

A latent dimension of Condorcet’s jury theorem for multiple AI advisers

Kazutoshi Sasahara¹, Aoi Naito^{1,2}, and Ryo Fujie¹

¹Institute of Science Tokyo, Tokyo, Japan

²Carnegie Mellon University, Pittsburgh, PA, USA

Contact: sasahara.k.6efa@m.isct.ac.jp

ABSTRACT

When the same question is asked of multiple AI advisers, as in self-consistency and LLM-as-a-judge panels, Condorcet’s jury theorem predicts that adding independent, competent advisers makes the majority more reliable. The theorem, however, has a latent dimension when viewed from the user’s vantage: adding advisers also makes disagreement more visible. A binomial model reveals that this “visible dissent” becomes nearly inevitable as the number of advisers grows, and that reliability and disagreement both approach certainty but at different convergence rates. The two rates cross at an adviser accuracy of $4/5$ (0.8). Below this value, visible dissent approaches certainty faster than reliability and, with enough advisers, becomes more likely than a correct majority. Even ideal panels of independent and competent advisers can be correct in aggregate but appear divided; such disagreement does not by itself indicate aggregation failure. The way advisers split also provides a common basis for predictive multiplicity, reconciliation load, and reliance miscalibration. These results indicate two distinct decisions when using multiple AI advisers: how many advisers to consult and how their verdicts should be presented and interpreted.

Introduction

To decide what to believe, people rely on experts, institutions, and other sources of guidance, which serve as “trust anchors”. Yet that role is shifting to AI advisers,¹ especially for questions that people cannot reliably assess on their own. However, no AI model is perfectly accurate, and its answers are often claims that people cannot directly verify.^{2,3} In such cases, people consult a second model and then a third, or query the same model repeatedly, in effect convening a panel of AI advisers. For example, ask several diagnostic systems about one medical scan, and one may answer “nothing concerning”, while another raises the possibility of cancer.⁴ Although asking more advisers can make the majority more reliable, it can also make their disagreements harder for the user to ignore.

Asking more is Condorcet’s jury theorem at work.⁵ In its modern form,^{6–9} if K independent advisers are each correct with probability greater than chance, the aggregate reliability (P_{agg}) approaches certainty as the panel grows (Supplementary Note 1). The same logic underlies the “wisdom of crowds”^{10,11} and, two centuries on, systems that aggregate multiple AI advisers. Examples include self-consistency, which relies on a majority vote over repeated samples from a single model;¹² LLM-as-a-judge, in which a strong model evaluates a candidate response;¹³ and multi-model aggregation.^{14,15}

The theorem, however, is usually read from the system’s vantage (Fig. 1, left). It scores the majority against the ground truth, collapsing the panel to a single bit (the majority position) and discarding the vote distribution the user observes. That distribution is a second observable. Asking more advisers therefore creates a new kind of uncertainty that no single AI has. This is one form of what we call the “multi-trust-anchor problem”, the new uncertainties that arise once trust is distributed across many anchors. The starting point is that improving the reliability of the majority does not, by itself, settle how multiple advisers should be used. Even under ideal conditions, adding advisers changes the disagreement users encounter, raising a separate question about how their verdicts should be presented and interpreted.

Here we revisit the theorem from the user’s rather than the system’s vantage. Unlike the system, the user does not observe the ground truth at the time of consultation; only the advisers’ verdicts are available. This epistemic shift makes another quantity relevant: the probability of “visible dissent” (P_{vis}), or how often the advisers appear to split. As the panel grows, both aggregate reliability and visible dissent approach one. But they do not converge at the same rate. The two convergence rates cross at a particular adviser accuracy. This crossing provides a quantitative reference for studying how systems should present multiple advisers’ verdicts.

To show that the disclosure problem arises even under conditions favorable to aggregation, we begin with the following base model. Each question has two possible verdicts, with one of them correct, so the analysis applies to questions with a ground truth. This still covers many high-stakes settings, such as medical screening and fact-checking, where the decision is effectively binary. We consider K independent advisers, each answering a given question correctly with the same probability p ,

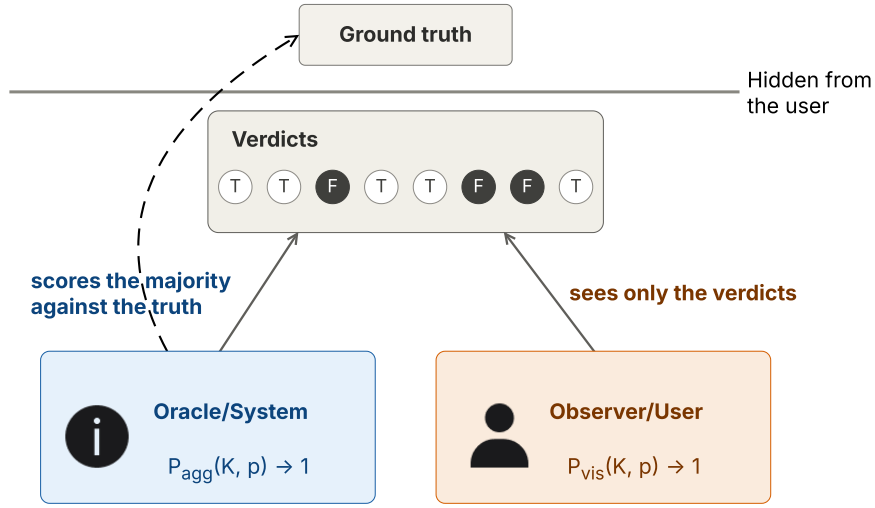


Figure 1. Two observables from the same verdicts. The same verdicts are read from two vantages. The system (left), an oracle with the ground truth, scores the majority against it. The user (right) sees only the verdicts. As the panel grows, aggregate reliability approaches certainty ($P_{\text{agg}} \rightarrow 1$), while the user increasingly sees a split ($P_{\text{vis}} \rightarrow 1$).

