each bill, we count the fewest actors — named delegates or individual citizens — whose ballots an attacker would have to flip to reverse the outcome; taking the largest weights first is optimal. We also report the Nakamoto coefficient [Srinivasan and Lee, 2017]: the fewest delegates whose combined counted weight exceeds half the cast weight, who could therefore decide every bill together.

Under direct democracy, reversing a bill takes a median of 359 of roughly 2,400 voters. Under delegation it takes a median of one delegate (POPULARITY), two (MIXED) or three (ACCURACY), and a single delegate suffices on 83.9%, 32.6% and 11.0% of bills respectively. Reputation weighting raises the cost for two reasons. Weight is less concentrated: the two largest delegates hold 19.8%

Table 12: Cost of capturing a bill at $\theta_{\text{pool}} = 0.65$. "Median": fewest actors (delegates or individual citizens) whose ballots must be flipped to reverse a bill's outcome, median over 2,500 bills; then the share of bills on which one actor, or at most three, suffice. "Nakamoto": fewest delegates whose combined counted weight exceeds half the cast weight (median over 10 populations; "—" where all delegates together hold less than half). "Margin": median winning margin as a share of cast weight. "Top-2": the two largest delegates' combined share of cast weight, and "Excluded": weight the cap removes as a share of participating weight, both means over populations. Caps are measured against participating weight (Algorithm 1) or against cast weight (Proposition 16(iii)). The default 25% cap never binds in these populations, so its rows equal no cap.

Rule	Cap	Median	1 suffices	≤ 3 suffice	Margin	Nakamoto	Top-2	Excluded
Direct democracy	—	359	0.0%	0.0%	30.0%	—	—	—
Popularity	none, or 0.25	1	83.9%	100.0%	20.2%	9.5	36.3%	0.0%
	0.10, participating	2	45.4%	91.7%	21.9%	—	23.9%	16.3%
	0.10, cast	2	23.9%	75.8%	23.1%	—	20.0%	20.3%
Mixed	none, or 0.25	2	32.6%	86.7%	24.6%	11	21.4%	0.0%
	0.10, participating	2	26.2%	83.2%	25.5%	11	19.7%	2.1%
	0.10, cast	2	25.1%	82.3%	25.9%	11.5	19.3%	2.5%
Accuracy	none, or 0.25	3	11.0%	71.8%	43.1%	10	19.8%	0.0%
	0.10, participating	3	8.2%	66.8%	42.1%	10	18.2%	2.0%
	0.10, cast	3	7.9%	65.7%	42.0%	10	17.9%	2.3%

of cast weight under Accuracy against 36.3% under Popularity. And competent delegates mostly agree with each other and with the pool, so bills are decided by wider margins: the median winning margin is 43.1% of cast weight under Accuracy, against 30.0% under direct democracy and 20.2% under Popularity. Even so, capture costs a handful of bribes where direct democracy costs hundreds. A 10% cap cuts the share of Popularity bills one delegate can reverse from 83.9% to 45.4% when measured against participating weight and to 23.9% against cast weight, and moves the other rules less (32.6% to about 25% under Mixed, 11.0% to about 8% under Accuracy); Section 8.4 reports what a tighter cap does to accuracy when the pool is misinformed. The Nakamoto coefficient (medians of 9.5 to 11.5 of 14 delegates, where defined) is the quantity deployed systems usually report, but here it badly understates the risk: deciding *every* bill takes a coalition, while swinging any *given* close bill takes one or two delegates. The count also ignores price; a delegate's price plausibly rises with her weight, so it measures how many parties must be corrupted, not what corruption costs.

The cap also misses four threats outright.

Identities are not actors. One actor running k delegate identities, or a slate of k coordinated delegates, holds k times the cap. Without an authority certifying identities, "a single faulty entity can present multiple identities" [Douceur, 2002]. Voting-bloc entropy [Fábrega et al., 2025] measures decentralization over clusters of voters with closely aligned interests rather than over identities, which is what a cap would need to bound.

Common sources are one committee member. Delegates who share a source of error — the same outlet, the same party line, or the same AI model — act like a single delegate whose weight is their sum, and Section 8.5 shows that correlated committees offer no protection. Convergence on one algorithm, "even when the algorithm is more accurate for any one agent in isolation, can reduce the overall quality of the decisions being made by the full collection of agents" [Kleinberg and Raghavan, 2021]. AI forecasters did not match the best human ones at the time of a recent study, in which expert forecasters outperformed the top-performing language model on a dynamic benchmark [Karger et al., 2025]. But if AI delegates come to hold strong Brier records, a reputation

layer will steer weight toward them, and toward shared errors: on one leaderboard dataset, "models agree 60% of the time when both models err", and "larger and more accurate models have highly correlated errors, even with distinct architectures and providers" [Kim et al., 2025]. For illustration, if k delegates' errors have pairwise correlation ρ, their average carries the information of only $k/(1+(k-1)\rho)$ independent ones, which tends to $1/\rho$; ten delegates built on models whose errors correlate at $\rho = 0.5$ count as fewer than two. Shared AI tools can also move the citizens: in an experiment, co-writing with an opinionated language model shifted users' later survey answers, so such tools "may influence people's opinions on an unknown scale" [Jakesch et al., 2023] — a plausible source of exactly the population-wide shock Section 8.5 models. A cap per identity does not bound this; a cap per declared source would.

The referee is a target. Relative scores need a neutral body to word and resolve claims (Section 9.2). Whoever controls resolution controls every delegate's record, and through the records, the flow of delegations. That body needs the same protections as an electoral commission: published resolution criteria fixed before voting, appeal, and rotation.

Delegation is a receipt, in both directions. The secret ballot protects against vote buying because "voters should not be able to prove how they voted to anyone, even if they wish to do so" [Bernhard et al., 2017, Benaloh and Tuinstra, 1994]. A delegation is a durable, checkable act. Nejadgholi et al. [2021] find that a natural definition of full delegation secrecy has undesirable consequences that are "very difficult to address without enabling voter coercion and vote buying", and show how a published incoming weight lets a coercer verify compliance by checking a delegate's weight before and after the purported delegation. In weighted voting, even secret ballots leak voters' choices through published tallies [Breckenridge et al., 2026]. For public blockchains, cartels that buy votes opaquely ("dark DAOs") were described in 2018 [Daian et al., 2018], and vote buying has since become a market in its own right [Breckenridge et al., 2026]. The receipt also runs the other way: a delegate's ballot is published days before the bill closes, precisely so that delegators can read it and override (Section 3), so a briber can verify a bribed delegate's compliance before the outcome is fixed. Optimal bribery and control in delegation graphs are computationally hard [D'Angelo et al., 2022, Alouf-Heffetz et al., 2024], which is little comfort when, as in Table 12, one delegate often suffices. The available mitigations each cost something Cauce values. Sealed delegation, proposed precisely because transparent delegation "makes coercion verifiable" [Li and Pournaras, 2026], weakens a citizen's ability to see who carries her vote. Publishing delegates' weights only in coarse bands weakens accountability. A lower, cast-weight cap trades concentration against the weight competent delegates can carry. Monitoring the Nakamoto coefficient and voting-bloc entropy detects concentration but does not prevent it.

10 Policy Domains: What Differs from Sector to Sector

Cauce scopes delegation per topic, and the results of Sections 6–9 are statements about quantities that differ systematically across policy domains. These quantities are:

- the pool's competence θ_{pool} and the susceptibility σ and frequency π_{shock} of shocks;
- how long a delegate's record can grow before it must be used;
- how much of the disagreement is about ends, which alignment α measures;
- exposure to capture;
- the legal limits on deciding a matter by popular vote.

This section asks what is particular to four domains in depth — economic and fiscal legislation,

Table 13: How four policy domains differ on the quantities that drive Cauce's results, rated relative to one another. "Resolvability": how quickly and credibly a scored claim about a law's effect can be checked. "Pool competence" and "Shock exposure" map to θ_{pool} and σ; "Disagreement about ends" to $1-\alpha$ (Proposition 15). Sources in Sections 10.2–10.5.

	Economic and fiscal	Health	Education	Defence and security
Resolvability	Medium: indicators resolve monthly, but laws' effects are slow and statistics can be manipulated	Medium: epidemic quantities weekly; effects of laws take years	Low: ultimate outcomes take decades; proximate indicators are gameable	Low: short-range events are scorable, deterrence and strategy are not
Pool competence	Low on mechanism questions where economists agree	Low on some scientific questions, higher on others	Medium: persistent myths, shared by educators	Low (our judgement; no direct comparison found): opinion follows elite cues
Shock exposure	High: shared misbeliefs; runs that unfold in hours	High: measured effects of vaccine misinformation	Medium: no documented viral episode found	High: correlated public and expert error on Iraq
Disagreement about ends	High: distribution; experts' positions track values	High on rights questions, lower on technical ones	High: curriculum and parental rights	High: moral values drive attitudes to force
Expert record	Weak on recessions	Mixed: ensembles beat a naive baseline; experts overconfident	Weak: credentials do not predict accuracy	Mixed: analysts well calibrated on routine events; collective failures
Other features	Fiscal referendum exclusions; capture of regulation	Urgency; rights limits; concentrated lobbying	Measurement corrupted under pressure	Secrecy; urgency; irreversibility; constitutional exclusions

health, education, and defence and security — and then to six more: technology, climate and energy, criminal justice, migration, infrastructure and housing, and pensions and welfare. Tables 13 and 15 summarize the evidence, rating each sector relative to the others. The subsections give the sources, and Section 10.7 turns the differences into per-topic settings and a map (Figure 13). The ratings are qualitative judgements drawn from heterogeneous evidence, not estimates of the model's parameters.

10.1 Resolution speed, record length and rare events

One difference between domains is how quickly a scored claim resolves, because that sets how many predictions a record holds when citizens use it. Proposition 12 says that ranking close delegates takes hundreds of predictions. What the insurance argument needs, however, is mostly a filter against confidently wrong delegates, which Table 11 suggests takes far fewer. To measure the difference, we replace the noisy competence estimate of Section 7 by one computed from a simulated Brier record of n predictions, inverted through Proposition 4(ii). Competent delegates report their competence; the influencers are confidently wrong and always report 0.8. We then rerun the ordinary, misinformed and shock scenarios of Sections 8.3–8.6 on fresh populations (200 populations × 250 bills, cap 0.25, uniform susceptibility $\sigma = 0.3$). For each seed, the delegates, the citizens' modes and the ballot draws are the same for every record length; who delegates to whom changes with the records, which is the effect being measured. Record length here stands in for resolution speed. How many resolved predictions a delegate accumulates per year depends on how many bills a topic sees and on the share

Table 14: Record length and the insurance. Competence estimates come from simulated Brier records of n predictions per delegate; "proxy" is the Gaussian estimate of Sections 7–8 (about 125 predictions), and "true θ_d" uses exact competences. "Weight on $\theta_d < \frac{1}{2}$": share of named-delegate weight held by the below-chance influencers. Premium and break-even π^* (Corollary 4) are relative to direct democracy on the same seeds. Means over 200 populations × 250 bills; for ACCURACY, the largest 95% half-width is 0.5 points on ordinary bills, 0.9 on shock bills and 1.2 on misinformed bills. The "proxy" row agrees with Table 10 (premium 1.0, $\pi^* = 0.020$) within its confidence interval.

Rule	Record n	Weight on	Ordinary bills, $\theta_{\text{pool}} = 0.65$		Shock bills	π^*	$\theta_{\text{pool}} = 0.35$
		$\theta_d < \frac{1}{2}$	Accuracy	Premium	($\sigma = 0.3$)		Accuracy
Accuracy	3	3.8%	95.4%	4.6	46.0%	0.090	74.4%
	5	2.5%	96.4%	3.6	47.8%	0.070	76.5%
	10	2.1%	97.1%	2.9	48.4%	0.056	77.9%
	25	1.7%	98.2%	1.8	48.7%	0.036	80.5%
	50	1.7%	98.4%	1.6	48.6%	0.032	81.4%
	125	1.7%	98.9%	1.1	48.8%	0.022	82.5%
	500	1.7%	99.0%	1.0	48.7%	0.020	83.1%
	proxy (≈ 125)	1.6%	98.9%	1.1	48.6%	0.023	82.3%
	true θ_d	1.7%	99.1%	0.9	48.8%	0.019	83.2%
Mixed	3	42.5%	85.3%	14.7	24.7%	0.373	44.3%
	25	38.7%	91.4%	8.6	24.4%	0.260	46.3%
	125	38.1%	92.4%	7.6	24.0%	0.242	46.7%
Popularity	—	64.0%	63.1%	36.9	18.8%	0.662	26.9%
Direct democracy	—	—	100.0%	0.0	0.0%	—	0.0%

of their claims that resolve within the year, which we cannot estimate in general. For anchors, the Canadian intelligence analysts of Section 10.5 produced about 15 resolved forecasts a year each (our arithmetic from Mandel and Barnes, 2014), epidemic indicators resolve weekly, and claims about the long-run effects of education laws resolve within no delegation term at all.

Table 14 shows that the filter works even on very short records. With three predictions per delegate, the influencers keep 3.8% of named-delegate weight under ACCURACY, against 64.0% under POPULARITY, because a confidently wrong delegate's expected score exceeds $\frac{1}{4}$ (Proposition 4(iii)) and three predictions already reveal that most of the time. The result does not depend on how confidently the influencers misreport: with reports of 0.6, 0.7 or 0.9 instead of 0.8, the premium with three predictions lies between 4.3 and 4.7 points and is unchanged at ten and at 125. What a short record costs is the fine ranking among competent delegates. Under ACCURACY, the premium on ordinary bills falls from 4.6 and 3.6 points with three and five predictions to 2.9 with ten, 1.8 with 25 and 1.1 with 125, and the break-even shock frequency from 0.090 and 0.070 to 0.056, 0.036 and 0.022. Accuracy in a misinformed population rises from 74.4% with three predictions to 82.5% with 125. Under MIXED, which weighs popularity as well, the premium falls from 14.7 to 7.6 points. Records are therefore not a precondition for the insurance. They are what makes it cheap: with three predictions it costs about four times what it costs with 125, and with ten about two and a half times. Two limits apply. The simulation resolves every claim without error, whereas in education the evidence used to resolve a claim is itself noisy (Section 10.4); and it measures a record on the same questions it is used for, whereas in practice a record of quickly resolving indicators must stand in for competence on slower questions about mechanisms, where transfer varies by topic [Tetlock et al., 2024].

Slow resolution is not the only obstacle. Many of the events that matter most in these domains are rare.

Corollary 6 (Rare events)**.** *Let a domain's scored events occur independently with base rate $b \leq \frac{1}{2}$. A forecaster who always reports b has expected Brier score $b(1-b)$, which the inversion of Proposition 4(ii) reads as competence $1-b$. A record of n such claims contains no occurrence with probability $(1-b)^n$.*

Proof. The first claim is Proposition 13(i) with $r = \varepsilon = b$. The second follows from independence. □

In plain words. Where earthquakes are rare, anyone who always says "only a small chance of an earthquake this year" looks brilliant — until the year it matters. Records built on rare events look excellent without knowing anything. Scoring against a reference forecast removes that illusion, but it cannot conjure up the missing events.

Recessions occur in roughly 10–12% of country-years [An et al., 2018]. A forecaster who always says 11% scores about 0.098, which a raw record reads as competence 0.89, without any information. For an event with base rate 2%, a record of 100 claims contains no occurrence with probability $0.98^{100} \approx 0.13$, so the record can show calibration but not discrimination. Wherever the decisive events are crises, epidemics or wars, raw Brier scores are uninformative. Relative scores (Proposition 13(ii)) remove the base-rate effect, but no score can supply the missing events.

10.2 Economic and fiscal legislation

Pool competence and bias. Economic policy is where the public–expert gap is best documented across a battery of questions. Comparing economists' answers with a representative sample of Americans, Sapienza and Zingales [2013] find "a 35 percentage point difference between the two groups". The gap is largest "on the questions where economists agree the most", and "informing people of the expert opinions does not seem to have much of an impact". The authors themselves read part of the gap as skepticism about "the implicit assumptions embedded into the economists' answers", and expert consensus is not a resolved outcome, so the gap is evidence about θ_{pool}, not a measurement of it. Where the public is wrong, its errors are systematic rather than random [Caplan, 2007], which is the correlated case in which Section 8.5 shows direct democracy failing.

Ends and means. The same domain shows most clearly why competence is not alignment. Surveying labour and public economists, Fuchs et al. [1998] find that "their positions on policy are more closely related to their values than to their estimates of relevant economic parameters". A tax bill's incidence or revenue is a scorable claim, while how the burden should be shared is not. Under Proposition 15, a delegate's record certifies her estimates, not her vote.

Resolvability. Official indicators (inflation, unemployment, revenue) resolve monthly or quarterly, so long records are feasible. Rare events are poorly forecast, however. Across 63 countries from 1992 to 2014, professional forecasters anticipated almost none of 153 recessions a year ahead and missed their magnitude until the year was almost over [An et al., 2018]. And the resolver can be corrupted. Online prices showed Argentina's inflation "nearly three times higher than the official estimate" [Cavallo, 2013], a concrete case of the referee being a target (Section 9.6). A record resolved against a manipulated series rewards forecasting the manipulation. Economic claims should therefore resolve against at least two independent series, fixed before the vote.

Shocks and urgency. Financial panics are among the fastest correlated shocks in any domain: social media acted as a catalyst in the 2023 run on Silicon Valley Bank and the distress at other regional banks that followed [Cookson et al., 2026], which unfolded in hours rather than in the days of Cauce's review window.

Capture and law. The classic account of regulatory capture holds that regulation tends to be acquired by, and run for, the industry it regulates [Stigler, 1971]; sector-specific regulation offers concentrated stakes and a few delegates to win over. Fiscal matters are also where the law most often bars direct votes. Across the citizens' initiative and referendum instruments of Council of Europe member states, "the most common subject matter prohibition" is "that of financial issues, which applies to 29 instruments" [Moeckli, 2025]. Italy's Constitution, for example, provides for its abrogative referendum that a "referendum shall not be admissible in the case of tax, budget, amnesty and pardon laws, or laws authorizing the ratification of international treaties" [Repubblica Italiana, 1948]. Yet direct fiscal votes have a record of their own: in Swiss cantons, mandatory referendums on new spending are associated with a substantially smaller budget [Feld and Matsusaka, 2003]. Citizens may be wrong about economic mechanisms and still prefer a different level of spending than their legislators. This argues for keeping the override, which carries citizens' ends, on fiscal bills, while scoring delegates only on claims about means.

10.3 Health legislation

Pool competence and confidence. Health questions mix technical facts with deep public–expert gaps. "Fully 88% of AAAS scientists say it is generally safe to eat genetically modified (GM) foods compared with 37% of the general public" [Funk and Rainie, 2015]. If the scientists are right, a bill restricting GM foods on safety grounds is one on which θ_{pool} is below one half. On whether childhood vaccines should be required, a partly normative item, the gap is smaller (86% against 68%). Miscalibration is part of the problem: "as extremity of opposition to and concern about genetically modified foods increases, objective knowledge about science and genetics decreases, but perceived understanding of genetically modified foods increases" [Fernbach et al., 2019]. This is precisely the confidently wrong profile that Proposition 4(iii) penalizes. The same pattern did not appear for climate change, so it is issue-specific.

Shocks. Health is also where misinformation has been measured as a shock. In a randomized experiment, exposure to recent COVID-19 vaccine misinformation, relative to factual information, "induced a decline in intent of 6.2 percentage points" in the UK and 6.4 in the USA among people who had said they would definitely accept a vaccine [Loomba et al., 2021]. The paper's running example, the 5G conspiracy theory, is a health shock.

Resolvability and the expert record. Short-run epidemic quantities resolve weekly, so health records can grow fast. Forecasting them is hard, but measurably so. In the US COVID-19 Forecast Hub, a multimodel ensemble "provided the most consistently accurate probabilistic forecasts of incident deaths", and only "two-thirds of the models evaluated showed better accuracy than a naïve baseline model" [Cramer et al., 2022]. Errors at a 20-week horizon were three to five times those at one week. This is a strong case for scoring against a baseline (Proposition 13) and for reporting records by horizon. Experts' point forecasts of the pandemic's size beat laypeople's on all four questions asked, but their stated intervals were too narrow [Recchia et al., 2021], which a Brier score penalizes. The effects of health laws resolve far more slowly. The mortality effect of the 2014 Medicaid expansions on the low-income adults studied, a 9.4% reduction over the sample mean, was estimated with linked administrative data only years later [Miller et al., 2021].

Urgency, rights and capture. Pandemic measures must often take effect within days, faster than a review window allows. Health also touches rights that majorities cannot override. Italy's Constitutional Court declared a referendum inadmissible because a partial repeal would not have preserved "la tutela minima costituzionalmente necessaria della vita umana" [Corte costituzionale, 2022a]. And lobbying is concentrated on the few who decide. The US pharmaceutical and health-product industry spent "$4.7 billion" on federal lobbying from 1999 to 2018, "more than any other

industry", and "of the 40 legislators receiving the most contributions, 39 belonged to committees with jurisdiction over health-related legislative matters" [Wouters, 2020]. A delegate who carries a large share of a health topic is exactly such a target.