$1/2 < p < 1$, and otherwise free of incompetence and bias. The equal-accuracy assumption here is made for clarity and is not required by the theorem. In other words, the model starts from advisers close to the ideal that efforts toward trustworthy AI seek to achieve. Any disagreement that appears therefore need not be a sign of failure: it can arise even among ideal advisers.

A latent dimension: what the user observes

From the user’s vantage (Fig. 1, right), the question is how often the panel looks divided. We focus on the minimal split, with at least two dissenters on each side. Two rationales support this choice. Psychologically, even an ideal panel will often include a lone dissenter, since each adviser commits errors at a rate of $q (= 1 - p)$. The user can dismiss one dissenting vote as random noise. Research on social influence in human groups supports this distinction. A unanimous majority differs qualitatively from one with a single defector,¹⁶ and a minority is perceived as a legitimate position only when it is consistent,^{17,18} and when at least one other member shares it.¹⁹ This distinction has recently been observed with panels of AI advisers as well.²⁰ Mathematically, having two dissenters on each side represents the smallest split that cannot be attributed solely to the error of a single adviser (Supplementary Note 2). This “two” threshold is an operational definition of visible dissent, not a claim that lone dissent never matters.

We use this minimal split throughout; in the generic form, the count of two is replaced by a fixed fraction r of the panel on each side (Supplementary Notes 3 and 4). The complement of this event is the four near-unanimous configurations: zero, one, all but one, or all advisers on one side. Under this definition, the probability of visible dissent is expressed in closed form as

$$P_{\text{vis}}(K, p) = 1 - q^{K-1}(Kp + q) - p^{K-1}(Kq + p), \quad q = 1 - p,$$

which approaches 1 for all $p < 1$; equivalently, near-unanimous panels become rare. Both facts follow from the binomial distribution alone (Supplementary Note 2). The closed form of visible dissent is symmetric in p and q : a split indicates that the panel disagrees, but the event itself is side-blind.

Both aggregate reliability and visible dissent increase toward certainty. Which dominates and for how long depends on adviser accuracy. At $p = 0.8$, a split never overtakes a correct majority (Fig. 2a); at lower accuracies, it does and at smaller panels (Fig. 2b,c). Even panels with near-perfect accuracy routinely exhibit splits. This structural decoupling across panel sizes and accuracies is captured by P_{vis} (Fig. 2d). At $p = 0.9$ and $K = 20$, the majority is almost certain to be correct (above 99.999%), yet the user sees a split about 61% of the time. Frequent agreement on easy questions does not contradict this. Here p is adviser accuracy on a specific item, and competent advisers are nearly always right on easy ones. For any fixed imperfect accuracy in the base model, splits approach certainty as the panel grows, although higher accuracy delays this increase. Users

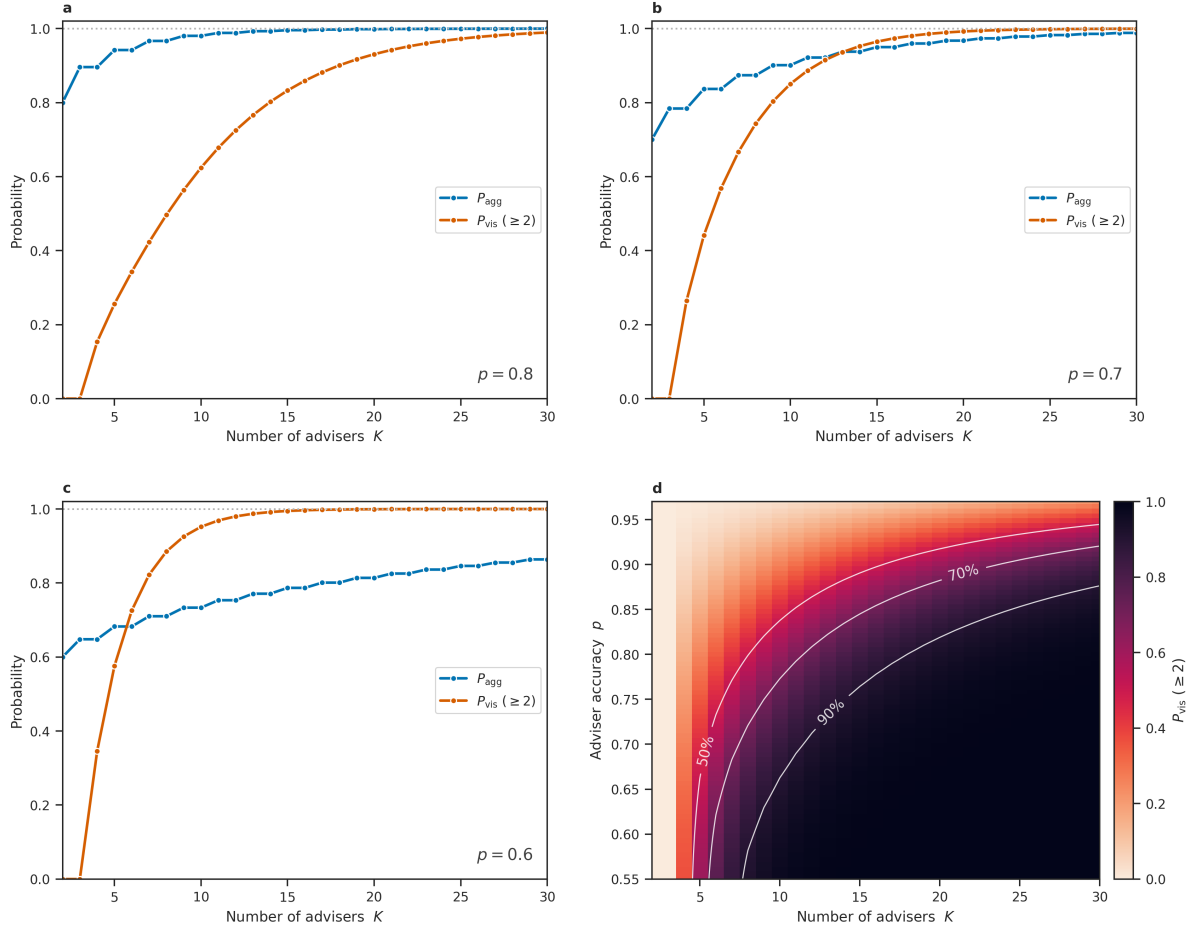


Figure 2. Disagreement is inevitable and, below $p = 4/5$, eventually overtakes reliability. K independent advisers each correct with probability p , $1/2 < p < 1$. The user sees only the verdicts. A split counts at least two dissenters on each side. (a–c) aggregate reliability P_{agg} (fair tie-breaking) and visible dissent P_{vis} both tend to 1 yet decouple, with the first crossover moving with accuracy. At $p = 0.8$ (a), reliability stays ahead and the user’s view never overtakes; at $p = 0.7$ (b) it overtakes at $K^* = 14$; and at $p = 0.6$ (c) already at $K^* = 6$. (d) $P_{\text{vis}}(K, p)$ across the plane with contours. A split is common once the panel is even moderately large (the 50% contour) and persists at high accuracy.

can also observe the dissenting fraction (m). As the panel grows, this fraction approaches q even when the majority is correct. We write this structural baseline as $m \rightarrow q$ (Supplementary Note 3).

When visible dissent outpaces reliability

Which approaches certainty faster, aggregate reliability or visible dissent? Both approach it exponentially fast as K grows, and the two convergence rates can be compared directly. The former is the classical error exponent of majority voting, and the latter the standard large-deviation rate at which near-unanimous panels become rare²¹ (Supplementary Note 4). This comparison shows that the two rates cross at a single adviser accuracy $p = 4/5$ (0.8, or odds of four to one). Above this line, aggregate reliability converges faster and stays ahead; below it, the order reverses, and visible dissent converges faster (Fig. 3a). Figure 2a sits exactly on this line ($p = 0.8$), where reliability keeps its lead at every finite panel size; the lower-accuracy panels of Fig. 2b,c fall below it ($p = 0.7$ and $p = 0.6$, respectively). This crossing of two rates is robust. Adjusting the fixed-count split size affects only the panel size at which the crossover occurs. (Across fraction thresholds in the generic form, the rate boundary remains below $p = 4/5$ for every fixed positive fraction and approaches $p = 4/5$ as the fraction shrinks to zero; Supplementary Fig. 1.) No split definition that excludes unanimity places its rate boundary above $4/5$; the minimal split counted here attains that bound. Neither fixed residual weighting nor the equivalent formulation in terms of disjoint outcomes (Supplementary Note 4) moves this upper bound. The $4/5$ bound is thus intrinsic to this class of split events.

Below the $4/5$ line the probability ordering of the two quantities can reverse, and two effects must be kept apart. Visible

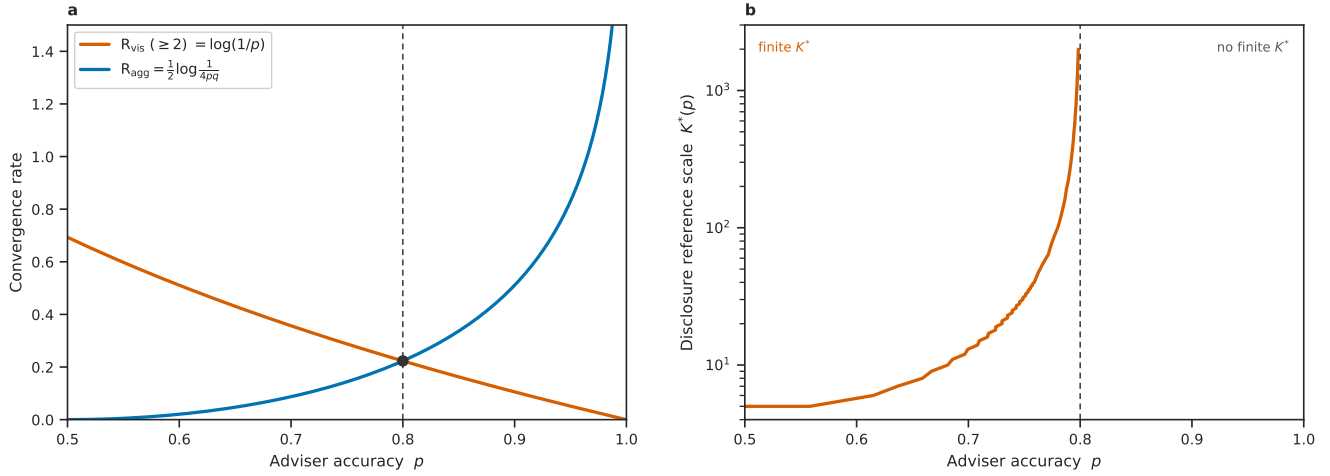


Figure 3. The two convergence rates cross at $p = 4/5$, fixing a disclosure reference scale. (a) The crossing fixes whether a finite disclosure reference scale K^* exists (yes below the line, no above). **(b)** $K^*(p)$, the panel size at which the dissent–reliability crossover occurs (a split becomes at least as likely as a correct majority), diverging as $p \rightarrow 4/5$ (log scale). $K^* = 6, 14, 32$ at $p = 0.6, 0.7, 0.75$ under fair tie-breaking; the values shift under a strict-majority convention (Supplementary Note 5). The dashed line separates the two regimes (finite K^* to its left, none to its right). Benchmark averages are not the per-item p and do not locate a panel on this curve (Supplementary Note 6). The rate boundary and the first crossover are properties of the base model’s probabilities; the effectiveness of the presentation policies compared in Box 1 remains to be tested.

dissent rises with K at every accuracy. The second effect is the reversal of the probability ordering, $P_{vis} \geq P_{agg}$, which is a statement about the base model rather than a claim of net behavioural harm (Supplementary Note 4). At or above the line, a correct majority stays more likely than a split, no matter how large the panel grows. We call the value $4/5$ the “rate-crossover boundary” of the base model.