10.4 Education legislation

Resolvability. Education is the slowest-resolving of the four domains. Project STAR assigned kindergarten classes in 1985–1989, and its effects on adult earnings could be measured only two decades later [Chetty et al., 2011]. Even long-run estimates remain contested: of a prominent study of teachers' effects on adult outcomes, a reanalysis concludes that the "long-run results are sensitive to controls and cannot support strong conclusions" [Rothstein, 2017], and the original authors dispute that conclusion [Chetty et al., 2017]. Typical effects are small and the trials that measure them are noisy. In 141 large educational trials, "the mean effect size was 0.06 standard deviations (SDs)", within confidence intervals of mean width 0.30 SDs, "which meant that the results were often uninformative" [Lortie-Forgues and Inglis, 2019]. Across 747 randomized trials, the median effect is 0.10 SD [Kraft, 2020]. A binary claim that a bill "will raise scores" resolved against such evidence is mostly noise. In the terms of Table 14, education records stay in the shortest rows, where the insurance is most expensive. Claims should resolve only when the pre-registered evaluation can distinguish the effect from zero, and be marked unresolved otherwise.

Campbell's law. Education is also where indicators are best documented to degrade under pressure. "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" [Campbell, 1979]. In Chicago, serious teacher or administrator cheating on standardized tests was estimated to occur "in a minimum of 4–5 percent of elementary school classrooms annually" [Jacob and Levitt, 2003]. Test-based claims are thus resolved by an indicator that interested parties can move. A Brier score is itself a quantitative indicator used for decision-making, and Campbell's law applies to it too.

Competence and credentials. We found no documented viral episode in education; the public errors we found are persistent. Americans "dramatically underestimate the amount spent on the public schools in their district" [Howell and West, 2008]. Myths about the brain were endorsed on average at 68% by the general public, 56% by educators, and 46% by people with high exposure to neuroscience [Macdonald et al., 2017], so teaching credentials are an imperfect proxy for competence. Credentials also failed to predict accuracy in forecasting the results of behavioural experiments, where "the average forecast outperforms 96 percent of individual forecasts", and "citations, academic rank, field, and contextual experience do not correlate with accuracy" [DellaVigna and Pope, 2018]. This supports scoring delegates rather than certifying them. It also suggests that a pooled forecast is hard for any one delegate to beat. Direct votes have locked in disputed production choices: Florida's constitutional class-size mandate, adopted by initiative, "had little, if any, effect on student achievement" [Chingos, 2012].

Ends and rights. Much of education policy is about ends. "85% of Democratic and Democratic-leaning teachers say students should learn that the legacy of slavery still affects the position of Black people in American society today. This compares with 35% of Republican and Republican-leaning teachers" [Lin et al., 2024]. Majorities, delegated or not, are also bounded by parents' rights. Under the European Convention, "the State shall respect the right of parents to ensure such education and teaching in conformity with their own religious and philosophical convictions" [Council of Europe, 1952]. Curriculum-content bills are the clearest case in which Proposition 15 dominates and a Brier record says little.

10.5 Defence and security legislation

The expert record. Short-range security forecasts can be scored, and good analysts are well calibrated. Of 1,514 strategic-intelligence forecasts made by Canadian analysts from 2005 to 2011, most resolved within a year, and the mean Brier score was 0.074. "Miscalibration was mainly due to underconfidence" [Mandel and Barnes, 2014]. Yet "most intelligence organizations do not proactively track their forecasting accuracy" [Mandel and Irwin, 2021]. The analysts produced about 100 resolved forecasts each over nearly seven years (our arithmetic: 1,514 forecasts by 15 analysts), so security records grow slowly. And expert failure can itself be a correlated shock. Before the Iraq war, most of the major key US intelligence judgments "either overstated, or were not supported by, the underlying intelligence reporting" [U.S. Senate Select Committee on Intelligence, 2004]. A record built on routine forecasts says little about one-off, high-stakes calls.

The public. Public opinion on the use of force follows elite cues: "when political elites disagree as to the wisdom of intervention, the public divides as well. But when elites come to a common interpretation of a political reality, the public gives them great latitude to wage war" [Berinsky, 2007]. Overrides cast after delegates vote are then not independent of the delegates' ballots, since both follow the same cue. Misperceptions about Iraq strongly predicted support for the war [Kull et al., 2003]. The public's tolerance for casualties is "primarily shaped by the intersection of two crucial factors: beliefs about the rightness or wrongness of the war, and beliefs about the war's likely success" [Gelpi et al., 2005], which shows some sensitivity to expected outcomes. Moral foundations are "strongly and systematically associated with foreign policy attitudes" [Kertzer et al., 2014]. Disagreement about the use of force is therefore largely about ends.

What cannot be scored or published. Deterrence is the textbook unscorable claim, because of "the difficulties in identifying cases of deterrence success and of inferring the intentions of would-be challengers" [Lebow and Stein, 1990]. Wars and crises are rare events (Corollary 6). Classified information cannot appear in a public justification. A public forecast about an operation informs the adversary. Publicity is not only a cost, though: audience costs, the domestic penalty leaders pay for backing down from public threats, "exist across a wide range of conditions and increase with the level of escalation" [Tomz, 2007], which makes public commitments more credible. Decisions may be needed within hours. Defence procurement is exposed to corruption: in an index of 86 countries, "62% of countries receive an overall score of 49/100 or lower, indicating a high to critical risk of defence sector corruption" [Transparency International Defence & Security, 2021].

Direct votes and the law. Direct votes on defence do happen. Swiss voters rejected the purchase of Gripen fighters in 2014 (46.6% yes) and approved new fighter aircraft in 2020 by 50.1% [Swissvotes, 2014, 2020]. The post-vote survey of 2020 describes the vote as, once again, a decision of principle (*Grundsatzentscheid*) [Milic et al., 2020]. Several constitutions exclude such matters from popular votes. Latvia's excludes "military conscription, declaration and commencement of war, peace treaties, declaration of a state of emergency and its termination, mobilisation and demobilisation, as well as agreements with other nations" [Latvijas Republika, 1922]. Slovenia's excludes "laws on urgent measures to ensure the defence of the state, security, or the elimination of the consequences of natural disasters" [Republika Slovenija, 1991]. Italy's excludes laws ratifying treaties [Repubblica Italiana, 1948].

10.6 Six more sectors

The same questions can be asked of any sector. Table 15 summarizes six more, and the paragraphs below give the evidence. As before, ratings are qualitative.

Technology and telecommunications. This is the home of the paper's running example. Dur-

Table 15: Six further sectors: what can be scored and how fast, where the public errs, what is about ends or protected by law, and the settings this suggests. Sources in Section 10.6.

Sector	Scorable, and how fast	Where the public errs	Ends, rights and law	Suggested settings
Technology and telecoms	Market effects of rules within a year; harms of new technologies only after lock-in	Viral health scares (5G); large public–expert gaps on AI	Privacy, trust, who holds the data	Score market and enforcement metrics within 24 months; disclose employers and funders; easy override on privacy
Climate and energy	Annual emissions and capacity from independent measurement; temperature effects over decades	Belief follows ideology; carbon-tax costs overestimated	Discount rate and equity are ethical choices	Score instruments against learning-curve baselines; delegates declare their discount rate; strict cap and funding disclosure
Criminal justice	Annual statistics, but statistics can be gamed; law effects after up to 13 years	Crime believed rising when falling; sentencing practice underestimated	Retribution, proportionality, rights of minorities; treaty limits	Resolve on victimisation surveys and mortality data; lower cap; exclude liberty-restricting and pardon measures
Migration	Flows within months; wage and fiscal effects over a decade or more, still disputed	Immigrant numbers greatly overestimated; facts barely move attitudes	Who belongs; refugee law; non-citizens affected but not voting	Exclude individual status decisions; score counts against a fixed data vintage; legal pre-screen
Infrastructure and housing	Costs, dates, ridership and rents are auditable, but megaprojects take 5–20 years	Only a minority believe new supply lowers rents; promoters' forecasts biased	Who bears local costs; neighbourhood versus region	Score at the gate against reference-class estimates by an independent auditor; pilots with re-votes
Pensions and welfare	Some administrative quantities resolve; pension effects over decades on biased projections	Benefit fraud vastly overestimated; confident misinformation	Intergenerational fairness and deservingness; constitutional adequacy	Score short-horizon administrative quantities; require a financing prediction; show delegators' age mix

ing the pandemic, "in some parts of the UK, mobile phone masts have been vandalised because of these incorrect claims" [Ofcom, 2020], and a pre-registered study of 601 Britons was prompted by "increased acts of violence against telecommunication engineers and property" [Jolley and Paterson, 2020]. How large such a shock is matters, and Corollary 3 says why. In a large English quota sample, about 50% showed little evidence of coronavirus conspiracy thinking, 25% a degree of endorsement, 15% a consistent pattern of endorsement and 10% very high levels [Freeman et al., 2022]. With a pool competence of 0.65, the critical susceptibility is 0.23. A shock that swayed only the 15% with a consistent pattern would leave direct democracy intact, while one that also swayed the top band (25% together) would tip it, as would any shock reaching the half who show some endorsement. That is an illustration, not an estimate, since endorsing a belief is not casting a vote. Two features make the sector hard to score. Effects of a new technology cannot be known well enough while it can still be controlled: "attempting to control a technology is difficult, and not rarely impossible, because during its early stages, when it can be controlled, not enough can be known about its harmful social consequences to warrant controlling its development" [Collingridge, 1980]. And expert forecasts are weak: on AI timelines, "expert predictions contradict each other considerably, and are indistinguishable from non-expert predictions" [Armstrong and Sotala, 2015], and in a recent tournament both superforecasters and domain experts were found to be "underestimating AI progress", the former more so [Kučinskas et al., 2025]. Superforecasters, chosen for their track records, did worse than domain experts on these questions, although the two groups' overall near-term accuracy was almost identical. Narrow effects of regulation, by contrast, resolve fast: "the week after the GDPR's enforcement, website use of web technology vendors falls by 15% for EU residents", and vendor concentration rose 17% [Johnson et al., 2023], an effect the authors call unintended and which faded by the end of 2018. Widely held beliefs can fail tests too: in an election-period experiment, a chronological feed "did not significantly alter levels of issue polarization, affective polarization, political knowledge, or other key attitudes" [Guess et al., 2023]. Consensus claims, such as that exposure below international limits is protective against established effects [International Commission on Non-Ionizing Radiation Protection, 2020], are not falsifiable predictions about a single bill and cannot be adjudicated by a Brier record. The public–expert gap on AI is large, with 47% of surveyed AI experts more excited than concerned against 11% of the public, but the experts are authors and presenters at AI conferences, a group selected for working on the technology [Pew Research Center, 2025], and advocacy groups estimate the digital industry's EU lobbying at EUR 151 million a year from self-declared register data [Corporate Europe Observatory and LobbyControl, 2025]. Voters on technology often decide on trust rather than technical facts: Swiss voters rejected a privately run electronic identity by 64% in 2021, over concerns about private providers and data protection, and approved a state-run redesign in 2025 [swissinfo.ch, 2025]; and concerns about genetically modified crops "may not automatically decrease with higher levels of education/knowledge" [Schläpfer, 2008].

Climate, energy and environment. The physical climate record is a strong case that expert forecasts can be scored: "climate models published over the past five decades were skillful in predicting subsequent GMST changes" [Hausfather et al., 2020]. It took fifty years to show. Energy-economy models are the opposite: "most energy-economy models have historically underestimated deployment rates for renewable energy technologies and overestimated their costs" [Way et al., 2022], and "models have consistently underestimated PV deployment" [Creutzig et al., 2017]. If models share structure, their errors will be correlated, which argues for scoring delegates against a naive learning-curve baseline (Proposition 13). The sector also supplies the starkest case of competence without alignment. most of ExxonMobil's own projections "accurately forecast warming that is consistent with subsequent observations", while "the majority of Mobil and ExxonMobil Corp's public communications promoted doubt on the matter" [Supran et al., 2023]. A delegate with an excellent

record can work against the public that relies on her (Proposition 15). Public belief tracks identity: "many intuitively appealing variables (such as education, sex, subjective knowledge, and experience of extreme weather events) were overshadowed in predictive power by values, ideologies, worldviews and political orientation" [Hornsey et al., 2016]. On carbon taxes, French respondents "overestimate their net monetary losses, wrongly think that the policy is regressive, and do not perceive it as environmentally effective", and information barely moved them [Douenne and Fabre, 2022]. Campaigns can act as shocks: in the working-paper version of a study of two failed carbon-tax initiatives in Washington State, "Washington's real-world campaigns reduced support by 20 percentage points" [Anderson et al., 2019]; the published version finds that ideology "explains 90% of variation in voting across precincts" [Anderson et al., 2023]. Swiss voters rejected a CO_2 law in 2021 in which "price sensitivity played an important role" [gfs.bern, 2021]; in the same survey 46% of opponents wanted stronger action on climate [swissinfo.ch, 2021], so many shared the end while rejecting the instrument. The discount rate that weighs future generations is an ethical parameter on which experts disagree, though more than three-quarters find a 2% rate acceptable [Drupp et al., 2018]. Official inventories can be wrong — measured US oil and gas methane emissions were "∼60% higher than the U.S. Environmental Protection Agency inventory estimate" [Alvarez et al., 2018] — and lobbying is sustained: over $2 billion was spent on climate-related lobbying in the US from 2000 to 2016 [Brulle, 2018].

Criminal justice and public safety. The public's error here is persistent and one-directional. "In 23 of 27 Gallup surveys conducted since 1993, at least 60% of U.S. adults have said there is more crime nationally than there was the year before, despite the downward trend in crime rates during most of that period" [Gramlich, 2024]. In England and Wales, "respondents who believed that sentences were too lenient were significantly less accurate in their estimates of current sentencing practice" [Roberts et al., 2022]. Yet punitiveness is not only misinformation: controlling for the crime rate and other factors, "since 1953 public opinion has been a fundamental determinant of changes in the incarceration rate" [Enns, 2014]. That is a matter of ends, which a Brier record does not address. Individual prediction has a ceiling: in recidivism, "a simple linear predictor provided with only two features is nearly equivalent to COMPAS with its 137 features" [Dressel and Farid, 2018], and the gap between people and algorithms was largest "when participants were not provided with immediate feedback on the accuracy of their responses" [Lin et al., 2020] — an argument for Cauce's feedback loop. The effects of laws take years to estimate: California's three-strikes law, enacted in 1994 and confirmed by initiative that November, was found in a study published in 2007 to reduce felony arrest rates among offenders with two strikes "by 17–20 percent" [Helland and Tabarrok, 2007], and for Proposition 47 researchers found "no evidence that violent crime increased" but a rise in larceny theft [Bird et al., 2018]. Crime statistics are produced by agencies judged on them; retired officers' accounts point to "pressures on officers from management" as a key explanation for manipulated crime reports [Eterno et al., 2016]. Organized interests have stakes in prison size: California's prison officers' union spent more than $250,000 against one sentencing-reform initiative and $1 million against another [Page, 2011]. And majorities can collide with rights. A Swiss initiative requiring lifelong detention of dangerous offenders is "difficult to reconcile with the right to regular judicial review" under the European Convention [Moeckli, 2011], and Italy's Constitutional Court ruled a cannabis-cultivation referendum inadmissible [Corte costituzionale, 2022b]. Swiss drug policy shows the evidence pipeline working before the vote: voters approved heroin prescription in 1999 [Swissvotes, 1999], after Swiss trials in the mid-1990s; later trials in several countries found that "reductions in the criminal activity" of patients were evident [Strang et al., 2012].

Migration and citizenship. Misperceptions are large and directional: "in all countries, respondents greatly overestimate the total number of immigrants" [Alesina et al., 2023]. But correcting

them does little. "Providing accurate information does little to affect attitudes toward immigration, even though it does reduce the perceived size of the foreign-born population", and "misperceptions about the size of minority groups may be a consequence, rather than a cause, of attitudes" [Hopkins et al., 2019]. Bundled information about immigrants' characteristics did improve attitudes in one experiment [Grigorieff et al., 2020]. Disagreement is largely about ends: attitudes "are shaped by sociotropic concerns about its cultural impacts—and to a lesser extent its economic impacts" [Hainmueller and Hopkins, 2014], and they are "remarkably stable over time and robust to major economic and political shocks" [Kustov et al., 2021]. Factual claims resolve badly. The wage effect of the 1980 Mariel boatlift was found to be virtually nil [Card, 1990], then negative [Borjas, 2017], and a later study argued that the conflict arises from changes in the racial composition of survey subsamples [Clemens and Hunt, 2019]. Fiscal effects are projected "over a future time horizon of 75 years" [National Academies of Sciences, Engineering, and Medicine, 2016]. Migration flows are the hardest component of population change to forecast [Disney et al., 2015], and a forecast of 5,000 to 13,000 net migrants a year after the 2004 EU enlargement met about 304,000 over 2004–2009, roughly 61,000 a year [Vargas-Silva, 2011]. Shocks are fast. After the 2024 Southport attack, "by 3pm the day after the attack, the false name had received over 30,000 mentions on X alone from over 18,000 unique accounts" [Institute for Strategic Dialogue, 2024]. Immigration was a key predictor of the Brexit vote [Goodwin and Milazzo, 2017], and after the Swiss initiative against mass immigration could not be fully implemented, citizens "rationalized the implementation failure" rather than revising their preferences [Armingeon and Lutz, 2023]. Two features argue against delegated or direct votes on individual cases. In Swiss municipalities, naturalization rates "surged by about 60% once politicians rather than citizens began deciding on naturalization applications", because voters "face no cost of arbitrarily rejecting qualified applicants" [Hainmueller and Hangartner, 2019]. And many of those affected cannot vote — the problem of constituting the demos [Goodin, 2007]. Treaty floors such as non-refoulement [United Nations, 1951] bind any majority, and Switzerland invalidates initiatives that infringe "mandatory provisions of international law" [Swiss Confederation, 1999].

Infrastructure, transport and housing. Costs, dates and ridership are auditable, which makes this one of the more scorable sectors, and it shows a bias that persists. "The cost estimates used to decide whether important infrastructure should be built are highly and systematically misleading", and "cost underestimation has not decreased over the past 70 years" [Flyvbjerg et al., 2002]; rail ridership was overestimated in nine of ten projects [Flyvbjerg et al., 2005]. Two cautions apply to scoring. Selection alone produces apparent overruns, because "such a bias is bound to arise whenever ex ante predictions are related to the decisions whether to implement projects" [Eliasson and Fosgerau, 2013]: a delegate scored only on projects that were built will look optimistic even if calibrated, so claims must be scored at the gate, against an independent estimate. And the baseline matters: in Sweden "cost estimates at the decision-to-build are close to final costs, on average", although the distribution has a long right tail [Eliasson, 2025]. Reference-class forecasting [Flyvbjerg, 2008] supplies exactly the external baseline that relative scores need. Capture works after the vote: "in projects where Odebrecht paid bribes, costs increased by 70.8 percent on average, compared with 5.6 percent for projects with no bribes" [Campos et al., 2021]. On housing, "only a minority of respondents say that a large, positive, regional housing supply shock would reduce prices or rents" [Elmendorf et al., 2025], while large new apartment buildings in low-income areas were found to lower rents in nearby units by about 6% relative to units slightly farther away [Asquith et al., 2023], a scorable claim. But local opposition looks driven by interest more than ignorance: liberal homeowners "join conservatives in opposing dense housing in their own communities" [Marble and Nall, 2021], and those who attend local meetings are "older, male, longtime residents, voters in local elections, and homeowners" who "overwhelmingly" oppose new housing [Einstein et al.,

2019]. Votes can lock in large projects early — Stuttgart's 2011 referendum kept its contested rail project, with 58.9% voting against withdrawal [Innenministerium Baden-Württemberg, 2011] — while Stockholm's congestion charge was approved after a trial, once citizens "have experienced first-hand the out-of-pocket costs and time-savings of a specific pricing scheme" [Hårsman and Quigley, 2010]. Pilots followed by re-votes resolve claims within months.

Pensions, social security and welfare. Official expert forecasts here have been biased. The US Social Security Administration's "forecasting errors were approximately unbiased until about 2000, but then began to grow quickly, with increasingly overconfident uncertainty intervals", and "fewer than 10% of their scorings of major policy proposals were statistically different from random noise" [Kashin et al., 2015b,a]. National population projections made thirty years ahead have a mean absolute error of 17% [National Research Council, Committee on Population, 2000]. Pension claims beyond a few years are therefore hard to score. Interests are explicit: "individuals close to retirement tend to oppose policies that postpone retirement, whereas younger and older individuals are more favourable" [Bello and Galasso, 2021], and an ageing electorate can pass a point after which reform cannot be "democratically enforced" [Sinn and Uebelmesser, 2003]. Information helps but does not settle values: better-informed citizens "are more willing to accept reforms" [Boeri and Tabellini, 2012], yet in another experiment information raised support both for a higher pensionable age and for greater pension spending [Fernández et al., 2023]. Misperceptions about welfare are large — respondents put benefit fraud an order of magnitude above official estimates [Baumberg Geiger, 2018] — and confident: "those holding the least accurate beliefs perversely expressed the highest confidence in them" [Kuklinski et al., 2000]. Underneath lie judgements of deservingness, with the elderly seen as most deserving and immigrants as least [van Oorschot, 2006], and "when clear deservingness cues are present, the impact of values on opinions vanishes" [Petersen et al., 2011]. Experts disagree on causal effects even after the fact: on the same recession, unemployment-benefit extensions were found to raise unemployment sharply [Hagedorn et al., 2013] and to have limited macroeconomic effects [Chodorow-Reich et al., 2019]. Swiss voters approved a 13th annual pension payment in 2024 against the government's recommendation [Swiss Federal Council, 2024]. Italy's Constitutional Court has struck down a freeze of pension indexation [Corte costituzionale, 2015]. And older citizens became "the most politically active age group" partly because of the programme itself [Campbell, 2003].