We use “dissent–reliability crossover” for the first panel size at which visible dissent becomes at least as likely as a correct majority (Fig. 3b); this size defines the “disclosure reference scale” $K^*(p)$. At this first crossover, panels in repeated consultations are divided at least as often as their majority is correct—and a correct majority can still come from a divided panel. The rate-crossover boundary at $p = 4/5$ determines whether such a finite size exists. K^* anchors comparisons of presentation policies before and after the first probability crossover. The probability ordering can depend on parity just above this size, so comparisons for a given panel should use the probabilities at that size (Supplementary Note 5). For every accuracy below $4/5$, visible dissent eventually remains more likely than a correct majority (Supplementary Note 4). For $p \geq 4/5$, an analytic proof rules out any crossover (Supplementary Note 5). Near the line, K^* grows rapidly, reaching hundreds of advisers, with both probabilities extremely close to one; at $p = 0.6$ and 0.7 , by contrast, K^* is 6 and 14.

The design question therefore arises even before independence or competence fails: how should a system present verdicts that become more reliable in aggregate yet more likely to appear divided? The rate boundary and disclosure reference scale provide quantitative reference points for investigating this question. We next connect these results to prior work, human responses to disagreement, and dependence among advisers.

Two vantages on the same disagreement

The disagreement itself is not new. In predictive multiplicity, “ambiguity” refers to the fraction of inputs on which models of comparable accuracy disagree.^{22–24} Earlier, Poisson and, later, Gelfand and Solomon examined how verdicts split across a panel.^{25,26} These lines of work ask different questions about agreement and disagreement (Table 1). A closely related approach conditions on the realized vote to infer competence and truth,²⁷ treating the observed margin as evidence rather than asking how often dissent will be seen at all. Our comparison is different: we take the probability that the user encounters a split as the object and compare its convergence with that of aggregate reliability. Under independence, its probability has an exact closed form, which leads to the $4/5$ crossover. This is also the setting in which the advice to “ask more” is on its strongest footing. The framework therefore quantifies the dissent a user encounters when following that advice.

Our contribution is to make the user’s exposure to disagreement an explicit quantity in the classical jury model and compare its convergence with that of majority reliability. Its practical value lies in two reference quantities: the adviser accuracy

Table 1. Two vantages on disagreement among advisers. View identifies the analytical perspective emphasized here: evaluating collective performance (Oracle/System) or interpreting the verdicts available to an observer (Observer/User). These perspectives can overlap within a study. This work examines exposure to disagreement under the classical ideal assumptions, comparing visible-dissent probability with majority reliability.

View	Focus	Question	Representative work
Oracle/System	Dependence and diversity	What shapes the reliability of collective judgements?	Common-cause failure ²⁸ ; algorithmic monoculture ²⁹ ; correlation neglect ³⁰ ; ensemble diversity ^{31,32} ; interpreted versus generated signals ³³ ; pooling medical judgments ³⁴
Oracle/System and Observer/User	Vote distributions and multiplicity	How do verdicts differ across advisers and models?	Jury verdict-split models ^{25,26} ; ensemble diversity measures ³⁵ ; predictive multiplicity ²² ; positive dissensus ³⁶
Observer/User	Uncertainty and inference from votes	What can agreement patterns tell us about uncertainty, competence, or truth?	Deep ensembles ³⁷ ; disagreement-based UQ ³⁸ ; abstention on arbitrary predictions ³⁹ ; no-consensus-probability voting ⁴⁰ ; confidence-calibrated jury ⁴¹ ; posterior updating from jury size and margin ²⁷ ; algebraic evaluation from answer patterns ⁴² ; consensus-accuracy laws across models ⁴³ ; unanimity as evidence of systematic failure ⁴⁴ ; decisions with misleading correlated signals ⁴⁵
Observer/User	Visible dissent under the classical ideal assumptions (this work)	How often will users encounter a split as majority reliability improves?	Visible-dissent probability of independent, equal-accuracy, better-than-chance advisers compared with majority reliability; rate-crossover boundary $p = 4/5$; disclosure reference scale K^*

($p = 4/5$) below which visible dissent eventually surpasses aggregate reliability, and the corresponding disclosure reference scale (K^*). The result is not specific to AI: it applies to any collection of independent, imperfect advisers with the same accuracy, including jurors, sensors, and models. AI is unusual in that users can assemble such panels on demand. To our knowledge, this comparison of the convergence of visible dissent and majority reliability, together with its use as a reference for disclosure design, has not been developed in prior work.

Three perspectives on the same split

Why not simply ignore the split and act on the majority? The majority remains the best prediction, but this does not make the split irrelevant. Three separate literatures approach observed disagreement from different angles. All three begin from the same information available to the user: the vote split (Fig. 4). Predictive multiplicity—mentioned above—reads it as the probability of observing any disagreement. Cognitive load—the cost of reconciling conflicting advice^{20,46}—is linked here to the dissenting fraction (m). Reliance miscalibration—either under-reliance or over-trust in the model panel^{47–51}—is influenced both by exposure to disagreement and by “false consensus”, the all-wrong core (q^K) of apparent unanimity ($p^K + q^K$). While each literature has measured its quantity after the fact, the base model gives each a predictive counterpart in advance; testing these predictions is the experimental agenda this shift opens (Supplementary Note 7).

The predictive-multiplicity reading is exact under the base model’s assumptions. For the other two interpretations, the base model specifies the stimulus available to the user, while the behavioural response remains outside the base model. To fill this gap, we assume that each behavioural response does not decrease as its corresponding stimulus increases. We call this the “monotone-response hypothesis” and leave its empirical test to future work (Supplementary Note 7). Under this hypothesis, expected reconciliation load does not decrease with greater dissent. Exposure to a split may also increase under-reliance, since many splits occur despite a correct majority and may be interpreted as evidence of uncertainty. In the base model, the observed margin is itself evidence. At $p = 0.7$ and $K = 14$, majority accuracy is 0.938 before observing the vote, but 0.845 conditional on an 8–6 split. Lowering trust after a narrow split can therefore be a calibrated response rather than under-reliance (Supplementary Note 7). Over-trust has a different source. An all-wrong panel occurs with probability q^K under independence, and presents the same unanimous vote as an all-correct panel. The base model also fixes how these stimuli order with panel size and accuracy, which experiments can test (Supplementary Note 7). All three considerations can therefore be placed on the same panel scale, with K^* marking the first panel size at which a split becomes at least as likely as a correct majority.