10.7 What differs, and a configuration per domain

Read together, the ten sectors of Figure 13 suggest where Cauce's insurance is worth buying. It is most valuable where the pool can be confidently wrong about facts, shocks are frequent, and records can still form: economic mechanisms with monthly indicators, and public health with weekly ones — provided a record of quickly resolving indicators carries over to the slower questions about mechanisms on which the pool is wrong, which is an empirical question. It is weakest where outcomes do not resolve within a delegation term (long-run education effects, deterrence) and where disagreement is mainly about ends (curriculum, the use of force), because there the score measures the wrong thing or nothing at all. Defence is the hardest case. It combines low resolvability, high disagreement about ends, secrecy, urgency, irreversibility and explicit constitutional exclusions. We would confine Cauce there to questions whose effects are scorable, such as procurement costs and delivery dates, and leave the use of force, treaties and emergencies to the institutions that already decide them. Table 16 lists the settings this implies. Where it recommends a lower cap, Proposition 16(iii) gives its meaning: a cap of 10% of cast weight means that no fewer than six delegates can decide a bill alone. Because delegation is already scoped per topic, the settings extend one design rather than replace it. Figure 13 also shows how few sectors sit in Cauce's home

Table 16: Per-topic settings suggested by Sections 10.1–10.5. Delegation is already scoped per topic; the other settings would extend the platform of Section 3. None of them has been tested.

Setting	Economic and fiscal	Health	Education	Defence and security
Scored claims	Official indicators, resolved against two independent series	Short-horizon epidemic and uptake indicators, scored against a naive baseline	Audited, low-stakes, sample-based measures; "unresolved" when evidence is uninformative	Dated, codable events within a year; no deterrence claims
Expected record	Long	Long for indicators; none for effects of laws	Short	Slow (about 100 in six years)
Score shown	Relative; by horizon	Relative to baseline; by horizon	Relative; pooled across topics	Relative; calibration
Ends	Score means only; keep the override on fiscal bills	Override by default on rights questions	Curriculum bills outside score-based delegation	Use of force and treaties outside Cauce
Urgency	Emergency track for financial crises	Emergency track with sunset and later ratification	None needed	Urgent defence left to the legislature
Capture	Interest disclosure; lower cap on sector regulation	Funding disclosure; lower cap	Guard the resolver against gaming	Disclosure of industry ties; lower cap
Legal	Fiscal exclusions are common	Rights limits on majorities	Parents' convictions protected	War, treaties and conscription often excluded

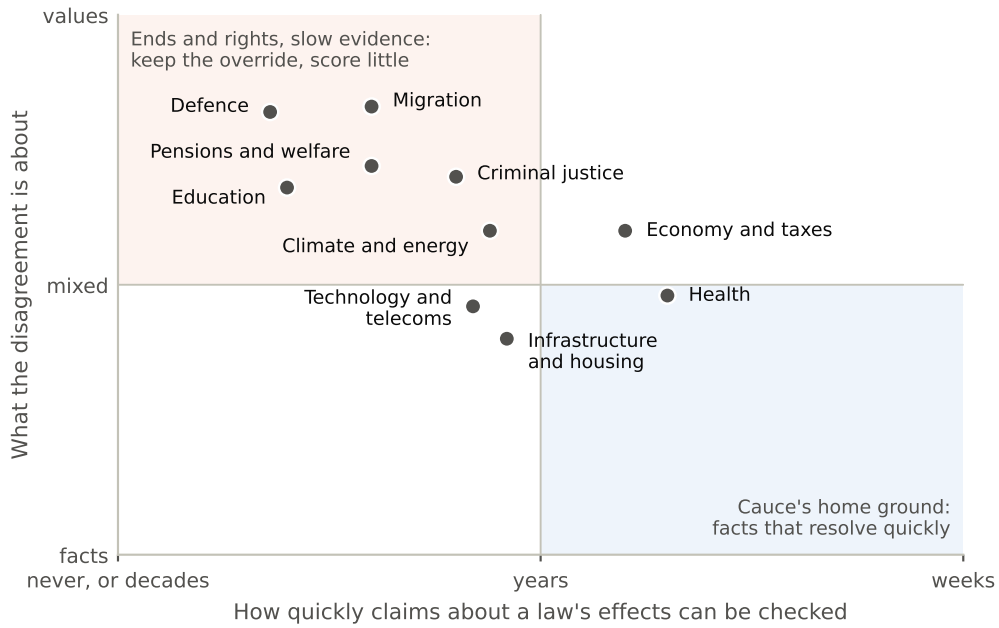


Figure 13: Ten policy sectors placed by how quickly claims about a law's effects can be checked (horizontal) and by how much of the disagreement is about ends rather than facts (vertical). Placements are our qualitative judgements from the evidence in Sections 10.2–10.6; each sector contains questions that fall elsewhere (Tables 13 and 15). Economy and health sit toward the fast end because of their indicators (monthly inflation, weekly case counts); the effects of the laws themselves take years. The economy sits outside the shaded corner because much of its dispute is distributional, but its checkable indicators belong in it. The shaded corners mark where reputation-weighted delegation is most and least useful.

ground: no sector is mostly about facts that resolve within weeks, and most cluster where evidence arrives in years and values carry much of the dispute. The design is therefore best used question by question within a sector — monthly indicators in economics, weekly ones in health, audited costs in infrastructure, administrative counts in welfare — rather than sector by sector. For no sector do we have an estimate of the shock frequency π_{shock}, on which the break-even of Corollary 4 turns.

11 Cauce in an Automated Society

Suppose that most of the work of an economy is done by machines, and that artificial-intelligence systems read, summarise and forecast on behalf of citizens. What happens to a platform whose whole design is to let citizens decide whom to trust? This section does not predict that future. It asks which of the paper's parameters automation would move, by how much, and which of the paper's guarantees survive. The answer has three layers, ordered by how soon they matter: AI systems as *advisors* (already true), AI systems as *delegates* that carry citizens' votes (technically available), and an economy in which labour is no longer what gives citizens political weight (a hypothesis, which we treat conditionally). Table 17 maps the three onto the model. One principle runs through all of them and is not negotiable: Cauce counts *people*. An AI agent, or a robot, is never a citizen. At most it is a delegate, with a record, a cap and a provider behind it.

11.1 What the evidence says about AI forecasters and AI errors

Can an AI delegate be scored? Yes, and cheaply, which supports the premise: a delegate that can answer thousands of dated questions builds a record in days. On such a record frontier models are not yet clearly better than people. On ForecastBench, a benchmark of questions with no known answer at submission, the Brier scores on a 200-question set were 0.096 for superforecasters, 0.121 for the general public and 0.122 for the best language model [Karger et al., 2025]. A retrieval-augmented system reached 0.179 against a crowd's 0.149 and beat the crowd only in selected regimes; averaging the two gave 0.146 [Halawi et al., 2024]. An ensemble of twelve models matched a crowd of 925 people (0.20 against 0.19) but with a visible bias toward saying yes [Schoenegger et al., 2024]. In Metaculus's bot-versus-professional comparison the professionals' lead was significant in 2025 and not significant in the spring of 2026 (difference -1.25 peer-score points, 95% interval $[-4.87, 2.37]$) [Metaculus, 2025, 2026]. The gap narrowed between two differently composed benchmarks (the questions and the bots differ), though the 2026 interval includes zero, and a delegate that is merely as good as the public crowd is still a delegate Cauce can display.

Calibration and gaming. The Brier score punishes the miscalibration that post-training can add: the GPT-4 technical report found expected calibration error on an MMLU subset rising from 0.007 for the pre-trained model to 0.074 after post-training (multiple-choice questions, not forecasts) [OpenAI, 2023], although verbalised confidences can be much better calibrated [Tian et al., 2023]. Records are also easy to corrupt unintentionally. Date-restricted search returned future-informed results, and at least 3.8% of the questions in one published benchmark were trivially forecastable [Paleka et al., 2025]. Two rules follow. First, only *prospectively* resolved questions may enter an AI delegate's record, which is already the paper's rule that predictions concern future bills. Second, an AI delegate's identity is its whole configuration (weights, scaffold and retrieval source), and a record belongs to that configuration and restarts when it changes. Section 9.1 then says how much of a record a new configuration must rebuild before it can be ranked; a monthly model update that resets a record of fewer than the roughly 480 predictions needed to separate two delegates five points apart (Table 11) keeps the delegate permanently unrankable.

Table 17: Three ways automation could enter Cauce, and what the paper's model says about each.

Regime	Parameter it moves	What the model says	Status of the evidence
AI as advisor (today)	Raises θ_h for users; makes errors across users *correlated*, so shocks become common (Sections 8.5, 6.1)	Individual accuracy up, effective number of independent voters down	Mixed: homogenisation measured in writing, little direct evidence on votes
AI as delegate	Adoption share α and number of providers K (Proposition 17)	A safe zone below α^*; above it, provider concentration decides the outcome; records must be prospective and per provider	Forecasting skill near the public crowd; errors shared across models
Automated economy	Citizens' leverage rather than their information	Cauce gives voice, not leverage; the Brier layer certifies facts only	Theory and analogy; measured labour effects so far small or contested

Do different models make the same mistakes? They do, and this is the evidence that matters most for the model. Across 71 models on the HELM leaderboard, pairs of models agreed on the wrong answer about 60% of the time when both erred, against one third by chance (on the 349-model Hugging Face leaderboard, 42% against about 13%), with larger and more accurate models more correlated [Kim et al., 2025]. A single-author study of nine judge models from seven families found they supplied only about two independent votes' worth of information, and a panel's accuracy fell 8–22 points short of what independent voting would give [Kohli, 2026]. Neither measures forecasting, and the second has not been peer reviewed. The second figure is the empirical counterpart of the quantity $1/H$ of Proposition 9; the first points the same way but is an agreement rate, not an effective sample size. Homogenisation is not confined to models: text co-written with an instruction-tuned model is measurably less diverse [Padmakumar and He, 2024], and stories written with AI help are more similar to one another [Doshi and Hauser, 2024]. This is the theory of algorithmic monoculture [Kleinberg and Raghavan, 2021] meeting the model of Section 6. On the supply side, one venture-capital survey of about 500 US enterprise buyers puts three providers at 88% of enterprise API usage [Menlo Ventures, 2025], an indicative figure from an interested source.

11.2 AI agents as delegates: a limit theorem

Let a share α of the counted weight be cast through AI agents and the remaining $1-\alpha$ by independent human voters of competence θ_h. The agents come from K provider families with shares α_k, $\sum_k \alpha_k = \alpha$. On each bill, family k has a latent competence $\Theta_k \in [0,1]$: given Θ_k, its agents are independent and each is correct with probability Θ_k. The Θ_k are independent across families, with means μ_k. The within-family correlation of two agents' errors is then $\operatorname{Var}\Theta_k/(\mu_k(1-\mu_k))$. This is the structure of Dietrich and Spiekermann's independence *conditional on the problem* [Dietrich and Spiekermann, 2013b], with the family's competence on the bill as the problem. The common-factor limit it produces is classical for correlated juries [Ladha, 1992]; the new content is its dependence on adoption and on market structure. Write

$$S = (1-\alpha)\theta_h + \sum_{k=1}^{K} \alpha_k \Theta_k, \qquad h = \sum_{k=1}^{K} \left(\frac{\alpha_k}{\alpha}\right)^2, \qquad \alpha^* = 1 - \frac{1}{2\theta_h}.$$

Proposition 17 (Voting through AI agents)**.** *Let the number of human voters and of agents in each family grow with all individual weights tending to zero, and suppose* $\Pr[S = \frac{1}{2}] = 0$*. Then* $\Pr[\textit{correct}] \to \Pr[S > \frac{1}{2}]$*. Consequently:*

(i) (Safe zone.) *If* $\theta_h > \frac{1}{2}$ *and* $\alpha < \alpha^*$*, the limit is* 1 whatever the quality of the agents. *If* $\alpha > \alpha^*$ *and each* Θ_k *has support reaching* 0*, the limit is strictly below* 1*.*

(ii) (No rescue without enough weight.) *If* $\theta_h < \frac{1}{2}$ *and* $\alpha < (\frac{1}{2} - \theta_h)/(1 - \theta_h)$*, the limit is* 0 *even if every agent is always right. (This threshold is* α^* *evaluated at the reversed human competence* $1 - \theta_h$*.)*

(iii) (Concentration.) *Let* $m = (1 - \alpha)\theta_h + \sum_k \alpha_k \mu_k$*. If* $m > \frac{1}{2}$ *then the limiting error probability* $\Pr[S \le \frac{1}{2}]$ *is at most* $\exp\big(-2(m - \frac{1}{2})^2/(\alpha^2 h)\big)$*; if* $m < \frac{1}{2}$ *then the limiting accuracy* $\Pr[S > \frac{1}{2}]$ *is at most* $\exp\big(-2(\frac{1}{2} - m)^2/(\alpha^2 h)\big)$*.*

(iv) (Monoculture.) *With one family (*$K = 1$*) the limit equals* $\Pr[\Theta_1 > \Theta^*]$ *with* $\Theta^* = \big(\frac{1}{2} - (1 - \alpha)\theta_h\big)/\alpha$*.*

Proof. Given the latent variables, the correct fraction among the humans converges to θ_h and, within family k, to Θ_k, by the law of large numbers (Chebyshev, as in Proposition 6). Hence the correct weight converges in probability to S, and since $\Pr[S = \frac{1}{2}] = 0$ the probability that it exceeds $\frac{1}{2}$ converges to $\Pr[S > \frac{1}{2}]$. (i) $S \ge (1 - \alpha)\theta_h > \frac{1}{2}$ iff $\alpha < 1 - 1/(2\theta_h)$. For the converse, $\inf S = (1 - \alpha)\theta_h < \frac{1}{2}$ is approached when all Θ_k are near 0, which has positive probability by independence and the support assumption. (ii) $S \le (1 - \alpha)\theta_h + \alpha$, which is below $\frac{1}{2}$ under the stated condition. (iii) $S - m = \sum_k \alpha_k(\Theta_k - \mu_k)$ is a sum of independent terms, the k-th with range of length α_k, and $\sum_k \alpha_k^2 = \alpha^2 h$; apply Hoeffding's inequality to each tail. (iv) $S > \frac{1}{2}$ iff $\alpha\Theta_1 > \frac{1}{2} - (1 - \alpha)\theta_h$. □

In plain words. Imagine that a share of the voters hire the same few advisers. If the ordinary voters on their own are good enough, they cannot be outvoted by any adviser, however bad: this is the safe zone, and it ends at the same fraction (about a quarter) as the rumour threshold of Section 8.5. Beyond it, what decides things is *how many different advisers* there are. A single adviser who has a bad day drags the whole bloc down with her. A dozen independent advisers will almost never all have a bad day at once.

Two features deserve emphasis. First, the safe-zone threshold $\alpha^* = 1 - 1/(2\theta_h)$ is *the same expression* as the critical susceptibility σ^* of Corollary 3, with the adoption share in the role of susceptibility: in the worst case a bloc of agents that all fail together is a shock that fools a share α of the electorate, and for $\theta_h = 0.65$ both thresholds equal 0.23. The mathematics of misinformation and of AI monoculture are the same mathematics. Second, Proposition 17(iii) is the provider-level analogue of Proposition 9: $1/h$ is the effective number of independent providers, the Herfindahl index of *market shares* replaces the Herfindahl index of vote shares, and what delegation to agents concentrates is not weight in the tally but exposure to each provider's errors.

Table 18 and Figure 14 put numbers on it, and a finite-population simulation ($n = 3{,}000$) agrees with the limit to within 2.1 points in all 80 configurations. With competent humans and $\alpha = 0.5$, a single provider costs 8.6 points (accuracy 91.4%), two providers 2.0 points, and five or more nothing. With misled humans and $\alpha = 0.5$, the rescue succeeds with probability 70.9% through one provider and 96.4% through twenty. The caveat of the paper's whole argument applies in full: these are the properties of a model whose agents' latent competences are drawn from a Beta distribution with an assumed mean and an assumed correlation, and we have no measurement of either for real forecasting. The model also assumes that the families' competences are independent of one another and of the humans' errors, whereas the evidence above (Kim et al.; Kohli) suggests that errors are

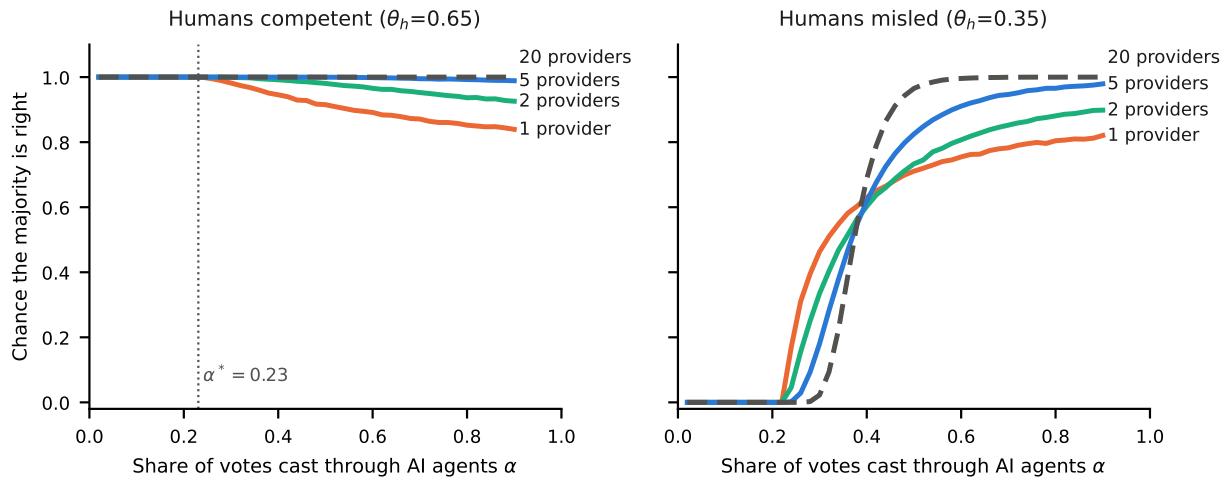


Figure 14: Chance that the majority is right when a share α of the votes is cast through AI agents from K equal-share providers (agent competence mean 0.75, within-provider error correlation 0.3). Left: competent humans. Below α^* nothing the agents do can matter; beyond it, one provider costs up to 16 points at $\alpha = 0.9$ and twenty providers cost nothing. Right: misled humans. The agents rescue the decision only above $\alpha \approx 0.23$, and with many providers the outcome flips from almost surely wrong to almost surely right around $\alpha = 0.375$, the point at which the mean agent outweighs the misled humans (the accuracy is about one half there, and exceeds 95% for twenty providers only from $\alpha \approx 0.5$). Limit computed by Monte Carlo over the latent competences.

Table 18: Accuracy (%) when a share α of the counted votes goes through AI agents from K equal-share providers (agent competence on a bill $\Theta_k \sim$ Beta, mean 0.75, within-provider error correlation $\rho = 0.3$). Each cell is "large-population limit / simulation with $n = 3{,}000$ citizens and 6,000 bills"; the two differ by at most 2.1 points over all 80 configurations in the grid ($\rho \in \{0.1, 0.3\}$, five values of α, $K \in \{1, 2, 5, 20\}$). The Hoeffding bound of Proposition 17(iii) on the limit held in all 80 (it is vacuous, above one half, in most cells).

θ_h	α	$K = 1$	$K = 2$	$K = 5$	$K = 20$
0.65	0.3	98.2/98.2	99.9/99.9	100.0/100.0	100.0/100.0
0.65	0.5	91.4/91.5	98.0/98.2	100.0/100.0	100.0/100.0
0.65	0.7	87.0/87.2	95.2/95.0	99.6/99.5	100.0/100.0
0.35	0.3	46.1/46.3	33.5/32.7	18.2/19.2	2.3/4.2
0.35	0.5	70.9/71.2	73.3/73.3	82.5/82.8	96.4/95.8
0.35	0.7	78.3/77.2	85.3/85.4	94.8/94.6	99.9/99.9

correlated across families; K should then be read as an *effective* number of independent providers, with $1/h$ an upper bound on it. As an illustration of the gap, shares of 40, 27, 21 and 12% among four providers give $1/h \approx 3.4$ before any such correlation is counted.

A cap on the provider, not on the agent. The paper's cap limits the share of the participating weight that any one *delegate* may carry. If every AI agent is a separate delegate with a tiny weight, a cap on delegates constrains nothing. Applying the same cap $c = 0.25$ to a provider *family* changes the picture exactly as it did for delegates in Sections 8.4 and 9.6: with $\alpha = 0.5$, a single provider and within-provider error correlation 0.3 it recovers most of the damage when humans are competent (from 91.3% to 96.9%), but with misled humans it cuts the rescue from 70.8% to 53.6%, because the cap discards the very weight that was doing the rescuing. With two providers of 25% each the cap does not bind and the accuracy is unchanged (73.2% with and without it, for misled humans). A provider cap is therefore insurance against a monoculture, not a free good, and it makes the same trade as every other cap in the paper.