The stimuli for reconciliation load and under-reliance are already substantial at this scale, whereas the all-wrong probability q^K has become negligible under independence. Pairwise disagreement among frontier models of comparable accuracy has been measured at 16–38% on multiple-choice benchmarks.⁵² These rates cannot be converted into per-item accuracy using the binary model. Multiple-choice disagreement includes pairs that give different wrong answers. Models sharing training data or

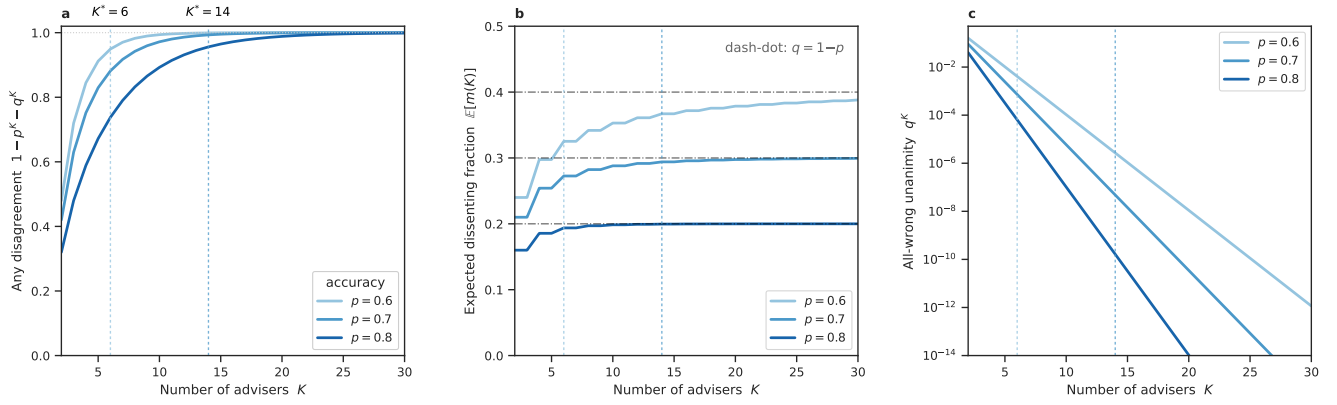


Figure 4. A common predictive substrate. All three user-side quantities are computed from the same vote split Y . **(a)** The probability of any disagreement, $1 - p^K - q^K$, which predictive multiplicity reads as the rate of conflicting predictions (Supplementary Note 7). **(b)** The expected dissenting fraction $\mathbb{E}[m(K)]$, with $m(K) = \min(Y, K - Y)/K$, considered here as a candidate stimulus for reconciliation load, converges to q . **(c)** The all-wrong core q^K of apparent unanimity, the modelled stimulus for over-trust (log scale); under independence it decreases exponentially with panel size; correlation can substantially inflate it (Supplementary Note 6). All three are explicit, testable functions of panel size and accuracy. The disclosure framework and the 4/5 line are stated on the split event P_{vis} (Figs. 2 and 3); the complements $1 - P_{\text{vis}}$ and $p^K + q^K$ vanish at the same rate, so rate-level statements do not depend on the reading of disagreement (Supplementary Note 7). Dotted vertical lines mark the disclosure reference scale $K^*(p)$ ($K^* = 6, 14$ for $p = 0.6, 0.7$; none exists for $p = 0.8$), shown for reference: at K^* the load and under-reliance stimuli are already substantial. The predictive-multiplicity reading in (a) is exact within the base model; the links to reconciliation load and reliance rest on the monotone-response hypothesis (Supplementary Note 7).

lineage also tend to make similar mistakes,^{53,54} so measured disagreement reflects the answer space and dependence as well as accuracy (Supplementary Note 7).

A null model for real panels, an exact law for independent machine-internal sampling

Real AI advisers are rarely independent or equally accurate. Even so, the base model provides an independence baseline: the distribution of vote splits an equal-competence panel would show if its members voted independently. A real panel can fall short of this baseline for reasons unrelated to dependence, such as easy questions or unequal accuracies. After adjusting for these factors, a remaining deviation that exceeds estimation and sampling error provides evidence of residual dependence; a shortfall of this kind we call the “dissent deficit” (Supplementary Fig. 2a). Extensions of the theorem to correlated votes show that dependence alters aggregate reliability and can erode the gains that independence provides.^{55,56} In the correlated panels we simulate, it also lowers the split rate, although this does not follow from positive pairwise correlation alone (one-factor Gaussian copula; Supplementary Fig. 2). This independence baseline can also help calibrate methods that use disagreement as a signal of uncertainty.³⁸

Positive dependence can also make agreement less reassuring. Advisers that make similar errors can produce a unanimous panel that is simply wrong, and such all-wrong panels can arise more often than the q^K predicted under independence (equal-competence case; Supplementary Fig. 2b). No adviser need have malfunctioned; individually reliable advisers can simply fail together. This is false consensus. Related work on the “paradox of unanimity” shows how allowing for systematic failure can make overwhelming agreement less reassuring.⁴⁴ Experiments with state-dependent signal correlation further show that both individuals and small groups struggle to use potentially misleading evidence, often treating correlated signals as independent.⁴⁵

The base model plays a different role in machine-internal sampling. A system that generates and aggregates multiple outputs can observe how they split without making any assumption about human responses. The observer need not be human: such a system stands at the same vantage, seeing the split but not the truth. The clearest case is repeated sampling from a fixed model. When the same item is sampled repeatedly under the same generation settings, the outputs are i.i.d. with probability p of being correct, and K denotes the number of samples drawn. Here the closed form for $P_{\text{vis}}(K, p)$ is not a baseline but an exact law. Changes in prompts, model versions, or serving conditions can cause the samples to no longer be i.i.d., taking the system outside the base model.

For $p < 4/5$, sampling beyond K^* can still improve aggregate reliability; K^* does not say when to stop sampling. This scale provides a reference for comparing how the samples are reconciled or presented.