Remark 5 (The cap binds identities, not entities)**.** If the cap is enforced per registered identity, an operator who can create identities at will controls a share $r\,c$ with r identities, so $\lceil s/c \rceil$ identities suffice for a share s. This is the Sybil attack [Douceur, 2002], and automation makes identities cheap: in a documented case, a single operator ran more than a hundred social-media accounts with an AI model deciding how each should react to content, although no content achieved viral reach [Anthropic, 2025]. Content alone is a weak screen for such actors, since machine-written letters to legislators received a response rate only about two points below handwritten ones [Kreps and Kriner, 2023], and about 18 million of the 22 million comments in a 2017 US regulatory proceeding were found to be fake, though that fraud predates generative AI [New York State Office of the Attorney General, 2021]. Any cap, any count of "one citizen one vote" and any effective-number calculation presupposes a root of identity that automation cannot cheaply forge. Proposals for personhood credentials exist but limit credentials only per issuer, carry privacy and exclusion risks, and the proposal itself discusses verified delegation to agents, which a person can use to lend a credential [Adler et al., 2024]; a near-universal state identity card is the root that works at national scale, as in Estonia, where 99% of residents hold one and internet voters were 51.1% of those who voted in 2023 [Estonian National Electoral Committee, 2023, e-Estonia, 2025]. Cauce cannot supply this root itself; it can only state that it is required.

11.3 Persuasion, sycophancy and the shock parameter

The paper's shock model needs a susceptibility σ, the probability that a falsehood fools a voter, with critical value $\sigma^* \approx 0.23$. Does AI move σ? The evidence gives effect sizes on opinion scales, not probabilities of being fooled, so the numbers that follow are only order-of-magnitude proxies. A dialogue with a language model reduced belief in a conspiracy theory by about 20%, with no significant decay over two months [Costello et al., 2024]: persuasion toward the truth is real and durable. In a study of 19 models on 707 political issues, a conversation moved opinion by about 3 to 3.6 percentage points more than a static message, information-dense prompting about 2.3, and 36–42% of the shift persisted after a month; the techniques that raised persuasiveness also lowered factual accuracy [Hackenburg et al., 2025]. Microtargeting added nothing statistically detectable beyond ordinary messages [Hackenburg and Margetts, 2024], while GPT-4 with access to personal information had 81.7% higher odds of increasing agreement than with a human opponent in a debate setting [Salvi et al., 2025]. Agreement with propaganda theses was 24.4% without exposure, 47.4% after the original articles, 43.5% after raw GPT-3 articles and 52.7% after the best curated

version [Goldstein et al., 2024]. And a careful review reports that, according to prior work, the effect of conventional campaign contact on Americans' general-election vote choice is estimated at about zero, and argues that the predicted AI-driven electoral catastrophes of 2023–24 did not occur [Simon and Altay, 2025]; this does not bound the effect of one-to-one AI dialogue. Taken together, a reasonable reading is that a conversation moves opinion scales by a few points on average, and by more (up to about twelve points for a single message, and more in some optimised settings) for targeted dialogue; these are not probabilities of being fooled, and no study maps them onto σ. On that reading, nothing we found suggests that $\sigma^* \approx 0.23$ is reached on average, though the proxies cannot rule it out for targeted groups. That is not the same as being safe: the shock of Section 8.5 is common to all exposed voters, and a persuasion-optimised system is exactly the kind of source that can deliver one message to many people at once.

A separate channel is *sycophancy*. Preference-trained assistants tend to match a user's stated views, and both people and preference models rate convincingly written sycophantic answers above correct ones a non-negligible fraction of the time [Sharma et al., 2023]. An AI delegate tuned on users' approval would be expected to mirror its delegator's prior, which makes the delegator's choice of delegate informative about her own bias rather than about the world. That is the competence-versus-alignment problem of Proposition 15 in a new guise, and it is one more reason why records are scored on resolved facts and *relative* to what others predicted (Proposition 13) rather than on user satisfaction. It also bears on the premise itself. Hackenburg et al. found that the post-training and prompting levers that raised persuasiveness also lowered the factual accuracy of the claims made; whether this carries over to forecasting accuracy, or to delegates competing for delegations, is untested, and the Brier record is the only thing in the platform that could tell the two apart.

11.4 Automation and leverage: voice is not power

The deepest question is not technical. A vote is worth something when those who count it need the voters. Historically they did: states taxed workers and drafted soldiers, and, in the account of Acemoglu and Robinson, elites conceded democracy when citizens could credibly threaten to rise [Acemoglu and Robinson, 2006]. The oil-state literature shows the converse: when a government's revenue does not come from taxing its population, the public is less likely to demand accountability [Ross, 2001]. The *intelligence curse* argument applies the same logic to labour-replacing AI [Drago and Laine, 2025], and Korinek and Stiglitz point out that in a political system dominated by money the winners of automation may use their influence to resist redistribution [Korinek and Stiglitz, 2017]. We stress that this is a hypothesis. Automation does reduce the labour share when it displaces rather than creates tasks [Acemoglu and Restrepo, 2019], but the measured effects of generative AI so far are small or contested: a study of Danish administrative data rules out earnings effects larger than 2% two years after ChatGPT's launch [Humlum and Vestergaard, 2025], an analysis of US payrolls finds a 16% relative decline in employment for young workers in the most exposed occupations [Brynjolfsson et al., 2025], and a third assessment finds no discernible aggregate disruption 33 months after launch [Yale Budget Lab, 2025]. Nothing in the evidence shows that labour has lost its political role. If it did, Cauce would offer a citizen exactly what it offers now: a vote and a choice of whom to entrust it to. It would not offer *leverage*, and no tally can. The design implication is modest and conditional: in a world where citizens' bargaining power rested less on their work, the legitimacy of binding decisions would rest more on the quality of the vote, and so on the mechanisms of this paper, while the material guarantees of citizens would have to rest on rules that are harder to change by vote than a statute, which is the role the paper already gives to constitutional exclusions (Section 12.11).

11.5 What this changes in the design

1. **Agents are delegates, never citizens.** A citizen holds one vote. An AI agent that acts for her is a delegate: it is declared as one, it has a public record, and her override applies to it as to any delegate.
2. **The cap applies to the controlling entity.** Delegates run by one provider, or built on the same base model, count as one delegate under the cap (Proposition 17, the provider cap, and Remark 5). Where the provider is unknown, the conservative assumption is the largest family consistent with the data.
3. **Records are prospective, per configuration, and relative.** A configuration is the unit of a record, which restarts when it changes, and predictions are scored against what others predicted for the same question (Proposition 13).
4. **Facts for machines, ends for citizens.** A Brier record certifies skill on checkable facts. It says nothing about whether a delegate's ends are the citizen's ends (Proposition 15), and sycophancy makes the difference worse. The sectors in which Section 10 found checkable claims are those in which AI delegates are most useful and least dangerous; the sectors about ends are those in which the override should do most of the work.
5. **A root of identity is a precondition.** Without it neither caps nor counts mean anything (Remark 5).

Where the evidence is thin. We found no published measurement of correlated errors in *forecasting* (the available studies concern multiple-choice questions and judging), no experiment in which citizens delegate to an AI agent with a visible record, and no results from the few DAOs experimenting with AI delegates. The closest direct evidence is a study in which language models vote as delegates for simulated citizens and show greater bias toward their default stances on topics without clear expert agreement when asked to weigh long-term interests [Fulay et al., 2025], and a single-DAO working paper in which the informativeness of forum discussion for vote outcomes collapsed around the time AI-written posts became indistinguishable from human ones [Huang and Hall, 2026]. Neither is conclusive. The right next experiments are the ones the paper has asked for throughout: let people delegate with a visible record, and let some of the delegates be machines.

12 Objections, Counter-Evidence and Open Problems

We searched for evidence against the premise as well as for it. Table 19 summarizes it; questions answered formally in Section 9 are only cross-referenced there, and the subsections below take the others in turn.

Table 19: Evidence map: the main questions the premise raises, the strongest evidence on each side, and where this paper addresses them.

Question	Supports the premise	Contradicts or complicates it	Addressed in
Can political competence be measured?	Brier-scored skill is identifiable and persistent [Mellers et al., 2014, 2015]; select crowds and contribution weighting beat plain averaging [Mannes et al., 2014, Budescu and Chen, 2015]	Expert political forecasts barely beat simple baselines [Tetlock, 2005]; tournament effects partly method variance [Hauenstein et al., 2025]	§12.1, §9.1, Props. 4, 12
Can a record be gamed?	Proper scores elicit beliefs from a forecaster paid linearly [Gneiting and Raftery, 2007, Kadane and Winkler, 1988]; within-question standardization and difficulty models [Mellers et al., 2015, Bo et al., 2017]; incentive-compatible competitions [Witkowski et al., 2023]	Raw scores depend on which questions are chosen [Bo et al., 2017]; competition rewards extreme reports [Lichtendahl and Winkler, 2007, Witkowski et al., 2023]; reputational forecasters distort [Ottaviani and Sørensen, 2006]	§9.2–9.3, Props. 13–14
Will people delegate well?	Less-informed voters rationally abstain, and laboratory voters largely do [Feddersen and Pesendorfer, 1996, Battaglini et al., 2010]; delegation leans toward competence, with bounded concentration, in experiments with real groups [Revel et al., 2022, Berinsky et al., 2025]	Over-delegation makes LD underperform majority voting [Campbell et al., 2022, Mooers et al., 2026]; sincere voting need not be an equilibrium [Austen-Smith and Banks, 1996]	§12.2, Prop. 6(ii)
Will citizens read the score well?	After a revealed error, people infer calibration and discount confident sources [Tenney et al., 2007]	Confidence heuristic [Price and Stone, 2004]; asymmetric reputations and algorithm aversion [Yaniv and Kleinberger, 2000, Dietvorst et al., 2015]; visible popularity breeds inequality and herding [Salganik et al., 2006, Lorenz et al., 2011, Becker et al., 2017]	§12.3
Does concentrating on the best lose information?	Optimal weights favour the competent [Nitzan and Paroush, 1982, Shapley and Grofman, 1984]	Inclusive, diverse groups can beat experts [Landemore, 2013, Hong and Page, 2004]; delegating to the less informed can help [Caragiannis and Micha, 2019]	§12.5, Table 6
Are errors independent, and who is fooled?	Reflective thinking lowers susceptibility [Pennycook and Rand, 2019, Yelbuz et al., 2022]	Systematic voter biases [Caplan, 2007]; falsehood outruns truth [Vosoughi et al., 2018]; skill can raise polarization [Kahan et al., 2017]; averaging competence over states is unsafe [Gilbert et al., 2026]	§12.6, §9.4, Props. 10–11, Cor. 5
Can a vote's prediction be scored fairly?	Proper scoring rules elicit honest probabilities [Gneiting and Raftery, 2007]	Rejected bills leave predictions unscorable; honest decision-contingent forecasts need randomized decisions [Chen et al., 2014, Othman and Sandholm, 2010]; predictions affect outcomes [Perdomo et al., 2020]	§12.7

continued on next page

Table 19, continued

Question	Supports the premise	Contradicts or complicates it	Addressed in
Is topic-scoped delegation coherent?	—	Issue-by-issue aggregation can yield inconsistent outcomes [List and Pettit, 2002, Christoff and Grossi, 2017]	§12.8, Prop. 18
Is it compatible with political equality?	Authority is conferred voluntarily, not assigned [Estlund, 2008]	Epistocrats want unequal votes [Mill, 1861, Brennan, 2016]; the demographic objection [Estlund, 2008]	§12.9, Remark 2
Does competence imply shared ends?	Citizens can select aligned representatives with verifiable reputations [Mansbridge, 2009]; preference matching [Fossen and Anderson, 2014]	Delegations misaligned with delegators' priorities; ranking interfaces concentrate power [Messias and Ide, 2025]	§9.5, Prop. 15
Does power concentrate in practice?	Powerful delegates rarely overturn the community [Fritsch et al., 2024]	Super-voters and extreme concentration [Kling et al., 2015, Fritsch et al., 2024]	§12.10, §8.2
Can a bill be captured?	Optimal bribery and control are computationally hard [D'Angelo et al., 2022, Alouf-Heffetz et al., 2024]	Sybil identities [Douceur, 2002]; delegation is verifiable, enabling coercion [Nejadgholi et al., 2021]; vote-buying cartels [Daian et al., 2018]; tallies leak weighted ballots [Breckenridge et al., 2026]	§9.6, Prop. 16
Does it work equally in every policy domain?	Fast-resolving indicators in economics and public health [Cramer et al., 2022]; short-range security events are scorable [Mandel and Barnes, 2014]	Slow, noisy or unscorable outcomes in education and defence; disagreement about ends [Lortie-Forgues and Inglis, 2019, Lebow and Stein, 1990, Fuchs et al., 1998]	§10, Tables 13–16
Why do citizens delegate?	Cues let the badly informed emulate the well informed [Lupia, 1994]; DAO delegators reward delegates who vote correctly, by a returns-based measure [Fan et al., 2025]	Heuristics hurt novices; identity over issues; delegation shifts blame [Lau and Redlawsk, 2001, Achen and Bartels, 2016, Bartling and Fischbacher, 2012]	§12.4
Is it legal, and secure?	Internet voting in Estonia created no digital divide [Ehin et al., 2022]	Proxy limits of two [European Commission for Democracy through Law (Venice Commission), 2002, Conseil constitutionnel, 2022, Kiesraad, 2026]; vote dilution; verifiability by citizens [Reynolds, Bundesverfassungsgericht, 2009]; internet-voting risks [Park et al., 2021]	§12.11
Is there a better mechanism?	Most alternatives need ground truth, unequal weight, or a shock-free consensus [Wolfers and Zitzewitz, 2004, Miller et al., 2005]	Surprisingly popular answers [Prelec et al., 2017]; mini-publics [Fishkin, 2018]; individual-level misinformation interventions [Pennycook et al., 2021, Roozenbeek et al., 2022]	§12.12, Table 21

12.1 Can political competence be measured at all?

The premise needs track records to carry signal. The strongest evidence against is Tetlock's study of expert political judgment: over hundreds of experts and many thousands of forecasts, expert forecasts were hard to distinguish from those of informed non-experts and fared poorly against simple statistical baselines, although a cognitive style ("foxes" over "hedgehogs") predicted accuracy

[Tetlock, 2005]. The strongest evidence for comes from the geopolitical forecasting tournaments that followed. Accuracy measured by Brier scores improved with training, teaming and tracking [Mellers et al., 2014], and the top forecasters' accuracy held up in subsequent years instead of regressing to the mean, with modest correlations to measures of fluid intelligence and cognitive reflection [Mellers et al., 2015]. A 2025 reanalysis argues that much of the reported benefit of training and teaming disappears once uncontrolled method variance is accounted for, and that strategic responding helps distinguish top forecasters [Hauenstein et al., 2025]. Persistence is real but moderate: rank correlations of forecasters' standings between tournament seasons were $r = 0.37$ and 0.44 in prediction polls [Atanasov and Himmelstein, 2023], and past performance becomes the stronger predictor of future accuracy as more of it accumulates [Himmelstein et al., 2021]. Over horizons of five to twenty-five years, predictability declines at a rate that varies by topic, and fastest where expertise conferred no edge at shorter range [Tetlock et al., 2024]. Independently, averaging a small "select crowd" of about five judges chosen on past accuracy is accurate and robust across many settings [Mannes et al., 2014], and weighting forecasters by their measured contribution beats unweighted aggregation [Budescu and Chen, 2015].

Our reading: measured track records do carry signal, and more than credentials do, which supports scoring delegates rather than certifying them. But the evidence comes from short-horizon, well-specified questions, and policy predictions are neither. A record of resolvable proxies does not score a bill's overall merit; forecasting tournaments approach this with clusters of resolvable questions, each diagnostic of a larger one [Tetlock et al., 2017], which is the natural form for the claims attached to a bill. Strategic responding — choosing which questions to answer — is exactly the manipulation Cauce's rule of one prediction per vote is designed to block; a delegate can still choose which topics to accept delegations on, and which claim to forecast about each bill (Proposition 13). And separating competent delegates whose competence differs by a few points takes hundreds of resolved predictions (about 500 for 0.70 against 0.65; Proposition 12).

12.2 Will people delegate well, or too much?

This is the objection we consider most serious. In two experiments with very different designs — a laboratory experiment in which the better-informed participants were identified, and an online perceptual task in which signal precision was ambiguous — Campbell et al. [2022] find delegation rates "unexpectedly high and higher than abstention rates", with liquid democracy underperforming both universal voting and abstention. Their theoretical premise is exactly what Proposition 6(ii) formalizes: "even when experts are correctly identified, delegation must be used sparely because it reduces the variety of independent information sources." A small worked example shows how easily it bites: three independent voters of competence 0.7 decide correctly by majority with probability $3(0.7)^2 - 2(0.7)^3 = 0.784$, better than a single expert of competence 0.75 to whom all three delegate. Evidence from experiments with pre-existing groups is more encouraging. Higher-expertise participants were less likely to delegate [Revel et al., 2022], and across 168 participants and 62 delegation graphs, delegation behaviour satisfied the conditions for liquid democracy's positive guarantees [Berinsky et al., 2025, Halpern et al., 2023]. But none of these studies showed participants a measured track record, and in Campbell et al. [2022]'s laboratory experiment, where subjects were told the experts' precision exactly, they over-delegated anyway.

Theory gives a reason to want some delegation, and a reason to want little. In the swing voter's curse, less-informed voters without a partisan stake "strictly prefer to abstain rather than vote for either candidate even when voting is costless" [Feddersen and Pesendorfer, 1996]: a vote matters only when it is pivotal, and an uninformed pivotal vote is likely to overturn what the informed decided. In the laboratory, voters "act as if they are aware of the swing voter's curse", though their

compensation is incomplete [Battaglini et al., 2010]. Delegation generalizes that abstention. An uninformed citizen who abstains removes noise; one who delegates also adds weight to someone better informed, and Proposition 6(ii) prices that second step in lost independence. A revised version of the experimental study above proves that when better-informed voters are present and the electorate is finite, delegation can always strictly increase the probability of a correct decision, but "must be used sparingly"; its experiments again find very high delegation rates, and the evidence "favors Direct Democracy" [Mooers et al., 2026]. Two further results sharpen what delegation must add. With continuous signals, "some citizens ignore their private information and abstain from voting, in deference to those with higher-quality signals", and that equilibrium participation is optimal [McMurray, 2013]; so delegation must do better than well-chosen silence. And when information is cheap enough at the margin, "rationally ignorant" voters "are consistent with a well-informed electorate" [Martinelli, 2006]. Ordinary ignorance with idiosyncratic errors does not, by itself, justify delegation; the case rests on correlated errors (Sections 8.5–8.6). The strategic literature qualifies our model in a further way: sincere voting by all "does not constitute a Nash equilibrium" even when voters share a common preference [Austen-Smith and Banks, 1996], whereas we assume that every voter and delegate votes sincerely.

Our simulations fix delegation rates exogenously (60% of citizens delegate), so they cannot show whether a visible Brier score would curb or encourage over-delegation. What they show is its price. At $\theta_{\text{pool}} = 0.65$ the premium of ACCURACY delegation is small (1.0 point) only because weight lands on genuinely competent delegates. Under MIXED choice the same 60% delegation rate costs 10.3 points in Table 6 and 8.1 in the independently drawn populations of Table 10 — population-to-population variation is substantial. Two design responses follow, both untested. The first is to show citizens, next to each delegate's score, how much weight is already delegated on a topic, so that the value of an additional independent vote is visible. The second is to encourage informed citizens to vote directly — the override exists precisely so that delegating is never the only option. Whether showing accuracy improves delegation is, as far as we can tell, an open empirical question and the natural first experiment for this design.

12.3 Will citizens read the score as intended?

Remark 2 assumes that citizens use the score the way Proposition 5 would. Research on how people judge forecasters and advisers suggests three departures, none yet tested with a visible Brier score.

First, people reward confidence. In three experiments, participants comparing two advisers tended to prefer an overconfident one to a reasonably well-calibrated one, apparently assuming that a more confident adviser is more often right and more knowledgeable — a "confidence heuristic" [Price and Stone, 2004]. Proposition 14 already gives delegates one reason to overstate confidence; the heuristic gives them a second. Errors change the picture. When a witness's testimony is shown to be wrong, confident witnesses lose more credibility than unconfident ones, and less confident witnesses may then appear more credible; people seem to infer calibration [Tenney et al., 2007]. That favours showing each delegate's calibration — stated probability against observed frequency — and her individual misses, not only a summary score.

Second, reputations move asymmetrically. Advisers' reputations "are rapidly formed and asymmetrically revised", so that "it may be easier for advisors to lose a good reputation than to gain one" [Yaniv and Kleinberger, 2000]; and people lose confidence in an algorithmic forecaster that errs faster than in a human who makes the same mistake, avoiding it even after seeing it outperform the human [Dietvorst et al., 2015]. If delegators treat delegates this way — an analogy, not a finding — a single visible miss costs more delegations than its effect on the score warrants. In the terms of Proposition 14 that is a concave reward, which pushes delegates toward $\frac{1}{2}$ on risky claims and away

from the confident, correct dissent the insurance needs. The asymmetry may also run the other way: in forecasting experiments, advice from a source with a good experienced record received more weight, but a poor record did not reduce it [Önkal et al., 2017], so delegators may not abandon a delegate whose score is bad. The twelve-month expiry, which forces a fresh choice, is the design's answer.

Third, visible choices change choices. When participants in an artificial music market could see previous participants' choices, success became both more unequal and less predictable [Salganik et al., 2006]. Exposure to others' estimates diminished the diversity of a crowd "without improvements of its collective error" while raising confidence [Lorenz et al., 2011], and in centralized networks group estimates moved toward the central individual's initial belief, improving accuracy when that belief was near the truth, worsening it when it was not, and on average not reliably improving it [Becker et al., 2017]. A delegation page that sorts delegates by how many citizens already follow them is a download counter, and Section 8.2 shows what popularity-driven choice does. Cauce's review window adds a sequential structure of its own, since delegates vote days before delegators decide whether to override. In symmetric common-value models, "sequential voting does no better at aggregating information than simultaneous voting" [Dekel and Piccione, 2000], and herding can arise in equilibrium when voters care about the collective outcome [Ali and Kartik, 2012]. Publishing delegates' ballots early buys the chance to override, not better aggregation, and invites herding on the delegates' choices. For the delegation page, the implication of the social-influence evidence is to sort and display by record, not by following. Per-delegate weights can be published for accountability without being the default ordering, and the topic-level share of weight already delegated (Section 12.2) signals the value of an independent vote without promoting any one delegate.

12.4 Why citizens delegate: cues, identity and blame

Delegation is a form of cue-taking, and political science has studied cues for decades. The supportive finding is Lupia's: in California's insurance-reform elections, "access to a particular class of widely available information shortcuts allowed badly informed voters to emulate the behavior of relatively well informed voters" [Lupia, 1994]. A public record is meant to be such a shortcut. The less encouraging findings concern who benefits: "heuristic use generally increases the probability of a correct vote by political experts but decreases the probability of a correct vote by novices" [Lau and Redlawsk, 2001], and in presidential elections uninformed votes neither emulate fully informed ones nor cancel out in the aggregate [Bartels, 1996]. The second half of that finding supports Cauce's premise of correlated error; the first half warns that choosing a delegate is itself a heuristic that novices may use badly.

Citizens may also choose by identity rather than by record. Realist accounts hold that voters rely on social identities and partisan loyalties more than on issues [Achen and Bartels, 2016], and elite polarization "intensifies the impact of party endorsements on opinions, decreases the impact of substantive information" and raises confidence in the resulting opinions [Druckman et al., 2013]. Demand for delegation exists — "even as the people call for giving more influence to people like them, they make it clear that they would prefer not to be much involved in political decision making" [Hibbing and Theiss-Morse, 2002] — but a preference for decisions by experts is a minority profile: technocratic attitudes are "pervasive", yet citizens with a distinctly technocratic profile were about 12% of a nine-country sample [Bertsou and Caramani, 2022]. And delegation can be sought for what it does to blame: "responsibility attribution can be effectively shifted", which "can constitute a strong motive for the delegation of a decision right" [Bartling and Fischbacher, 2012]. Uptake of delegation is therefore weak evidence of trust in the record.

Legitimacy may be less of an obstacle than one might fear, although no study we found tests delegated or record-weighted voting directly. "Outcome favorability is still overwhelmingly the strongest determinant of individuals' willingness to accept authoritative decisions" [Esaiasson et al., 2019], and across twelve experiments participatory processes were judged fairer than representative ones "even when the outcomes are unfavourable" [Werner and Marien, 2022]. Cauce's unconditional override plausibly keeps it on the participatory side of that comparison (our inference). All of this can be measured before a deployment. A conjoint experiment could estimate how much weight citizens actually give a delegate's Brier score against a party label or a sense of similarity, and Lupia's emulation test could compare delegators' effective votes with those of well-informed direct voters.

12.5 Diversity versus ability

Landemore [2013] argues that inclusive procedures make the many collectively smarter than the few, because collective intelligence depends on cognitive diversity as well as individual competence. Hong and Page [2004]'s result that randomly selected problem solvers can outperform the best-performing ones is often cited for this, though it concerns sequential search rather than voting, and Thompson [2014] argues that its central theorem is mathematically trivial once corrected. On the voting side, Caragiannis and Micha [2019] show that delegating to *less* informed voters can sometimes raise accuracy, so "delegate to the most accurate" is not optimal in general. Our results give the voting-model version of Landemore's point a precise scope. When the many are individually better than chance, their independent votes beat any finite committee (Proposition 6(ii); direct democracy's 100% at $\theta_{\text{pool}} = 0.65$). When they are not, no amount of independence helps and a competent committee is the only route to a correct outcome (part (iii)). Which regime a topic is in is an empirical question that differs across topics — one reason Cauce scopes delegation per topic rather than globally.

12.6 Correlated error, misinformation and motivated reasoning

Caplan [2007] argues that the "miracle of aggregation" fails because voters' errors are systematic rather than random. Misinformation spreads in exactly that way: on Twitter, false stories spread farther, faster and more broadly than true ones, and the truth took about six times as long to reach 1,500 people [Vosoughi et al., 2018]. The 5G–COVID-19 conspiracy theory that motivates this paper spread through such networks [Ahmed et al., 2020]. Sections 8.5–8.6 are our answer. A shock that fools everyone defeats every rule (Proposition 10); a partial shock defeats direct democracy abruptly and delegation gradually (Corollary 3). How much reputation weighting helps depends on whether susceptibility falls with competence. The evidence here is split. Reflective thinking predicts better discernment of false headlines, even ideologically congenial ones [Pennycook and Rand, 2019], and lower conspiracy belief ($r = -0.19$ across 145 samples; Yelbuz et al., 2022); both support the competence-linked profile. On a politically charged version of a numeracy task, however, polarization "did not abate among subjects highest in numeracy; instead, it increased" [Kahan et al., 2017]. That is closer to uniform, or even inverted, susceptibility. We did not simulate the inverted case; under it, reputation weighting would concentrate weight on exactly the voters most prone to identity-driven error on politicized topics, and could do worse than Table 10 suggests. Corollary 5 gives the exact condition under which a better record means a worse vote on a shock bill. There is also a timing problem. Reputation accrues over the months or years a policy prediction takes to resolve, while a viral falsehood spreads in days [Vosoughi et al., 2018], so the score can never react to a fresh shock — the shock has to be absorbed by records built beforehand.

12.7 Scoring a vote's prediction: counterfactuals, influence and delay

Proposition 4(i) guarantees honest probabilities only when the forecaster cannot influence which outcome is observed. A Cauce delegate can, in two ways, and has two further reasons to distort her record. First, a prediction about a bill's effect can only be checked if the bill passes; predictions attached to bills that failed are never scored, and a delegate's vote affects which branch occurs. Chen et al. [2014] prove that when a decision depends on experts' predictions, making honest reporting of action-contingent predictions strictly optimal requires the decision maker to choose *every* action with positive probability. A binding vote does not. In decision markets with deterministic rules, Othman and Sandholm [2010] show that a trader can gain by exaggerating the success probability of a suboptimal action. Second, predictions can shape the outcomes they forecast [Perdomo et al., 2020]. A delegate predicting a policy's failure can vote to make it fail. Third, a delegate still chooses which topics to accept delegations on and which claim to forecast (Section 12.1, Proposition 13). Fourth, because delegations reward rank, she has a reason to overstate her confidence even on claims she cannot influence (Proposition 14).

What the design can do is partial. The cap bounds any delegate's share of the tally and hence, though it does not remove, its pivotal influence over which branch is scored. Asking for forecasts on quantities resolved regardless of the vote — or for forecasts under both branches, scoring whichever occurs — narrows the gap without closing it; Chen et al. [2014]'s alternative of eliciting recommendations rather than predictions escapes the impossibility but produces no Brier score. For predictions that never resolve, reciprocal scoring is a partial remedy (Section 12.12). We regard the incentive design of the scoring layer as the main open technical problem this paper leaves. It is the liquid-democracy counterpart of the problem futarchy faces in scoring bets on policies never enacted [Hanson, 2013].

12.8 Topic-scoped delegation and collective inconsistency

Topic scoping is Cauce's most distinctive feature, and it has a price that Christoff and Grossi [2017] anticipated and that the literature on judgment aggregation makes precise [List and Pettit, 2002].

Proposition 18 (Topic-scoped delegation need not preserve consistency)**.** *There exist three consistent delegates, a consistent citizen, and a topic-scoped delegation of that citizen such that the ballots cast on the citizen's behalf are jointly inconsistent. Moreover, if three equal blocs of citizens follow the three delegates, issue-by-issue majority yields an inconsistent collective outcome.*

Proof. Consider the bills and delegates of Table 20, subject to the budget constraint that the three bills cannot all pass ($B_1 \wedge B_2 \Rightarrow \neg B_3$). Each delegate's ballot set satisfies the constraint, as does the citizen's own view. A citizen who delegates economy to E, health to H and fiscal policy to E has Yes cast on all three bills on her behalf. With three equal blocs following E, H and F on every topic, each bill passes two to one, again violating the constraint. □

> **In plain words.** Three friends each plan a trip that fits the budget. Take the flights from the first plan, the hotel from the second and the tours from the third, and the total can exceed what any of them would spend. Delegating each topic to its own best expert can likewise yield a set of decisions that no single person would sign. The same happens with plain issue-by-issue majorities.

By List and Pettit [2002], no aggregation rule that is anonymous and systematic, accepts every consistent profile and always produces complete collective judgments can rule this out; rules that do avoid it, such as supermajority thresholds, do so by leaving some questions undecided. The problem is therefore structural rather than an artifact of Cauce. Mitigations are organizational. Logically

Table 20: A discursive dilemma under topic-scoped delegation. The constraint is that bills B_1–B_3 cannot all pass. Every delegate is consistent; the ballots cast for a citizen who delegates economy and fiscal policy to E and health to H, and issue-by-issue majority over three equal blocs, are not.

	B_1: cut income tax (economy)	B_2: build hospital (health)	B_3: balanced-budget rule (fiscal)	Satisfies constraint?
Delegate E (economist)	Yes	No	Yes	yes
Delegate H (epidemiologist)	No	Yes	Yes	yes
Delegate F (fiscal analyst)	Yes	Yes	No	yes
Citizen's own view	No	Yes	Yes	yes
Cast on the citizen's behalf	Yes	Yes	Yes	**no**
Majority of three equal blocs	Yes	Yes	Yes	**no**

linked bills can be placed in one topic so that one delegate answers all of them; the review window can warn a citizen when the ballots cast on her behalf violate a constraint she declares; and budget questions can be voted as packages. Topic classification itself is a form of agenda power — whoever decides that a bill is "fiscal" rather than "health" decides whose delegate votes on it — and would need its own safeguards.

12.9 Political equality and the epistocracy debate

Mill [1861] proposed plural votes for the educated; Brennan [2016] revives the case, sorting citizens into "hobbits", "hooligans" and "vulcans" and defending forms of epistocracy from restricted suffrage to weighted voting. Remark 2 shows why the mathematics seems to point that way: the optimal rule gives misinformed voters zero or negative weight. Cauce declines to follow it. Every citizen keeps an equal vote and an unconditional override; the platform publishes information and assigns nothing. This answers Estlund [2008]'s "expert/boss fallacy" — that knowing better does not by itself confer authority — since a Cauce delegate's authority comes only from citizens' revocable consent. Formally, the equality Cauce offers is an equal endowment, not an anonymous aggregator. By May's theorem, simple majority is the only rule over two alternatives that satisfies anonymity, neutrality and positive responsiveness on every profile [May, 1952], and a delegated tally is not anonymous in the ballots cast; Cauce's claim to equality rests on each citizen's single unit and the unconditional override (Proposition 1). It does not answer his demographic objection: those best placed to build a strong track record (with the time, training and access to data to forecast well) may be demographically skewed in ways that bias what they get right. How competence is measured matters here too. About half of the gender gap in political-knowledge quizzes "is illusory, reflecting response patterns that work to the collective advantage of male respondents" [Mondak and Anderson, 2004]; under a proper scoring rule guessing earns nothing in expectation, so the response-style advantage behind that artifact disappears, and who appears competent under a Brier record may differ from who appears competent on a quiz. Under Accuracy choice our simulated influencers keep 2% of delegated weight (Table 5). They are marginalized voluntarily, by citizens' own choices, but marginalized all the same. Whether that is a feature or a harm depends on whether the record measures what matters, which Section 12.1 leaves open.

12.10 Concentration in deployed systems

Every large deployment we found concentrates power. In LiquidFeedback a few dozen super-voters held most delegations [Kling et al., 2015]. In three major blockchain governance systems, "the majority of voting power is concentrated in the hands of a small number of addresses" — by the authors' count, 8, 11 and 18 delegates respectively could together control governance votes in Compound, Uniswap and ENS — though these powerful entities rarely overturned the outcome preferred by the wider community [Fritsch et al., 2024]. Delegation is not always the concentrating force there: when large holders are inactive, "delegation tends (perhaps counterintuitively) to increase decentralization" as measured by voting-bloc entropy [Fábrega et al., 2025]. Token weighting there makes the concentration plutocratic rather than reputational, so it is only indirect evidence for Cauce, where each citizen carries one unit. Our own populations show reputation weighting concentrating power too (Gini 0.405), just on competent delegates. The cap bounds any single delegate but not a coalition of similar ones, nor a single bribed delegate on a close bill (Section 9.6), and Section 8.4 shows it offers no epistemic protection. A delegative power index [Zhang and Grossi, 2021] would be a natural quantity for a deployment to monitor and publish.

12.11 Is it legal, and can it be run securely?

Electoral law is wary of exactly what Cauce does. The Council of Europe's reference standard requires that "very strict rules must apply to voting by proxy; the number of proxies a single voter may hold must be limited", and, under secret suffrage, that "family voting and any other form of control by one voter over the vote of another must be prohibited" [European Commission for Democracy through Law (Venice Commission), 2002]. Where proxy voting is allowed, the limit is tiny. A French proxy-holder "ne peut disposer de plus de deux procurations" [Conseil constitutionnel, 2022], and in the Netherlands "de gemachtigde persoon mag maximaal twee volmachten aannemen" [Kiesraad, 2026]. A UK review of electoral fraud found proxy voting "far less regulated than postal voting" and recommended reducing to two the number of close relatives for whom one person may act [Pickles, 2016]. A cap of a quarter of the vote is of a different order. Cauce cannot claim proxy voting as precedent; it must argue that revocability, the per-bill override, publicity and expiry answer the concerns behind these limits.

Two constitutional standards bear directly on the design. Under the doctrine of equal suffrage, "the right of suffrage can be denied by a debasement or dilution of the weight of a citizen's vote just as effectively as by wholly prohibiting the free exercise of the franchise" [Reynolds]. Cauce's cap discards the excess weight of a capped delegate, so a citizen who delegated to her ends with less than one effective unit while a direct voter keeps a full one. A dilution argument of this kind, made by analogy from legislative apportionment, is a further reason beyond the arithmetic of Section 9.6 to return excess weight to its delegators. And Germany's Federal Constitutional Court held that when electronic voting devices are used "die wesentlichen Schritte der Wahlhandlung und der Ergebnisermittlung vom Bürger zuverlässig und ohne besondere Sachkenntnis überprüft werden können" must be possible [Bundesverfassungsgericht, 2009] — the essential steps must be verifiable by citizens without special expertise. Extending that standard from voting devices to tally software is our inference, but a fractional, transitive, capped tally is far harder to recheck by hand than a plurality count, so a deployment would have to publish everything needed to recompute it deterministically, and say which parts of the delegation graph stay secret and how they are audited.

Remote voting is the other obstacle. "Internet- and blockchain-based voting would greatly increase the risk of undetectable, nation-scale election failures" [Park et al., 2021], and an analysis of Estonia's system found "serious architectural limitations and procedural gaps that potentially

jeopardize the integrity of elections" [Springall et al., 2014]. On participation, Estonia's eleven elections with internet voting are the main counterweight: it "did not increase turnout in a context where voting was already accessible, nor did it create digital divides among users" [Ehin et al., 2022]. In a Brazilian policy referendum, by contrast, online voting "is mainly used by new participants rather than just for convenience by those who were already mobilized" [Spada et al., 2016], so who participates, and therefore who delegates and to whom, depends on the channel. The review window, in which delegators may override, is the most attractive point of attack: suppressing overrides shifts weight to delegates. For binding decisions in high-stakes domains, Cauce is most defensible as a consultative layer or with a paper-verifiable fallback; Section 10 lists the matters that several constitutions exclude from popular votes altogether.

12.12 Alternatives that target the same failure

Cauce is one answer to the problem that a majority can be confidently wrong. Table 21 sets it beside other mechanisms aimed at the same failure.

Four points stand out. First, no mechanism escapes the trade-off this paper studies. Those that need no outcomes score reports against other reports, so a shock that moves most reports also moves the standard. Peer prediction makes honest reporting an equilibrium, but its authors acknowledge that non-truthful equilibria exist [Miller et al., 2005], and later work calls "the existence of additional equilibria that payoff-dominate truthful behavior and reveal no useful information" its Achilles heel [Shnayder et al., 2016]. Reciprocal scoring asks forecasters to predict another group's median forecast; in its first test, forecasters scored this way were as accurate as forecasters predicting resolvable outcomes under a proper scoring rule [Karger et al., 2021]. It is the most practical way to score Cauce's predictions about bills that failed (Section 12.7), but what it rewards is anticipating a consensus. The relative score of Proposition 13 uses other forecasters only as a baseline and the outcome as the standard, which is why it rewards resisting a shock.

Second, the surprisingly popular principle is the closest competitor for the scenario that motivated Cauce. It selects "the answer that is more popular than people predict". Asked whether Philadelphia is the capital of Pennsylvania, most respondents say yes, but the minority who say no "probably possess an additional piece of evidence, that the capital is Harrisburg", and they expect the majority to say yes. Across 490 items, the method "reduced errors by 21.3% relative to simple majority vote" [Prelec et al., 2017]. It needs no track record and no unequal weight — only a second question on every ballot, each voter's estimate of the share voting YES, which a Cauce ballot could carry at little cost. Its guarantee rests on assumptions about how informed and uninformed voters predict one another, and we know of no test on politically charged binary questions, where identity-protective reasoning [Kahan et al., 2017] may shape the predictions as well as the answers. We see it as a complement rather than a rival: a platform could publish the surprisingly popular answer beside the tally, as a warning that a shock may be under way.

Third, markets and mini-publics change different things. Prediction market forecasts are "typically fairly accurate" and "outperform most moderately sophisticated benchmarks" [Wolfers and Zitzewitz, 2004], but they weigh beliefs by money, which is the epistocracy problem with a price attached. Cauce is closer to a prediction poll, and the head-to-head evidence favours polls with records: "team prediction polls outperformed prediction markets when forecasts were statistically aggregated using temporal decay, differential weighting based on past performance, and recalibration", with the largest advantage "at the beginning of long-duration questions" [Atanasov et al., 2017]. A later, larger comparison finds prediction polls and markets "statistically tied", while "small, elite crowds outperform larger ones" [Atanasov et al., 2025]; the best-performing aggregations gave the most accurate forecaster many times the median forecaster's weight, although those weights

Table 21: Mechanisms aimed at the failure that motivates Cauce: a majority that is confidently wrong. "Needs outcomes": whether participants are scored or paid against outcomes that must resolve.

Mechanism	Needs outcomes?	Equal say?	Against a shock that misleads most voters	Main weakness
Direct majority	No	Yes	Fails once susceptibility passes σ^* (Cor. 3)	Counts a ballot cast on a refuted belief like any other
Cauce: delegation with Brier records (this paper)	Yes	Yes; weight moves only by consent	Degrades gradually (Table 10)	Records are slow and gameable; cheap capture (§9)
Prediction markets, futarchy [Wolfers and Zitzewitz, 2004, Hanson, 2013]	Yes	No; influence scales with money	Informed traders profit by betting against the crowd	Plutocratic; bets on policies not enacted cannot be settled
Surprisingly popular answer, Bayesian truth serum [Prelec et al., 2017, Prelec, 2004]	No	Yes	Can recover a correct minority answer	Every respondent must also predict others' answers; untested on contested political questions
Peer prediction [Miller et al., 2005, Shnayder et al., 2016]	No	Yes	Scores agreement with others' reports, which a shock moves too	Uninformative equilibria can pay more than honesty
Reciprocal scoring [Karger et al., 2021]	No	— (scores forecasters)	Rewards predicting the consensus, which a shock moves too	May reward coordinating on a focal consensus (our inference)
Deliberative polls, mini-publics [Fishkin, 2018, Landemore, 2020]	No	In expectation, by random selection	Changes what participants know, not whose ballot counts	A small sample decides or advises
Voting advice applications [Fossen and Anderson, 2014]	No	Yes	None: matches preferences, not facts	Addresses alignment, not competence
Accuracy prompts, prebunking [Pennycook et al., 2021, Roozenbeek et al., 2022]	No	Yes	Lowers susceptibility σ directly	Effects may not carry over to real feeds [Wang et al., 2025]
Quadratic voting [Lalley and Weyl, 2018]	No	Equal budgets; intensity bought at a quadratic price	Credits spent cannot tell knowledge from confidence	May amplify intense, misinformed minorities (our inference)

apply to probability forecasts, not votes. Concentrating weight on the demonstrably accurate is what this evidence rewards, and a cap that binds limits it; where it binds, the cap is a price paid in accuracy for robustness against capture, not a free safeguard. Quadratic voting addresses a different failure: it would "robustly correct the failure of existing democracies to incorporate intensity of preference and knowledge" [Lalley and Weyl, 2018]. Spending vote credits, however, cannot distinguish knowledge from confidence (our inference). Deliberative polls change what a random sample of citizens knows rather than whose ballot counts [Fishkin, 2018], and Landemore [2020] discusses liquid representation as one form of democratic representation compatible with open democracy, alongside lottocratic and self-selected representation. These are complements too: a mini-public is a natural candidate for the neutral body that must word and resolve Cauce's claims (Section 9.2).