Box 1 | A disclosure framework

Base model. K independent advisers, each correct with probability p , $1/2 < p < 1$, on a binary question; the majority is taken with fair tie-breaking, and a split means at least two votes on each side.

What the model gives. For a given p , the model gives the probability of encountering a split, $P_{\text{vis}}(K, p)$, and the expected dissenting fraction, $\mathbb{E}[m(K)]$. It also gives the disclosure reference scale $K^*(p)$: the first panel size at which a split is at least as likely as a correct majority, finite only for $p < 4/5$.

From model to comparison. K^* anchors comparisons of presentation policies around the first probability crossover. The same policies are tested under both orderings below; the model does not determine which policy is best.

Policies to compare. Individual verdicts; verdicts with the vote margin and estimated confidence; an openly reconciled answer with estimated confidence and access to the full panel. Each policy explains the split expected under the base model.

A panel can be both correct and split; these are overlapping events.

Condition at the actual K	What to test
Correct majority more likely: $P_{\text{agg}} > P_{\text{vis}}$	How each policy affects judgement, reliance, and reconciliation load
Split at least as likely as a correct majority: $P_{\text{vis}} \geq P_{\text{agg}}$	Whether relative policy effectiveness differs between the two orderings

How to compare. For $p < 4/5$, include several panel sizes around K^* , accounting for panel size and the observed vote margin. Assign each size by its actual probability ordering, not simply by whether it exceeds K^* .

Conditions of use. If values within the uncertainty interval for p yield different rows at the actual K , leave the ordering unassigned. Allow for item difficulty and for unequal or dependent advisers. Apply only to questions with a ground truth. Keep safety-critical minority signals visible whatever the ordering.

The closed form does not require the model to be better than chance. On an item where p falls below $1/2$, the samples can remain i.i.d. even as majority voting becomes increasingly likely to select the wrong answer. The split cannot warn of this because its distribution is symmetric in p and q . If accuracy stops improving as more samples are drawn, the model may have reached a competence ceiling, or the samples may no longer be independent. After adjusting for item difficulty and unequal accuracies, a lower-than-expected split rate indicates the latter. In this setting, the framework helps diagnose whether further sampling adds independent information and whether reconciliation may be useful.

A framework for the disclosure layer

What should the system show to the user, and when? Even in the best case, where every adviser is reliable, well calibrated, and independent, two numbers need attention. One is the dissenting fraction expected by default, the structural baseline $m \rightarrow q$. The other is the disclosure reference scale K^* . Adding advisers and showing their verdicts are distinct design decisions. Both are handled at the disclosure layer, in how the system combines and presents its outputs (Box 1).

Users convene these panels, but only the system that presents the verdicts can act on the split. The application of the framework depends on who receives the verdicts. When a machine collects them and takes a majority vote, as in self-consistency, the criterion yields $K^*(p)$ exactly. Pipelines that do not vote, such as best-of- N selection or LLM-as-a-judge scoring, fall under the criterion only after their outputs are converted into verdicts and put to a majority vote.

For human-facing use, a domain-level accuracy estimate can be inserted into the homogeneous model to obtain a reference scale. A domain average, however, is not equivalent to a mixture of easy and hard questions, because both probabilities are nonlinear in p . At $K = 14$, an equal mixture of items with $p = 0.51$ and $p = 0.89$ (mean 0.7) reverses the ordering that holds when every item has $p = 0.7$ (Supplementary Note 6). This K^* is therefore a domain-level guide rather than an item-level trigger (Supplementary Note 6). The criterion's operational use depends on the estimate of p (Box 1). The realized split carries item-level information at run time, although it does not identify item difficulty.

Two properties clarify the comparison of P_{vis} with P_{agg} . First, the split event is side-blind. Near-agreement, by contrast, points to a side, since a near-unanimous panel is correct with probability approaching one (Supplementary Note 2). Second,

a designer may think a split matters more than a correct majority, or less. Because the complementary residuals decay at different exponential rates, fixed residual weighting shifts the finite crossover but does not remove the asymptotic rate reversal (Supplementary Note 4).

A well-aligned model should tell the user what its verdicts show, not a correct answer it does not have.⁵⁷ This has two consequences. First, the system should explain the disagreement expected under the base model while retaining the information carried by the observed vote margin. Second, a split in a reliable panel should not automatically generate doubt. Studies with two advisers point the same way. A second opinion is helpful only when users are not shown disagreement too often on questions that the AI has answered correctly.⁵⁸

The framework's criterion is exact within the base model, while the relative effectiveness of the presentation policies remains to be tested (Box 1). Among the candidates in Box 1, an openly reconciled answer is accompanied by the expected dissenting fraction, estimated confidence, and access to the full panel. Under dependence, the independence baseline can serve as a diagnostic reference (Supplementary Note 6), and the disclosure layer remains the lever.

Discussion

The multi-trust-anchor problem begins even in Condorcet's ideal world. Adding independent, competent advisers improves majority reliability, but also makes disagreement more likely to be encountered. Such disagreement does not, by itself, indicate a failure of aggregation. The gain in accuracy therefore does not settle the design of the consultation: how the verdicts are presented and interpreted remains a separate question. The 4/5 rate boundary and disclosure reference scale K^* make this question concrete. The same model also supplies an independence baseline for studying real adviser panels. A practical first step is to record and evaluate the pre-aggregation vote distribution and split rate alongside majority accuracy. Box 1 uses these quantities to frame comparisons of presentation policies, distinguishing model results from hypotheses to test. Whether presentation should change at the crossover remains an empirical question. The distinction between aggregation and presentation remains relevant even if policy effectiveness shows no particular change around K^* . The disclosure reference scale K^* anchors comparisons of the same presentation policies across the two probability orderings (Box 1). Model accuracy is only part of the design problem; applying the framework also depends on how closely real adviser panels match the assumptions of the base model.