Finally, interventions aimed at individuals act on a different parameter of our model. Subtly

shifting attention to accuracy "increases the quality of news that people subsequently share" [Pennycook et al., 2021], and short inoculation videos improved people's ability to discern trustworthy from untrustworthy content in randomized studies and a field study on YouTube [Roozenbeek et al., 2022]. In the terms of Corollary 3, these lower susceptibility σ, and pushing it below σ^* restores direct democracy's self-correction without moving any weight. The evidence on transfer is mixed — in a simulated social media feed, an inoculation effect disappeared once real posts were mixed in [Wang et al., 2025] — but these interventions are the cheapest complement to any of the mechanisms above, Cauce included.

13 Remaining Limitations

What the simulations do and do not establish. Kahng et al. [2021]'s negative result stands: no rule using only local information is *guaranteed* to beat direct voting. Our results are exact statements about a model (Propositions 6–11) and simulations of synthetic populations that confirm them. They are not measurements of a deployment. Delegation rates are fixed rather than chosen, so the simulations cannot show over-delegation (Section 12.2); delegates' errors on ordinary bills are independent; competence is fixed over time rather than learned, and the public estimate of it is about as precise as a record of roughly 125 resolved predictions would allow (Section 10.1 varies this); questions are binary with a ground truth; the influencers are a stark caricature; and both susceptibility profiles are stylized. The direct-democracy baseline assumes that citizens who delegate would otherwise vote independently with the pool's competence. Several of Cauce's features are specified and proved correct but not exercised by the accuracy experiments: every experiment concerns a single topic, no citizen overrides, and the accuracy sweeps have no forwarding. Topic scoping — the design's defining feature — is therefore never simulated; its costs appear only in the analysis of Section 12.8. Proposition 6 shows which way each parameter pushes the result, and the insurance calculation of Corollary 4 can be redone for any other estimate of premiums and gains.

Assumptions behind the epistemic results. Delegation choices are drawn from fixed rules rather than chosen strategically, although rational delegation games can have inefficient equilibria [Bloembergen et al., 2019]; every voter and delegate votes sincerely, although sincere voting need not be an equilibrium [Austen-Smith and Banks, 1996]; every bill has one side every citizen would choose if fully informed, the best case for the informativeness of the score (Proposition 15); delegates report calibrated, constant confidence, although competition for delegations rewards overstatement (Proposition 14); and susceptibility never rises with competence, the case in which an averaged record is safe (Corollary 5).

Identity, coercion and vote-buying. Nothing in the algorithm ensures that one citizen is one unit of weight — an identity-verification problem outside our scope. Nor does it prevent the buying or coercion of *delegates*, who are public and carry many citizens' weight. Section 9.6 quantifies the problem: one to three delegates suffice to reverse a typical bill, the default cap does not bind, and the mitigations — sealed delegation, coarse publication of weights, a cap measured against cast weight — each cost transparency or accuracy.

Cycles. Behrens [2015] argues from practice that delegation cycles are rare and rarely catastrophic. Our model detects and excludes them unconditionally (Definition 3, Proposition 3), since a platform guarantee should hold in the rare case too; our population-scale experiments seed no cycles, so they exercise this path only through the unit test.

14 Conclusion

Caminante, no hay camino,
se hace camino al andar.
(Wayfarer, there is no road;
the road is made by walking.)
— Antonio Machado, *Proverbios y cantares* XXIX, *Campos de Castilla* (1917 edition)

Cauce combines per-topic delegation, an unconditional override, a cap, a chain limit and an expiry with a public, Brier-scored track record for every delegate. We proved its resolution algorithm well-defined and terminating. We showed that, in expectation, a Brier score ranks calibrated delegates exactly as the optimal weighted-majority rule would, while leaving every weight in citizens' hands. We derived a limit theorem that makes the price of delegation exact, and finite-sample bounds showing that delegation cuts the effective number of independent voters from about 2,400 to about 29 in our populations. The results cut both ways, and we think the honest summary is this. Whenever delegates hold enough weight to decide outcomes: when a population is informed, letting everyone vote is better, and delegation costs accuracy; when it is misinformed, delegation to demonstrably competent delegates is the only way to be right. When misinformation strikes as a correlated shock, direct democracy collapses at a critical susceptibility we can compute, while reputation-weighted delegation degrades gradually — though it too falls to a coin flip or worse once a third or more of voters are fooled. Reputation weighting does not make liquid democracy better in normal times. In our simulations it makes it cheap insurance — a one-point premium against a gain of 48 to 63 points on shocked bills, worth buying once roughly one bill in fifty carries a shock that fools a quarter to a third of voters, and costlier against stronger shocks — where popularity-driven delegation is expensive insurance that almost never pays. The insurance has conditions the simulations take for granted. A record needs dozens of resolved predictions to separate delegates whose competence differs by fifteen points and hundreds when it differs by five, and it must be scored relative to other forecasters on claims a neutral body writes, or delegates can game it. Competition for delegations rewards overconfidence, and citizens' own preference for confident advisers rewards it again. An averaged record is safe only if the most competent are not also the most easily fooled. The score measures competence, not shared ends. The insurance is also not equally available everywhere. Across the ten sectors we examined, it is most valuable for checkable questions in economics and public health, where the pool can be confidently wrong and indicators resolve within weeks or months, provided a record on such indicators carries over to the slower questions of mechanism; records of only three predictions raise its premium about fourfold, which makes it likely to be costlier in slow domains such as education; and where disagreement is about ends — curriculum, the use of force, who may join the polity — or where claims cannot be scored or published, as in deterrence, the score measures the wrong thing. And delegation makes capture cheap: one to three bribed delegates reverse a typical bill, two delegates at the default cap together hold a majority of the counted vote, and measuring the cap against cast weight closes the second gap, at the price of excluding more weight, but not the first. If machines come to carry many votes, the same arithmetic applies to them: a bloc of AI agents is safe below the rumour threshold whatever its quality, and above it what protects the decision is the effective number of independent providers behind the agents, not the number of agents. Whether real citizens would use a visible track record to delegate sparingly and well, rather than over-delegating as laboratory subjects do, whether predictions about policies can be scored without distorting the votes attached to them, whether citizens read a score as evidence of calibration rather than of confidence, and whether susceptibility to misinformation falls or rises with competence on the questions that matter are the

empirical questions that decide whether the insurance is real. They are the experiments this design calls for next. In Machado's words, there is no road: a channel of trust, if it forms at all, will be made by walking it, one resolved prediction at a time.

Coda: A Ballad of the Riverbed

Cada voto es agua clara
que busca llegar al mar;
uno corre por su cuenta,
otro aprende en quién confiar.

Pues no es el cauce más ancho
el que lleva a buen lugar,
sino aquel que, año tras año,
supo el rumbo adivinar.

Y si el cauce se equivoca,
o se vende al que más da,
vuelve la gota a su orilla
en cuanto quiera votar.

Every vote is water, clear,
seeking out its way to sea;
one runs on its own account,
another learns whom to trust.

It is not the widest channel
that will lead the water right,
but the one that, year on year,
read the course and called it right.

And if the channel goes astray,
or sells itself to the highest bid,
the drop returns to its bank
whenever it chooses to vote.

The ballad is ours, written for this paper as a *romance*, the traditional Spanish ballad: lines of eight syllables, the main line of Manrique's *coplas* (which alternate it with a line of four), with every even line ending on the same stressed vowel, *-á*. The English is a free rendering and does not keep the metre.

References

Daron Acemoglu and Pascual Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, 2019.

Daron Acemoglu and James A. Robinson. *Economic Origins of Dictatorship and Democracy*. Cambridge University Press, Cambridge, 2006.

Christopher H. Achen and Larry M. Bartels. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton University Press, Princeton, NJ, 2016.

Steven Adler et al. Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online. arXiv:2408.07892, 2024. URL https://arxiv.org/abs/2408.07892.

Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. COVID-19 and the 5G conspiracy theory: Social network analysis of Twitter data. *Journal of Medical Internet Research*, 22(5):e19458, 2020. doi: 10.2196/19458.

David J. Aldous. A prediction tournament paradox. *The American Statistician*, 75(3):243–248, 2021. doi: 10.1080/00031305.2019.1604430.

Alberto Alesina, Armando Miano, and Stefanie Stantcheva. Immigration and redistribution. *Review of Economic Studies*, 90(1):1–39, 2023. doi: 10.1093/restud/rdac011.

S. Nageeb Ali and Navin Kartik. Herding with collective preferences. *Economic Theory*, 51(3): 601–626, 2012. doi: 10.1007/s00199-011-0609-7.

Shiri Alouf-Heffetz, Tanmay Inamdar, Pallavi Jain, Yash More, and Nimrod Talmon. Controlling delegations in liquid democracy. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2624–2632. IFAAMAS, 2024.

Ramón A. Alvarez, Daniel Zavala-Araiza, David R. Lyon, David T. Allen, et al. Assessment of methane emissions from the U.S. oil and gas supply chain. *Science*, 361(6398):186–188, 2018. doi: 10.1126/science.aar7204.

Zidong An, João Tovar Jalles, and Prakash Loungani. How well do economists forecast recessions? *International Finance*, 21(2):100–121, 2018. doi: 10.1111/infi.12130.

Soren Anderson, Ioana Marinescu, and Boris Shor. Can Pigou at the polls stop us melting the poles? NBER Working Paper 26146, National Bureau of Economic Research, 2019. URL https://www.nber.org/papers/w26146.

Soren Anderson, Ioana Marinescu, and Boris Shor. Can Pigou at the polls stop us melting the poles? *Journal of the Association of Environmental and Resource Economists*, 10(4):903–945, 2023. doi: 10.1086/722970.

Anthropic. Detecting and countering malicious uses of Claude: March 2025. https://anthropic.com/news/detecting-and-countering-malicious-uses-of-claude-march-2025, 2025.

Klaus Armingeon and Philipp Lutz. Citizens' response to a non-responsive government: The case of the Swiss initiative on mass immigration. *Comparative European Politics*, 21(1):133–151, 2023. doi: 10.1057/s41295-022-00306-4.

Ben Armstrong, Shiri Alouf-Heffetz, and Nimrod Talmon. Optimizing viscous democracy. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2643–2650, 2024. doi: 10.24963/ijcai.2024/292.

Stuart Armstrong and Kaj Sotala. How we're predicting AI – or failing to. In Jan Romportl, Eva Zackova, and Jozef Kelemen, editors, *Beyond Artificial Intelligence*. Springer, 2015. doi: 10.1007/978-3-319-09668-1_2.

Brian J. Asquith, Evan Mast, and Davin Reed. Local effects of large new apartment buildings in low-income areas. *Review of Economics and Statistics*, 105(2):359–375, 2023. doi: 10.1162/rest_a_01055.

Pavel Atanasov and Mark Himmelstein. Talent spotting in crowd prediction. In Matthias Seifert, editor, *Judgment in Predictive Analytics*, volume 343 of *International Series in Operations Research & Management Science*, pages 135–184. Springer, Cham, 2023. doi: 10.1007/978-3-031-30085-1_6.

Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3):691–706, 2017. doi: 10.1287/mnsc.2015.2374.

Pavel Atanasov, Jens Witkowski, Barbara Mellers, and Philip Tetlock. Crowd prediction systems: Markets, polls, and elite forecasters. *International Journal of Forecasting*, 41(2):580–595, 2025. doi: 10.1016/j.ijforecast.2023.12.009.

David Austen-Smith and Jeffrey S. Banks. Information aggregation, rationality, and the Condorcet Jury Theorem. *American Political Science Review*, 90(1):34–45, 1996. doi: 10.2307/2082796.

Larry M. Bartels. Uninformed votes: Information effects in presidential elections. *American Journal of Political Science*, 40(1):194–230, 1996.

Björn Bartling and Urs Fischbacher. Shifting the blame: On delegation and responsibility. *Review of Economic Studies*, 79(1):67–87, 2012. doi: 10.1093/restud/rdr023.

Marco Battaglini, Rebecca B. Morton, and Thomas R. Palfrey. The swing voter's curse in the laboratory. *Review of Economic Studies*, 77(1):61–89, 2010. doi: 10.1111/j.1467-937X.2009.00569.x.

Ben Baumberg Geiger. Benefit 'myths'? The accuracy and inaccuracy of public beliefs about the benefits system. *Social Policy & Administration*, 52(5):998–1018, 2018. doi: 10.1111/spol.12347.

Joshua Becker, Devon Brackbill, and Damon Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 114(26):E5070–E5076, 2017. doi: 10.1073/pnas.1615978114.

Jan Behrens. Circular delegations – myth or disaster? *The Liquid Democracy Journal*, 3, 2015.

Piera Bello and Vincenzo Galasso. The politics of ageing and retirement: Evidence from Swiss referenda. *Population Studies*, 75(1):3–18, 2021. doi: 10.1080/00324728.2020.1841270.

Josh Benaloh and Dwight Tuinstra. Receipt-free secret-ballot elections (extended abstract). In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing (STOC '94)*, pages 544–553. ACM, 1994. doi: 10.1145/195058.195407.

Adam J. Berinsky. Assuming the costs of war: Events, elites, and American public support for military conflict. *Journal of Politics*, 69(4):975–997, 2007. doi: 10.1111/j.1468-2508.2007.00602.x.

Adam J. Berinsky, Daniel Halpern, Joseph Y. Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D. Procaccia, and Manon Revel. Tracking truth with liquid democracy. *Management Science*, 71(9):7821–7839, 2025. doi: 10.1287/mnsc.2023.02470.

Matthew Bernhard, Josh Benaloh, J. Alex Halderman, Ronald L. Rivest, Peter Y. A. Ryan, Philip B. Stark, Vanessa Teague, Poorvi L. Vora, and Dan S. Wallach. Public evidence from secret ballots. In *Electronic Voting (E-Vote-ID 2017)*, volume 10615 of *Lecture Notes in Computer Science*, pages 84–109. Springer, 2017. doi: 10.1007/978-3-319-68687-5_6.

Francisco M. Bersetche. Generalizing liquid democracy to multi-agent delegation: a voting weight measure and equilibrium analysis. *International Journal of Game Theory*, 54(2):40, 2025. Preprint arXiv:2209.14128.

Eri Bertsou and Daniele Caramani. People haven't had enough of experts: Technocratic attitudes among citizens in nine European democracies. *American Journal of Political Science*, 66(1):5–23, 2022. doi: 10.1111/ajps.12554.

Mia Bird, Magnus Lofstrom, Brandon Martin, Steven Raphael, and Viet Nguyen. The impact of Proposition 47 on crime and recidivism. Technical report, Public Policy Institute of California, 2018.

Daan Bloembergen, Davide Grossi, and Martin Lackner. On rational delegations in liquid democracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1796–1803, 2019. doi: 10.1609/aaai.v33i01.33011796.

Christian Blum and Christina Isabel Zuber. Liquid democracy: Potentials, problems, and perspectives. *Journal of Political Philosophy*, 24(2):162–182, 2016.

Yuanchao Emily Bo, David V. Budescu, Charles Lewis, Philip E. Tetlock, and Barbara Mellers. An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment and Decision Making*, 12(2):90–103, 2017. doi: 10.1017/S1930297500005647.

Tito Boeri and Guido Tabellini. Does information increase political support for pension reform? *Public Choice*, 150(1):327–362, 2012. doi: 10.1007/s11127-010-9706-6.

Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. Viscous democracy for social networks. *Communications of the ACM*, 54(6):129–137, 2011. doi: 10.1145/1953122.1953154.

Dion Bongaerts, Thomas Lambert, Daniel Liebau, and Peter Roosenboom. Vote delegation in DeFi governance. arXiv:2503.11940, 2025.

George J. Borjas. The wage impact of the Marielitos: A reappraisal. *ILR Review*, 70(5):1077–1110, 2017. doi: 10.1177/0019793917692945.

Samuel Breckenridge, Dani Vilardell, Andrés Fábrega, Amy Zhao, Patrick McCorry, and Ari Juels. B-Privacy: Defining and enforcing privacy in weighted voting. In *35th USENIX Security Symposium (USENIX Security 26)*, pages 3279–3298. USENIX Association, 2026.

Jason Brennan. *Against Democracy*. Princeton University Press, Princeton, NJ, 2016.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Markus Brill and Nimrod Talmon. Pairwise liquid democracy. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 137–143, 2018. doi: 10.24963/ijcai.2018/19.

Markus Brill, Théo Delemazure, Anne-Marie George, Martin Lackner, and Ulrike Schmidt-Kraepelin. Liquid democracy with ranked delegations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4884–4891, 2022. doi: 10.1609/aaai.v36i5.20417.

Robert J. Brulle. The climate lobby: A sectoral analysis of lobbying spending on climate change in the USA, 2000 to 2016. *Climatic Change*, 149(3):289–303, 2018. doi: 10.1007/s10584-018-2241-z.

Erik Brynjolfsson, Bharat Chandar, and Ruyu Chen. Canaries in the coal mine? six facts about the recent employment effects of artificial intelligence. Technical report, Stanford Digital Economy Lab, 2025.

David V. Budescu and Eva Chen. Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280, 2015. doi: 10.1287/mnsc.2014.1909.

Bundesverfassungsgericht. Urteil des zweiten senats vom 3. märz 2009, 2 bvc 3/07, 2 bvc 4/07 (bverfge 123, 39). https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2009/03/cs20090303_2bvc000307.html, 2009.

Andrea Louise Campbell. *How Policies Make Citizens: Senior Political Activism and the American Welfare State*. Princeton University Press, Princeton, NJ, 2003.

Donald T. Campbell. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90, 1979.

Joseph Campbell, Alessandra Casella, Lucas de Lara, Victoria R. Mooers, and Dilip Ravindran. Liquid democracy. Two experiments on delegation in voting. NBER Working Paper 30794, National Bureau of Economic Research, 2022.

Nicolás Campos, Eduardo Engel, Ronald D. Fischer, and Alexander Galetovic. The ways of corruption in infrastructure: Lessons from the Odebrecht case. *Journal of Economic Perspectives*, 35(2):171–190, 2021. doi: 10.1257/jep.35.2.171.

Bryan Caplan. *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*. Princeton University Press, Princeton, NJ, 2007.

Ioannis Caragiannis and Evi Micha. A contribution to the critique of liquid democracy. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 116–122, 2019. doi: 10.24963/ijcai.2019/17.

David Card. The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review*, 43(2):245–257, 1990. doi: 10.1177/001979399004300205.

Alberto Cavallo. Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics*, 60(2):152–165, 2013. doi: 10.1016/j.jmoneco.2012.10.002.

Krishnendu Chatterjee, Seth Gilbert, Stefan Schmid, Jakub Svoboda, and Michelle Yeo. When is liquid democracy possible? On the manipulation of variance. In *Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC '25)*. ACM, 2025. doi: 10.1145/3732772.3733544.

Yiling Chen, Ian A. Kash, Michael Ruberry, and Victor Shnayder. Eliciting predictions and recommendations for decision making. *ACM Transactions on Economics and Computation*, 2(2):6:1–6:27, 2014. doi: 10.1145/2556271.

Raj Chetty, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, 126(4):1593–1660, 2011. doi: 10.1093/qje/qjr041.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the impacts of teachers: Reply. *American Economic Review*, 107(6):1685–1717, 2017. doi: 10.1257/aer.20170108.

Matthew M. Chingos. The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review*, 31(5):543–562, 2012. doi: 10.1016/j.econedurev.2012.03.002.

Gabriel Chodorow-Reich, John Coglianese, and Loukas Karabarbounis. The macro effects of unemployment benefit extensions: A measurement error approach. *Quarterly Journal of Economics*, 134(1):227–279, 2019. doi: 10.1093/qje/qjy018.

Zoé Christoff and Davide Grossi. Binary voting with delegable proxy: An analysis of liquid democracy. In *Proceedings of the 16th Conference on Theoretical Aspects of Rationality and Knowledge (TARK'17)*, volume 251 of *EPTCS*, pages 134–150, 2017. doi: 10.4204/EPTCS.251.10.

Michael A. Clemens and Jennifer Hunt. The labor market effects of refugee waves: Reconciling conflicting results. *ILR Review*, 72(4):818–857, 2019.

Rachael Colley, Théo Delemazure, and Hugo Gilbert. Measuring a priori voting power in liquid democracy. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2607–2615, 2023. doi: 10.24963/ijcai.2023/290.

David Collingridge. *The Social Control of Technology*. Frances Pinter, London, 1980.

Conseil constitutionnel. Quelles sont les règles pour voter par procuration? https://presidentielle2022.conseil-constitutionnel.fr/le-vote/etre-electeur-et-voter/regles-voter-procuration.html, 2022.

J. Anthony Cookson, Corbin Fox, Javier Gil-Bazo, Juan F. Imbet, and Christoph Schiller. Social media as a bank run catalyst. *Journal of Financial Economics*, 176, 2026. doi: 10.1016/j.jfineco.2025.104218.