The assumptions of the base model point to the next steps, starting with independence. Social influence and collective exposure diminish it.^{59,60} In the correlated panels we simulate, dependence suppresses visible dissent below the independence baseline, although this is not guaranteed for every dependent panel (Supplementary Fig. 2). A dissent deficit is diagnostic of dependence, but its magnitude must be calibrated before it can be used to infer how strong that dependence is. Frequent model updates invalidate existing calibrations.

The competence assumption may also fail. When advisers are worse than chance, the majority becomes more confidently wrong as the panel grows. Where per-adviser accuracies are available, heterogeneity under independence is absorbed exactly by the Poisson-binomial baseline (Supplementary Note 4). A further extension is the connection between visible dissent and human behaviour; only the human-facing use of the disclosure framework depends on that connection (Supplementary Note 7). The binary setting is less restrictive than it seems, since many systems turn open-ended responses into true-or-false decisions.

A harder case is when advisers do not answer in parallel. When agents see and revise one another's answers, interaction itself can shape agreement. Human dyad experiments highlight the role of communicating confidence in joint perceptual decisions.⁶¹ Recent work on LLM naming games shows how mutual learning from sampled outputs can generate consensus without external evidence.⁶² Extending the present framework to such settings requires tracking how interaction changes both the verdict distribution and its interpretation.

The framework does not apply equally to all questions. Reconciliation is appropriate when a question has a ground truth. For instance, whether two drugs interact has a ground truth, whereas questions of policy fairness are typically normative rather than objectively settled. When a system handles millions of queries, treating a contested issue as settled can hide legitimate disagreement from users. Deciding whether a question falls within scope is itself a governance decision and should not be left to an automated classifier alone.

Visible dissent is only the first layer of the multi-trust-anchor problem. Once the assumptions of the base model are relaxed, even agreement becomes harder to interpret: users can see that advisers agree, but not whether they reached the same answer independently or failed together. The null model offers a first diagnostic, but more complex systems introduce further layers of trust. One AI may be asked to judge which other AI should be trusted, and in agentic systems several components may combine outputs before anything reaches a human user. At each layer, the system must decide which outputs to retain, how to reconcile them, and what to disclose. As trust becomes distributed across multiple AI advisers, the problem shifts from the reliability of individual outputs to how agreement and disagreement are combined, interpreted, and disclosed.

Declarations

Author contributions. K.S. conceived and designed the study, developed the theory and the analysis, and wrote the manuscript. A.N. contributed to the design of the study by connecting the framework to research on human–AI reliance, and to the interpretation of the user-side quantities. R.F. contributed to the design of the theoretical analysis and checked the derivations in the Supplementary Information. All authors reviewed and approved the manuscript.

Competing interests. The authors declare no competing interests.

Data availability. No empirical datasets were generated or analysed in this study. The numerical analyses reported here derive from the closed forms and from synthetic simulations described in the Supplementary Information.

Code availability. The analytical results follow from the derivations in the Supplementary Information; accompanying scripts reproduce the numerical calculations and simulation figures. The code is not yet publicly available; a repository link will be added in a subsequent version.

Use of AI tools. Generative-AI assistance was used for language editing. All technical content is the authors' own.

References

1. Logg, J. M., Minson, J. A. & Moore, D. A. Algorithm appreciation: people prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* **151**, 90–103 (2019).
2. Augenstein, I. *et al.* Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence* **6**, 852–863 (2024).
3. Suzgun, M. *et al.* Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence* **7**, 1780–1790 (2025).
4. van Winkel, S. L. *et al.* AI as an independent second reader in detection of clinically relevant breast cancers within a population-based screening programme in the Netherlands: a retrospective cohort study. *The Lancet Digital Health* **7**, 100882 (2025).
5. Condorcet, M. d. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* (Imprimerie Royale, 1785).
6. Grofman, B., Owen, G. & Feld, S. L. Thirteen theorems in search of the truth. *Theory and Decision* **15**, 261–278 (1983).
7. Nitzan, S. & Paroush, J. *Collective Decision Making: An Economic Outlook* (Cambridge University Press, 1985).
8. List, C. & Goodin, R. E. Epistemic democracy: generalizing the Condorcet jury theorem. *Journal of Political Philosophy* **9**, 277–306 (2001).
9. Goodin, R. E. & Spiekermann, K. *An Epistemic Theory of Democracy* (Oxford University Press, 2018).
10. Galton, F. Vox populi. *Nature* **75**, 450–451 (1907).
11. Surowiecki, J. *The Wisdom of Crowds* (Doubleday, 2004).
12. Wang, X. *et al.* Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)* (2023).
13. Zheng, L. *et al.* Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks* (2023). ArXiv:2306.05685.
14. Du, Y. *et al.* Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*, PMLR 235, 11733–11763 (2024).
15. Chen, J. C.-Y. *et al.* ReConcile: round-table conference improves reasoning via consensus among diverse LLMs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 7066–7085 (2024).
16. Asch, S. E. Effects of group pressure upon the modification and distortion of judgments. In Guetzkow, H. (ed.) *Groups, Leadership and Men: Research in Human Relations*, 177–190 (Carnegie Press, Pittsburgh, PA, 1951).
17. Moscovici, S., Lage, E. & Naffrechoux, M. Influence of a consistent minority on the responses of a majority in a color perception task. *Sociometry* **32**, 365–380 (1969).