Corporate Europe Observatory and LobbyControl. Revealed: Tech industry now spending record €151 million on lobbying the EU. https://corporateeurope.org/en/2025/10/revealed-tech-industry-now-spending-record-eu151-million-lobbying-eu, 2025.

Corte costituzionale. Sentenza n. 70/2015. https://www.cortecostituzionale.it/scheda-pronuncia/2015/70, 2015.

Corte costituzionale. Sentenza n. 50/2022. Decided 15 February 2022, filed 2 March 2022, https://www.cortecostituzionale.it/scheda-pronuncia/2022/50, 2022a.

Corte costituzionale. Sentenza n. 51/2022 (referendum on the cultivation of narcotic plants). https://www.cortecostituzionale.it/actionSchedaPronuncia.do?param_ecli=ECLI:IT:COST:2022:51, 2022b.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 2024. doi: 10.1126/science.adq1814.

Council of Europe. Protocol no. 1 to the convention for the protection of human rights and fundamental freedoms, art. 2 (right to education). ETS No. 9, Paris, 1952.

Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, Johannes Bracher, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022. doi: 10.1073/pnas.2113561119.

Felix Creutzig, Peter Agoston, Jan Christoph Goldschmidt, Gunnar Luderer, Gregory Nemet, and Robert C. Pietzcker. The underestimated potential of solar energy to mitigate climate change. *Nature Energy*, 2(9):17140, 2017. doi: 10.1038/nenergy.2017.140.

Philip Daian, Tyler Kell, Ian Miers, and Ari Juels. On-chain vote buying and the rise of dark DAOs. Hacking, Distributed (blog), https://hackingdistributed.com/2018/07/02/on-chain-vote-buying/, 2018.

Gianlorenzo D'Angelo, Esmaeil Delfaraz, and Hugo Gilbert. Computation and bribery of voting power in delegative simple games. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 336–344. IFAAMAS, 2022.

Jonas Degrave. Resolving multi-proxy transitive vote delegation. Technical Report arXiv:1412.4039, arXiv, 2014.

Eddie Dekel and Michele Piccione. Sequential voting procedures in symmetric binary elections. *Journal of Political Economy*, 108(1):34–55, 2000. doi: 10.1086/262110.

Stefano DellaVigna and Devin Pope. Predicting experimental results: Who knows what? *Journal of Political Economy*, 126(6):2410–2456, 2018. doi: 10.1086/699976.

Franz Dietrich and Kai Spiekermann. Epistemic democracy with defensible premises. *Economics and Philosophy*, 29(1):87–120, 2013a. doi: 10.1017/S0266267113000096.

Franz Dietrich and Kai Spiekermann. Independent opinions? On the causal foundations of belief formation and jury theorems. *Mind*, 122(487):655–685, 2013b. doi: 10.1093/mind/fzt074.

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. doi: 10.1037/xge0000033.

George Disney, Arkadiusz Wiśniowski, Jonathan J. Forster, Peter W. F. Smith, and Jakub Bijak. Evaluation of existing migration forecasting methods and models. Technical report, ESRC Centre for Population Change, University of Southampton, for the Migration Advisory Committee, October 2015.

Anil R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), 2024. doi: 10.1126/sciadv.adn5290.

John R. Douceur. The Sybil attack. In *Peer-to-Peer Systems: First International Workshop, IPTPS 2002*, volume 2429 of *Lecture Notes in Computer Science*, pages 251–260. Springer, 2002. doi: 10.1007/3-540-45748-8_24.

Thomas Douenne and Adrien Fabre. Yellow vests, pessimistic beliefs, and carbon tax aversion. *American Economic Journal: Economic Policy*, 14(1):81–110, 2022. doi: 10.1257/pol.20200092.

Anthony Downs. *An Economic Theory of Democracy*. Harper & Brothers, New York, 1957.

Luke Drago and Rudolf Laine. The intelligence curse. Essay series, https://intelligence-curse.ai/, 2025.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018. doi: 10.1126/sciadv.aao5580.

James N. Druckman, Erik Peterson, and Rune Slothuus. How elite partisan polarization affects public opinion formation. *American Political Science Review*, 107(1):57–79, 2013. doi: 10.1017/S0003055412000500.

Moritz A. Drupp, Mark C. Freeman, Ben Groom, and Frikk Nesje. Discounting disentangled. *American Economic Journal: Economic Policy*, 10(4):109–134, 2018. doi: 10.1257/pol.20160240.

e-Estonia. Estonian e-identity: ID card. https://e-estonia.com/solutions/estonian-e-identity/id-card/, 2025.

Piret Ehin, Mihkel Solvak, Jan Willemson, and Priit Vinkel. Internet voting in Estonia 2005–2019: Evidence from eleven elections. *Government Information Quarterly*, 39(4):101718, 2022. doi: 10.1016/j.giq.2022.101718.

Katherine Levine Einstein, Maxwell Palmer, and David M. Glick. Who participates in local government? Evidence from meeting minutes. *Perspectives on Politics*, 17(1):28–46, 2019. doi: 10.1017/S153759271800213X.

Jonas Eliasson. Cost overruns of infrastructure projects – distributions, causes and remedies. *Transportation Research Part A: Policy and Practice*, 198:104532, 2025. doi: 10.1016/j.tra.2025.104532.

Jonas Eliasson and Mogens Fosgerau. Cost overruns and demand shortfalls – deception or selection? *Transportation Research Part B: Methodological*, 57:105–113, 2013. doi: 10.1016/j.trb.2013.09.005.

Christopher S. Elmendorf, Clayton Nall, and Stan Oklobdzija. The folk economics of housing. *Journal of Economic Perspectives*, 39(3):45–66, 2025. doi: 10.1257/jep.20241428.

Peter K. Enns. The public's increasing punitiveness and its influence on mass incarceration in the United States. *American Journal of Political Science*, 58(4):857–872, 2014. doi: 10.1111/ajps.12098.

Peter Esaiasson, Mikael Persson, Mikael Gilljam, and Torun Lindholm. Reconsidering the role of procedures for decision acceptance. *British Journal of Political Science*, 49(1):291–314, 2019. doi: 10.1017/S0007123416000508.

Bruno Escoffier, Hugo Gilbert, and Adèle Pass-Lanneau. The convergence of iterative delegations in liquid democracy in a social network. In *Algorithmic Game Theory (SAGT 2019)*, volume 11801 of *Lecture Notes in Computer Science*, pages 284–297. Springer, 2019. doi: 10.1007/978-3-030-30473-7_19.

David Estlund. *Democratic Authority: A Philosophical Framework*. Princeton University Press, Princeton, NJ, 2008.

Estonian National Electoral Committee. Statistics about internet voting in Estonia. https://www.valimised.ee/en/archive/statistics-about-internet-voting-estonia, 2023.

John A. Eterno, Arvind Verma, and Eli B. Silverman. Police manipulations of crime reporting: Insiders' revelations. *Justice Quarterly*, 33(5):811–835, 2016. doi: 10.1080/07418825.2014.980838.

European Commission for Democracy through Law (Venice Commission). Code of good practice in electoral matters: Guidelines and explanatory report. Technical Report CDL-AD(2002)023, Council of Europe, 2002. URL https://rm.coe.int/090000168092af01.

Andrés Fábrega, Amy Zhao, Jay Yu, James Austgen, Sarah Allen, Kushal Babel, Mahimna Kelkar, and Ari Juels. Voting-bloc entropy: A new metric for DAO decentralization. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 1299–1318. USENIX Association, 2025.

Chuxuan Fan, Tao Shu, and Fei Xie. Is there wisdom among the DAO crowd? Evidence from vote delegation. Finance Working Paper 1077/2025, European Corporate Governance Institute, 2025. Revised June 2026.

Timothy J. Feddersen and Wolfgang Pesendorfer. The swing voter's curse. *American Economic Review*, 86(3):408–424, 1996.

Lars P. Feld and John G. Matsusaka. Budget referendums and government spending: Evidence from Swiss cantons. *Journal of Public Economics*, 87(12):2703–2724, 2003.

Juan J. Fernández, Gema García-Albacete, Antonio M. Jaime-Castillo, and Jonas Radl. Priming or learning? The influence of pension policy information on individual preferences in Germany, Spain and the United States. *Journal of European Social Policy*, 33(3):337–352, 2023. doi: 10.1177/09589287231164347.

Philip M. Fernbach, Nicholas Light, Sydney E. Scott, Yoel Inbar, and Paul Rozin. Extreme opponents of genetically modified foods know the least but think they know the most. *Nature Human Behaviour*, 3(3):251–256, 2019. doi: 10.1038/s41562-018-0520-3.

James S. Fishkin. *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation*. Oxford University Press, Oxford, 2018. doi: 10.1093/oso/9780198820291.001.0001.

Bent Flyvbjerg. Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, 16(1):3–21, 2008. doi: 10.1080/09654310701747936.

Bent Flyvbjerg, Mette K. Skamris Holm, and Søren L. Buhl. Underestimating costs in public works projects: Error or lie? *Journal of the American Planning Association*, 68(3):279–295, 2002. doi: 10.1080/01944360208976273.

Bent Flyvbjerg, Mette K. Skamris Holm, and Søren L. Buhl. How (in)accurate are demand forecasts in public works projects? The case of transportation. *Journal of the American Planning Association*, 71(2):131–146, 2005. doi: 10.1080/01944360508976688.

Thomas Fossen and Joel Anderson. What's the point of voting advice applications? Competing perspectives on democracy and citizenship. *Electoral Studies*, 36:244–251, 2014. doi: 10.1016/j.electstud.2014.04.001.

Daniel Freeman, Felicity Waite, Laina Rosebrock, Ariane Petit, Chiara Causier, Anna East, Lucy Jenner, Ashley-Louise Teale, Lydia Carr, Sophie Mulhall, Emily Bold, and Sinéad Lambe. Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological Medicine*, 52(2):251–263, 2022. doi: 10.1017/S0033291720001890.

Robin Fritsch, Marino Müller, and Roger Wattenhofer. Analyzing voting power in decentralized governance: Who controls DAOs? *Blockchain: Research and Applications*, 5(3):100208, 2024. doi: 10.1016/j.bcra.2024.100208.

Victor R. Fuchs, Alan B. Krueger, and James M. Poterba. Economists' views about parameters, values, and policies: Survey results in labor and public economics. *Journal of Economic Literature*, 36(3):1387–1425, 1998.

Suyash Fulay, Jocelyn Zhu, and Michiel Bakker. From delegates to trustees: How optimizing for long-term interests shapes bias and alignment in LLMs. arXiv:2510.12689, 2025. URL https://arxiv.org/abs/2510.12689.

Cary Funk and Lee Rainie. Public and scientists' views on science and society. Technical report, Pew Research Center, January 2015. URL https://www.pewresearch.org/science/2015/01/29/public-and-scientists-views-on-science-and-society/.

Christopher Gelpi, Peter D. Feaver, and Jason Reifler. Success matters: Casualty sensitivity and the war in Iraq. *International Security*, 30(3):7–46, 2005. doi: 10.1162/isec.2005.30.3.7. Winter 2005/06.

gfs.bern. Analysis CO2 no. https://www.gfsbern.ch/en/news/analysis-co2-no/, 2021.

Seth Gilbert, Stefan Schmid, Santiago Schnell, Jakub Svoboda, and Michelle Yeo. Liquid democracy under vote correlation: On the fallacies of averaging and the excluded middle. arXiv:2608.16738, 2026.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Josh A. Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2):pgae034, 2024.

Paul Gölz, Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. The fluid mechanics of liquid democracy. *ACM Transactions on Economics and Computation*, 9(4):1–39, 2021. Conference version, "The Fluid Dynamics of Liquid Democracy," in Proceedings of the 13th Conference on Web and Internet Economics (WINE'18), pp. 188–202.

Robert E. Goodin. Enfranchising all affected interests, and its alternatives. *Philosophy & Public Affairs*, 35(1):40–68, 2007. doi: 10.1111/j.1088-4963.2007.00098.x.

Matthew Goodwin and Caitlin Milazzo. Taking back control? Investigating the role of immigration in the 2016 vote for Brexit. *British Journal of Politics and International Relations*, 19(3):450–464, 2017. doi: 10.1177/1369148117710799.

John Gramlich. What the data says about crime in the U.S. Pew Research Center, https://www.pewresearch.org/short-reads/2024/04/24/what-the-data-says-about-crime-in-the-us/, 2024.

James Green-Armytage. Direct voting and proxy voting. *Constitutional Political Economy*, 26(2): 190–220, 2015.

Alexis Grigorieff, Christopher Roth, and Diego Ubfal. Does information change attitudes toward immigrants? *Demography*, 57(3):1117–1143, 2020. doi: 10.1007/s13524-020-00882-8.

Bernard Grofman, Guillermo Owen, and Scott L. Feld. Thirteen theorems in search of the truth. *Theory and Decision*, 15(3):261–278, 1983.

Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656): 398–404, 2023. doi: 10.1126/science.abp9364.

Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24): e2403116121, 2024.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational AI. arXiv:2507.13919; *Science*, doi:10.1126/science.aea3884, 2025. URL https://arxiv.org/abs/2507.13919.

Marcus Hagedorn, Fatih Karahan, Iourii Manovskii, and Kurt Mitman. Unemployment benefits and unemployment in the great recession: The role of macro effects. NBER Working Paper 19499, National Bureau of Economic Research, 2013.

Jens Hainmueller and Dominik Hangartner. Does direct democracy hurt immigrant minorities? Evidence from naturalization decisions in Switzerland. *American Journal of Political Science*, 63 (3):530–547, 2019. doi: 10.1111/ajps.12433.

Jens Hainmueller and Daniel J. Hopkins. Public attitudes toward immigration. *Annual Review of Political Science*, 17:225–249, 2014. doi: 10.1146/annurev-polisci-102512-194818.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. arXiv:2402.18563, 2024. URL https://arxiv.org/abs/2402.18563.

Daniel Halpern, Joseph Y. Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D. Procaccia, and Manon Revel. In defense of liquid democracy. In *Proceedings of the 24th ACM Conference on Economics and Computation (EC '23)*. ACM, 2023. doi: 10.1145/3580507.3597817.

Robin Hanson. Shall we vote on values, but bet on beliefs? *Journal of Political Philosophy*, 21(2): 151–178, 2013.

Steve Hardt and Lia C. R. Lopes. Google votes: A liquid democracy experiment on a corporate social network. Technical Disclosure Commons, Defensive Publications Series, no. 79, 2015.

Björn Hårsman and John M. Quigley. Political and public acceptability of congestion pricing: Ideology and self-interest. *Journal of Policy Analysis and Management*, 29(4):854–874, 2010.

Clifford E. Hauenstein, Rick P. Thomas, David A. Illingworth, and Michael R. Dougherty. Rethinking the role of teams and training in geopolitical forecasting: The effect of uncontrolled method variance on statistical conclusions. *Psychological Science*, 36(1), 2025. doi: 10.1177/09567976241266481.

Zeke Hausfather, Henri F. Drake, Tristan Abbott, and Gavin A. Schmidt. Evaluating the performance of past climate model projections. *Geophysical Research Letters*, 47(1):e2019GL085378, 2020. doi: 10.1029/2019GL085378.

Eric Helland and Alexander Tabarrok. Does three strikes deter? A nonparametric estimation. *Journal of Human Resources*, 42(2):309–330, 2007. doi: 10.3368/jhr.XLII.2.309.

John R. Hibbing and Elizabeth Theiss-Morse. *Stealth Democracy: Americans' Beliefs About How Government Should Work*. Cambridge University Press, Cambridge, 2002.

Mark Himmelstein, Pavel Atanasov, and David V. Budescu. Forecasting forecaster accuracy: Contributions of past performance and individual differences. *Judgment and Decision Making*, 16(2): 323–362, 2021. doi: 10.1017/S1930297500008597.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004. doi: 10.1073/pnas.0403723101.

Daniel J. Hopkins, John Sides, and Jack Citrin. The muted consequences of correct information about immigration. *Journal of Politics*, 81(1):315–320, 2019. doi: 10.1086/699914.

Matthew J. Hornsey, Emily A. Harris, Paul G. Bain, and Kelly S. Fielding. Meta-analyses of the determinants and outcomes of belief in climate change. *Nature Climate Change*, 6(6):622–626, 2016. doi: 10.1038/nclimate2943.

William G. Howell and Martin R. West. Is the price "right"? *Education Next*, 8(3):37–41, 2008.

Victor Huang and Joseph Hall. The effect of AI writing on governance: Evidence from a $3.5 billion DAO. Working paper, https://josephphall.com/papers/ai-writing-dao-governance.pdf, 2026.

Anders Humlum and Emilie Vestergaard. Large language models, small labor market effects. NBER Working Paper 33777, National Bureau of Economic Research, 2025.

Innenministerium Baden-Württemberg. Amtliches endergebnis der volksabstimmung zum S 21-Kündigungsgesetz. https://im.baden-wuerttemberg.de/de/service/presse-und-oeffentlichkeitsarbeit/pressemitteilung/pid/amtliches-endergebnis-der-volksabstimmung-zum-s-21-kuendigungsgesetz-1, 2011.

Institute for Strategic Dialogue. From rumours to riots: How online misinformation fuelled violence in the aftermath of the Southport attack. Technical report, Institute for Strategic Dialogue, 2024. URL https://www.isdglobal.org/digital-dispatch/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/.

International Commission on Non-Ionizing Radiation Protection. Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz). *Health Physics*, 118(5):483–524, 2020. doi: 10.1097/HP.0000000000001210.

Brian A. Jacob and Steven D. Levitt. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3):843–877, 2003. doi: 10.1162/00335530360698441.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. doi: 10.1145/3544548.3581196.

Matt Jaquiery and Nick Yeung. Preferences for advisor agreement and accuracy. *PLOS ONE*, 19 (9):e0311211, 2024. doi: 10.1371/journal.pone.0311211.

Garrett A. Johnson, Scott K. Shriver, and Samuel G. Goldberg. Privacy and market concentration: Intended and unintended consequences of the GDPR. *Management Science*, 69(10):5695–5721, 2023. doi: 10.1287/mnsc.2023.4709.

Daniel Jolley and Jenny L. Paterson. Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology*, 59(3):628–640, 2020. doi: 10.1111/bjso.12394.

J. B. Kadane and Robert L. Winkler. Separating probability elicitation from utilities. *Journal of the American Statistical Association*, 83(402):357–363, 1988. doi: 10.1080/01621459.1988.10478605.

Dan M. Kahan, Ellen Peters, Erica C. Dawson, and Paul Slovic. Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, 1(1):54–86, 2017. doi: 10.1017/bpp.2016.2.

Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. Liquid democracy: An algorithmic perspective. *Journal of Artificial Intelligence Research*, 70:1223–1252, 2021. doi: 10.1613/jair.1.12261. Conference version in Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18), pp. 1095–1102.

Ezra Karger, Joshua Monrad, Barbara Mellers, and Philip E. Tetlock. Reciprocal scoring: A method for forecasting unanswerable questions. SSRN Working Paper No. 3954498, 2021.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025.

Konstantin Kashin, Gary King, and Samir Soneji. Systematic bias and nontransparency in US Social Security Administration forecasts. *Journal of Economic Perspectives*, 29(2):239–258, 2015a. doi: 10.1257/jep.29.2.239.

Konstantin Kashin, Gary King, and Samir Soneji. Explaining systematic bias and nontransparency in US Social Security Administration forecasts. *Political Analysis*, 23(3):336–362, 2015b. doi: 10.1093/pan/mpv011.

Joshua D. Kertzer, Kathleen E. Powers, Brian C. Rathbun, and Ravi Iyer. Moral support: How moral values shape foreign policy attitudes. *Journal of Politics*, 76(3):825–840, 2014. doi: 10.1017/s0022381614000073.

Kiesraad. Volmacht. https://www.kiesraad.nl/verkiezingen/tweede-kamer/stemmen/volmacht, 2026. Accessed 1 October 2026.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 30038–30066, 2025.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118.

Christoph Carl Kling, Jérôme Kunegis, Heinrich Hartmann, Markus Strohmaier, and Steffen Staab. Voting behaviour and power in online democracy: A study of LiquidFeedback in Germany's pirate party. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 9, pages 208–217, 2015.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. arXiv:2605.29800 (preprint), 2026. URL https://arxiv.org/abs/2605.29800.

Anton Korinek and Joseph E. Stiglitz. Artificial intelligence and its implications for income distribution and unemployment. NBER Working Paper 24174, National Bureau of Economic Research, 2017.

Matthew A. Kraft. Interpreting effect sizes of education interventions. *Educational Researcher*, 49 (4):241–253, 2020. doi: 10.3102/0013189X20912798.

Sarah Kreps and Douglas Kriner. The potential impact of emerging technologies on democratic representation: Evidence from a field experiment. *New Media & Society*, 2023. doi: 10.1177/14614448231160526.

Simas Kučinskas, Josh Rosenberg, Rebecca Ceppas de Castro, Zach Jacobs, Jordan Canedy, Philip E. Tetlock, and Ezra Karger. Assessing near-term accuracy in the existential risk persuasion tournament. Technical report, Forecasting Research Institute, 2025. URL https://forecastingresearch.org/research/near-term-xpt-accuracy.

James H. Kuklinski, Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. Misinformation and the currency of democratic citizenship. *Journal of Politics*, 62(3):790–816, 2000. doi: 10.1111/0022-3816.00033.

Steven Kull, Clay Ramsay, and Evan Lewis. Misperceptions, the media, and the Iraq war. *Political Science Quarterly*, 118(4):569–598, 2003.

Alexander Kustov, Dillon Laaker, and Cassidy Reller. The stability of immigration attitudes: Evidence and implications. *Journal of Politics*, 83(4):1478–1494, 2021. doi: 10.1086/715061.

Krishna K. Ladha. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992. doi: 10.2307/2111584.

Steven P. Lalley and E. Glen Weyl. Quadratic voting: How mechanism design can radicalize democracy. *AEA Papers and Proceedings*, 108:33–37, 2018. doi: 10.1257/pandp.20181002.

Hélène Landemore. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton University Press, Princeton, NJ, 2013.

Hélène Landemore. *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*. Princeton University Press, Princeton, NJ, 2020.

Latvijas Republika. Constitution of the Republic of Latvia (Satversme), art. 73. https://www.saeima.lv/en/legislative-process/constitution, 1922.

Richard R. Lau and David P. Redlawsk. Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science*, 45(4):951–971, 2001. doi: 10.2307/2669334.

Richard Ned Lebow and Janice Gross Stein. Deterrence: The elusive dependent variable. *World Politics*, 42(3):336–369, 1990. doi: 10.2307/2010415.

Zhuolun Li and Evangelos Pournaras. Resilient liquid democracy: Mitigating voting power imbalances via secure delegation networks. arXiv:2607.01730, 2026.

Kenneth C. Lichtendahl, Jr. and Robert L. Winkler. Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, 53(11):1745–1755, 2007. doi: 10.1287/mnsc.1070.0729.

Luona Lin, Juliana Menasce Horowitz, Kiley Hurst, and Dana Braga. Teachers', teens' and Americans' views about race and LGBTQ issues in K-12 schools. Technical report, Pew Research Center, 2024. URL https://www.pewresearch.org/social-trends/2024/02/22/race-and-lgbtq-issues-in-k-12-schools/.

Zhiyuan Jerry Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. The limits of human predictions of recidivism. *Science Advances*, 6(7):eaaz0652, 2020. doi: 10.1126/sciadv.aaz0652.

Christian List and Philip Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18(1):89–110, 2002. doi: 10.1017/S0266267102001098.

Sahil Loomba, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3):337–348, 2021. doi: 10.1038/s41562-021-01056-1.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011. doi: 10.1073/pnas.1008636108.

Hugues Lortie-Forgues and Matthew Inglis. Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3):158–166, 2019. doi: 10.3102/0013189X19832850.

Arthur Lupia. Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections. *American Political Science Review*, 88(1):63–76, 1994. doi: 10.2307/2944882.

Kelly Macdonald, Laura Germine, Alida Anderson, Joanna Christodoulou, and Lauren M. McGrath. Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. *Frontiers in Psychology*, 8:1314, 2017. doi: 10.3389/fpsyg.2017.01314.

David R. Mandel and Alan Barnes. Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, 111(30):10984–10989, 2014. doi: 10.1073/pnas.1406138111.

David R. Mandel and Daniel Irwin. Tracking accuracy of strategic intelligence forecasts: Findings from a long-term Canadian study. *Futures & Foresight Science*, 3(3–4):e98, 2021. doi: 10.1002/ffo2.98.

Albert E. Mannes, Jack B. Soll, and Richard P. Larrick. The wisdom of select crowds. *Journal of Personality and Social Psychology*, 107(2):276–299, 2014. doi: 10.1037/a0036677.

Jane Mansbridge. A "selection model" of political representation. *Journal of Political Philosophy*, 17(4):369–398, 2009. doi: 10.1111/j.1467-9760.2009.00337.x.

William Marble and Clayton Nall. Where self-interest trumps ideology: Liberal homeowners and local opposition to housing development. *Journal of Politics*, 83(4):1747–1763, 2021. doi: 10.1086/711717.

César Martinelli. Would rational voters acquire costly information? *Journal of Economic Theory*, 129(1):225–251, 2006. URL https://ideas.repec.org/a/eee/jetheo/v129y2006i1p225-251.html.

Kenneth O. May. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica*, 20(4):680–684, 1952.

Joseph C. McMurray. Aggregating information by voting: The wisdom of the experts versus the wisdom of the masses. *Review of Economic Studies*, 80(1):277–312, 2013. doi: 10.1093/restud/rds026.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E. Scott, Don Moore, Pavel Atanasov, Samuel A. Swift, Terry Murray, Eric Stone, and Philip E. Tetlock. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115, 2014. doi: 10.1177/0956797614524255.

Barbara Mellers, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3):267–281, 2015. doi: 10.1177/1745691615577794.

Menlo Ventures. 2025: The state of generative AI in the enterprise. https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/, 2025.

Johnnatan Messias and Ayae Ide. Fairness in token delegation: Mitigating voting power concentration in DAOs. arXiv:2510.05830, version 2 (May 2026), 2025.

Metaculus. Q2 AI benchmark results: Pros maintain clear lead. https://www.lesswrong.com/posts/Surnjh8A4WjgtQTkZ/q2-ai-benchmark-results-pros-maintain-clear-lead, 2025.

Metaculus. FutureEval spring results: Pros beat bots, but the gap is not significant. https://www.lesswrong.com/posts/wZBbDqzfBjYG58CxK/futureeval-spring-results-pros-beat-bots-but-the-gap-is, 2026.

Robert Michels. *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. Hearst's International Library Co., New York, 1915. Translated by Eden and Cedar Paul; originally published in German in 1911.

Thomas Milic, Alessandro Feller, and Daniel Kübler. VOTO-studie zur eidgenössischen volksabstimmung vom 27. September 2020. Technical report, FORS, ZDA and LINK, 2020. URL https://swissvotes.ch/vote/635.00/nachbefragung-de.pdf.

John Stuart Mill. *Considerations on Representative Government*. Parker, Son, and Bourn, London, 1861.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005. doi: 10.1287/mnsc.1050.0379.

Sarah Miller, Norman Johnson, and Laura R. Wherry. Medicaid and mortality: New evidence from linked survey and administrative data. *Quarterly Journal of Economics*, 136(3):1783–1829, 2021. doi: 10.1093/qje/qjab004.

Daniel Moeckli. Of minarets and foreign criminals: Swiss direct democracy and human rights. *Human Rights Law Review*, 11(4):774–794, 2011. doi: 10.1093/hrlr/ngr026.

Daniel Moeckli. Beyond the reach of the people? Admissibility requirements and procedures for citizens' initiatives in comparative perspective. *European Constitutional Law Review*, 21(2):230–273, 2025. doi: 10.1017/S1574019625000100.

Jeffery J. Mondak and Mary R. Anderson. The knowledge gap: A reexamination of gender-based differences in political knowledge. *Journal of Politics*, 66(2):492–512, 2004. doi: 10.1111/j.1468-2508.2004.00161.x.

Victoria Mooers, Joseph Campbell, Alessandra Casella, Lucas de Lara, and Dilip Ravindran. Liquid democracy or direct democracy? One theoretical result and two experiments. arXiv:2212.09715, revised June 2026, 2026.

A. H. Murphy. Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, 12(1):215–223, 1973. doi: 10.1175/1520-0450(1973)012<0215:HASSFP>2.0.CO;2.

National Academies of Sciences, Engineering, and Medicine. *The Economic and Fiscal Consequences of Immigration*. The National Academies Press, Washington, DC, 2016. URL https://nap.nationalacademies.org/catalog/23550.

National Research Council, Committee on Population. *Beyond Six Billion: Forecasting the World's Population*. National Academy Press, Washington, DC, 2000. doi: 10.17226/9828.

Mahdi Nejadgholi, Nan Yang, and Jeremy Clark. Short paper: Ballot secrecy for liquid democracy. In *Financial Cryptography and Data Security: FC 2021 International Workshops*, Lecture Notes in Computer Science. Springer, 2021. doi: 10.1007/978-3-662-63958-0_26.

New York State Office of the Attorney General. Attorney general James issues report detailing millions of fake comments, revealing secret campaign to influence FCC's 2017 repeal of net neutrality rules. https://ag.ny.gov/press-release/2021/attorney-general-james-issues-report-detailing-millions-fake-comments-revealing, 2021.

Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–297, 1982. doi: 10.2307/2526438.

Ofcom. Clearing up the myths around 5G and the coronavirus. https://www.ofcom.org.uk/spectrum/electromagnetic-fields/clearing-up-myths-5g-and-coronavirus, 2020.

Dilek Önkal, M. Sinan Gönül, Paul Goodwin, Mary Thomson, and Esra Öz. Evaluating expert advice in forecasting: Users' reactions to presumed vs. experienced credibility. *International Journal of Forecasting*, 33(1):280–297, 2017. doi: 10.1016/j.ijforecast.2015.12.009.

OpenAI. GPT-4 technical report. arXiv:2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

Abraham Othman and Tuomas Sandholm. Decision rules and decision markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, pages 625–632, 2010.

Marco Ottaviani and Peter Norman Sørensen. The strategy of professional forecasting. *Journal of Financial Economics*, 81(2):441–466, 2006. doi: 10.1016/j.jfineco.2005.08.002.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? ICLR 2024; arXiv:2309.05196, 2024. URL https://arxiv.org/abs/2309.05196.

Joshua Page. Prison officer unions and the perpetuation of the penal status quo. *Criminology & Public Policy*, 10(3):735–770, 2011. doi: 10.1111/j.1745-9133.2011.00737.x.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. arXiv:2506.00723, 2025. URL https://arxiv.org/abs/2506.00723.

Georgios Papasotiropoulos and Ulrike Schmidt-Kraepelin. Liquid democracy: An annotated reading list. *ACM SIGecom Exchanges*, 22(2):130–135, 2025. doi: 10.1145/3821540.3821549.

Sunoo Park, Michael Specter, Neha Narula, and Ronald L. Rivest. Going from bad to worse: From Internet voting to blockchain voting. *Journal of Cybersecurity*, 7(1):tyaa025, 2021. doi: 10.1093/cybsec/tyaa025.

Alois Paulin. An overview of ten years of liquid democracy research. In *Proceedings of the 21st International Conference on Digital Government Research (DGO)*, 2020.

Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019. doi: 10.1016/j.cognition.2018.06.011.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, 2021. doi: 10.1038/s41586-021-03344-2.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020.

Michael Bang Petersen, Rune Slothuus, Rune Stubager, and Lise Togeby. Deservingness versus values in public opinion on welfare: The automaticity of the deservingness heuristic. *European Journal of Political Research*, 50(1):24–52, 2011. doi: 10.1111/j.1475-6765.2010.01923.x.

Pew Research Center. How the U.S. public and AI experts view artificial intelligence. Technical report, Pew Research Center, 2025. URL https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/.

Eric Pickles. Securing the ballot: Report of sir eric pickles' review into electoral fraud. HM Government, London, https://assets.publishing.service.gov.uk/media/5a8058bd40f0b62305b8a9d3/eric_pickles_report_electoral_fraud.pdf, August 2016.

Hanna F. Pitkin. *The Concept of Representation*. University of California Press, Berkeley, 1967.

Dražen Prelec. A Bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004. doi: 10.1126/science.1102081.

Dražen Prelec, H. Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017. doi: 10.1038/nature21054.

Paul C. Price and Eric R. Stone. Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, 17(1):39–57, 2004. doi: 10.1002/bdm.460.

Gabriel Recchia, Alexandra L. J. Freeman, and David Spiegelhalter. How well did experts and laypeople forecast the size of the COVID-19 pandemic? *PLOS ONE*, 16(5):e0250935, 2021. doi: 10.1371/journal.pone.0250935.

Repubblica Italiana. Constitution of the Italian Republic, art. 75. English translation by the Constitutional Court of the Italian Republic, https://www.cortecostituzionale.it/documenti/download/pdf/Costituzione_italiana_english_version_ott2023.pdf, 1948.

Republika Slovenija. Constitution of the Republic of Slovenia, art. 90. https://www.varuh-rs.si/en/about-us/legal-framework/the-constitution-of-the-republic-of-slovenia, 1991.

Manon Revel, Daniel Halpern, Adam Berinsky, and Ali Jadbabaie. Liquid democracy in practice: An empirical analysis of its epistemic performance. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*, 2022. Non-archival.

Reynolds. Reynolds v. sims. 377 U.S. 533 (Supreme Court of the United States), 1964.

Julian V. Roberts, Jonathan Bild, Jose Pina-Sánchez, and Mike Hough. Public knowledge of sentencing practice and trends. Technical report, The Sentencing Academy, 2022.

Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34):eabo6254, 2022. doi: 10.1126/sciadv.abo6254.

Michael L. Ross. Does oil hinder democracy? *World Politics*, 53(3):325–361, 2001.

Jesse Rothstein. Measuring the impacts of teachers: Comment. *American Economic Review*, 107 (6):1656–1684, 2017. doi: 10.1257/aer.20141440.

Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006. doi: 10.1126/science.1121066.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 2025. doi: 10.1038/s41562-025-02194-6.

Paola Sapienza and Luigi Zingales. Economic experts versus average Americans. *American Economic Review*, 103(3):636–642, 2013. doi: 10.1257/aer.103.3.636.

Felix Schläpfer. Determinants of voter support for a five-year ban on the cultivation of genetically modified crops in Switzerland. *Journal of Agricultural Economics*, 59(3):421–435, 2008. doi: 10.1111/j.1477-9552.2008.00167.x.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10, 2024. doi: 10.1126/sciadv.adp1528.

Lloyd Shapley and Bernard Grofman. Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 43(3):329–343, 1984. doi: 10.1007/BF00118940.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, et al. Towards understanding sycophancy in language models. arXiv:2310.13548, 2023. URL https://arxiv.org/abs/2310.13548.

Irina G. Shevtsova. An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, 82(3):862–864, 2010. doi: 10.1134/S1064562410060062.

Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation (EC '16)*. ACM, 2016. doi: 10.1145/2940716.2940790.

Felix M. Simon and Sacha Altay. Don't panic (yet): Assessing the evidence and discourse around generative AI and elections. Knight First Amendment Institute at Columbia University, 2025. URL https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections.

Hans-Werner Sinn and Silke Uebelmesser. Pensions and the path to gerontocracy in Germany. *European Journal of Political Economy*, 19(1):153–158, 2003. doi: 10.1016/S0176-2680(02)00134-9.

Paolo Spada, Jonathan Mellon, Tiago Peixoto, and Fredrik M. Sjoberg. Effects of the Internet on participation: Study of a public policy referendum in Brazil. *Journal of Information Technology & Politics*, 13(3):187–207, 2016. doi: 10.1080/19331681.2016.1162250.

Drew Springall, Travis Finkenauer, Zakir Durumeric, Jason Kitcat, Harri Hursti, Margaret MacAlpine, and J. Alex Halderman. Security analysis of the Estonian Internet voting system. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*, pages 703–715, 2014. doi: 10.1145/2660267.2660315.

Balaji S. Srinivasan and Leland Lee. Quantifying decentralization. 21.co blog, https://news.earn.com/quantifying-decentralization-e39db233c28e, July 2017.

George J. Stigler. The theory of economic regulation. *Bell Journal of Economics and Management Science*, 2(1):3–21, 1971.

John Strang, Teodora Groshkova, and Nicola Metrebian. New heroin-assisted treatment: Recent evidence and current practices of supervised injectable heroin treatment in Europe and beyond. Technical report, European Monitoring Centre for Drugs and Drug Addiction, 2012.

Marilyn Strathern. 'Improving ratings': Audit in the British University system. *European Review*, 5(3):305–321, 1997.

G. Supran, S. Rahmstorf, and N. Oreskes. Assessing ExxonMobil's global warming projections. *Science*, 379(6628):eabk0063, 2023. doi: 10.1126/science.abk0063.

Swiss Confederation. Federal constitution of the Swiss Confederation, art. 139(3). English translation, https://www.constituteproject.org/constitution/Switzerland_2014, 1999.

Swiss Federal Council. Popular initiative for a 13th OASI pension payment: Vote of 3 march 2024. https://www.admin.ch/gov/en/start/documentation/votes/20240303/iniziativa-per-una-13esima-mensilita-avs.html, 2024.

swissinfo.ch. Strong turnout influenced rejection of key environmental law. https://www.swissinfo.ch/eng/politics/strong-turnout-influenced-rejection-of-key-environmental-law/46829490, 2021.

swissinfo.ch. Five lessons from Swiss voters' acceptance of e-ID. https://www.swissinfo.ch/eng/swiss-politics/five-lessons-from-swiss-voters-acceptance-of-e-id/90082224, 2025.

Swissvotes. Federal decree on the medical prescription of heroin (13 june 1999). https://swissvotes.ch/votes?policy_area=10.101.1014, 1999.

Swissvotes. Vote no. 584.00: Federal act on the fund for the procurement of the Gripen fighter aircraft (18 may 2014). https://swissvotes.ch/vote/584.00, 2014.

Swissvotes. Vote no. 635.00: Federal decree on the procurement of new fighter aircraft (27 september 2020). https://swissvotes.ch/vote/635.00, 2020.

Elizabeth R. Tenney, Robert J. MacCoun, Barbara A. Spellman, and Reid Hastie. Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, 18(1):46–50, 2007. doi: 10.1111/j.1467-9280.2007.01847.x.

Philip E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton, NJ, 2005.

Philip E. Tetlock, Barbara A. Mellers, and J. Peter Scoblic. Bringing probability judgments into policy debates via forecasting tournaments. *Science*, 355(6324):481–483, 2017. doi: 10.1126/science.aal3147.

Philip E. Tetlock, Christopher Karvetski, Ville A. Satopää, and Kevin Chen. Long-range subjective-probability forecasts of slow-motion variables in world politics: Exploring limits on expert judgment. *Futures & Foresight Science*, 6(1):e157, 2024. doi: 10.1002/ffo2.157.

Abigail Thompson. Does diversity trump ability? An example of the misuse of mathematics in the social sciences. *Notices of the American Mathematical Society*, 61(9):1024–1030, 2014. doi: 10.1090/noti1163.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv:2305.14975, 2023. URL https://arxiv.org/abs/2305.14975.

Michael Tomz. Domestic audience costs in international relations: An experimental approach. *International Organization*, 61(4):821–840, 2007.

Transparency International Defence & Security. Government defence integrity index 2020. Technical report, Transparency International, 2021. URL https://ti-defence.org/government-defence-integrity-index-2020-corrption-index/.

Bao Tran Truong, Siqi Wu, Alessandro Flammini, Filippo Menczer, and Alexander J. Stewart. Community notes are vulnerable to rater bias and manipulation. arXiv:2511.02615, 2025.

United Nations. Convention relating to the status of refugees. Geneva, 28 July 1951, Art. 33, 1951.

U.S. Senate Select Committee on Intelligence. Report on the U.S. intelligence community's prewar intelligence assessments on Iraq. Technical report, United States Senate, 2004. 108th Congress, 2nd Session.

Markus Utke and Ulrike Schmidt-Kraepelin. Anonymous and copy-robust delegations for liquid democracy. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. URL https://openreview.net/forum?id=bbL2OOupi4.

Wim van Oorschot. Making the difference in social Europe: Deservingness perceptions among citizens of European welfare states. *Journal of European Social Policy*, 16(1):23–42, 2006. doi: 10.1177/0958928706059829.

Carlos Vargas-Silva. Seven years after the Eastern European enlargement. COMPAS, University of Oxford, https://www.compas.ox.ac.uk/article/seven-years-after-the-eastern-european-enlargement, September 2011.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. doi: 10.1126/science.aap9559.

Sze Yuh Nina Wang, Samantha C. Phillips, Kathleen M. Carley, Hause Lin, and Gordon Pennycook. Limited effectiveness of psychological inoculation against misinformation in a social media feed. *PNAS Nexus*, 4(6):pgaf172, 2025. doi: 10.1093/pnasnexus/pgaf172.

Rupert Way, Matthew C. Ives, Penny Mealy, and J. Doyne Farmer. Empirically grounded technology forecasts and the energy transition. *Joule*, 6(9):2057–2082, 2022. doi: 10.1016/j.joule.2022.08.009.

Hannah Werner and Sofie Marien. Process vs. outcome? how to evaluate the effects of participatory processes on legitimacy perceptions. *British Journal of Political Science*, 52(1):429–436, 2022. doi: 10.1017/S0007123420000459.

Jens Witkowski, Rupert Freeman, Jennifer Wortman Vaughan, David M. Pennock, and Andreas Krause. Incentive-compatible forecasting competitions. *Management Science*, 69(3):1354–1374, 2023. doi: 10.1287/mnsc.2022.4410.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2): 107–126, 2004. doi: 10.1257/0895330041371321.

Olivier J. Wouters. Lobbying expenditures and campaign contributions by the pharmaceutical and health product industry in the United States, 1999–2018. *JAMA Internal Medicine*, 180(5): 688–697, 2020. doi: 10.1001/jamainternmed.2020.0146.

Yale Budget Lab. Evaluating the impact of AI on the labor market: Current state of affairs. https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs, 2025.

Ilan Yaniv and Eli Kleinberger. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2):260–281, 2000. doi: 10.1006/obhd.2000.2909.

Büşra Elif Yelbuz, Ecesu Madan, and Sinan Alper. Reflective thinking predicts lower conspiracy beliefs: A meta-analysis. *Judgment and Decision Making*, 17(4):720–744, 2022. doi: 10.1017/S1930297500008913.

Yuzhe Zhang and Davide Grossi. Power in liquid democracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5822–5830, 2021. doi: 10.1609/aaai.v35i6.16729.