18. Latané, B. & Wolf, S. The social impact of majorities and minorities. *Psychological Review* **88**, 438–453 (1981).
19. Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied* **70**, 1–70 (1956).
20. Tsuchiya, Y. & Baba, Y. More isn't always better: balancing decision accuracy and conformity pressures in multi-AI advice. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21 (ACM, 2026).
21. Dembo, A. & Zeitouni, O. *Large Deviations Techniques and Applications* (Springer, 1998), 2nd edn.
22. Marx, C., Calmon, F. & Ustun, B. Predictive multiplicity in classification. In *International Conference on Machine Learning (ICML)* (2020).
23. Watson-Daniels, J., Parkes, D. C. & Ustun, B. Predictive multiplicity in probabilistic classification. In *AAAI Conference on Artificial Intelligence*, vol. 37, 10306–10314 (2023).
24. Black, E., Raghavan, M. & Barocas, S. Model multiplicity: opportunities, concerns, and solutions. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 850–863 (2022).
25. Poisson, S. D. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile* (Bachelier, 1837).
26. Gelfand, A. E. & Solomon, H. A study of Poisson's models for jury verdicts in criminal and civil trials. *Journal of the American Statistical Association* **68**, 271–278 (1973).
27. Romeijn, J.-W. & Atkinson, D. Learning juror competence: a generalized Condorcet Jury Theorem. *Politics, Philosophy & Economics* **10**, 237–262 (2011).
28. Knight, J. C. & Leveson, N. G. An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering* **SE-12**, 96–109 (1986).
29. Kleinberg, J. & Raghavan, M. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* **118**, e2018340118 (2021).
30. Enke, B. & Zimmermann, F. Correlation neglect in belief formation. *Review of Economic Studies* **86**, 313–332 (2019).
31. Krogh, A. & Vedelsby, J. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7 (NIPS 1994)*, 231–238 (1995).
32. Kim, D. Are diversity metrics measuring diversity? A capability-controlled audit of majority-vote gain in LLM ensembles (2026). ArXiv:2607.20768.
33. Hong, L. & Page, S. E. Interpreted and generated signals. *Journal of Economic Theory* **144**, 2174–2196 (2009).
34. Kurvers, R. H. J. M. *et al.* Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences* **113**, 8777–8782 (2016).
35. Kuncheva, L. I. & Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* **51**, 181–207 (2003).
36. Landemore, H. & Page, S. E. Deliberation and disagreement: problem solving, prediction, and positive dissensus. *Politics, Philosophy & Economics* **14**, 229–254 (2015).
37. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017).
38. Jiang, B. DiscoUQ: structured disagreement analysis for uncertainty quantification in LLM agent ensembles (2026). ArXiv:2603.20975.
39. Cooper, A. F. *et al.* Arbitrariness and social prediction: the confounding role of variance in fair classification. In *AAAI Conference on Artificial Intelligence*, vol. 38, 22004–22012 (2024).
40. Nair-Kanneganti, A. *et al.* Increasing LLM response trustworthiness using voting ensembles (2025). ArXiv:2510.04048.

41. Karge, J. Epistemic filtering and collective hallucination: a jury theorem for confidence-calibrated agents (2026). ArXiv:2602.22413.
42. Corrada-Emmanuel, A. Algebraic evaluation theorems (2024). ArXiv:2412.16238.
43. Liu, N. LLMs as a jury: cross-model consensus can outperform process reward models for LLM reasoning (2026). ArXiv:2607.10139.
44. Gunn, L. J. *et al.* Too good to be true: when overwhelming evidence fails to convince. *Proceedings of the Royal Society A* **472**, 20150748 (2016).
45. Fehrler, S., Hochleitner, A. & Janas, M. Too good to be true: individual and collective decision-making with misleading signals. *Management Science* (2026). Published online May 4, 2026; doi:10.1287/mnsc.2024.06504.
46. Yaniv, I. & Milyavsky, M. Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes* **103**, 104–120 (2007).
47. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* **144**, 114–126 (2015).
48. Zhang, Y., Liao, Q. V. & Bellamy, R. K. E. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 295–305 (2020).
49. Buçinca, Z., Malaya, M. B. & Gajos, K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* **5**, 1–21 (2021).
50. Passi, S. & Vorvoreanu, M. Overreliance on AI: literature review. Tech. Rep. MSR-TR-2022-12, Microsoft Research (2022).
51. Schemmer, M., Kühl, N., Benz, C., Bartos, A. & Satzger, G. Appropriate reliance on AI advice: conceptualization and the effect of explanations. In *ACM Conference on Intelligent User Interfaces (IUI)*, 410–422 (2023).
52. Yang, E. & Wang, D. Benchmark illusion: disagreement among LLMs and its scientific consequences (2026). ArXiv:2602.11898.
53. Kuai, C. *et al.* How independent are large language models? A statistical framework for auditing behavioral entanglement and reweighting verifier ensembles (2026). ArXiv:2604.07650.
54. Kohli, G. Nine judges, two effective votes: correlated errors undermine LLM evaluation panels (2026). ArXiv:2605.29800.
55. Ladha, K. K. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science* **36**, 617–634 (1992).
56. Boland, P. J. Majority systems and the Condorcet jury theorem. *The Statistician* **38**, 181–189 (1989).
57. Steyvers, M. *et al.* What large language models know and what people think they know. *Nature Machine Intelligence* **7**, 221–231 (2025).
58. Lu, Z., Wang, D. & Yin, M. Does more advice help? The effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction* **8 (CSCW1)**, 1–31 (2024).
59. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* **108**, 9020–9025 (2011).
60. Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: a randomized experiment. *Science* **341**, 647–651 (2013).
61. Bahrami, B. *et al.* Optimally interacting minds. *Science* **329**, 1081–1085 (2010).
62. Tanaka, H. When is collective intelligence a lottery? Multi-agent scaling laws for memetic drift in LLMs (2026). ArXiv:2603.24676.

Supplementary Information for “A latent dimension of Condorcet’s jury theorem for multiple AI advisers”

Kazutoshi Sasahara¹, Aoi Naito^{1,2}, Ryo Fujie¹

¹Institute of Science Tokyo, Tokyo, Japan ²Carnegie Mellon University, Pittsburgh, PA, USA

Supplementary Notes 1–7 provide the derivations behind the main text: the two closed forms and their domains (Notes 1 and 2), the dissenting fraction (Note 3), the convergence rates and the constant $p = 4/5$ (Note 4), the disclosure reference scale K^* (Note 5), the independence baseline and its corrections (Note 6), and the three user-side readings of the split (Note 7).

Unless noted otherwise, the base model is homogeneous: K advisers answer a binary question independently, and each is correct with probability p , $1/2 < p < 1$ ($q = 1 - p$). Let $Y \sim \text{Bin}(K, p)$ count the correct verdicts. The user does not observe which side is correct, but the user-side quantities P_{vis} , m , and U depend only on how the verdicts split. Only P_{agg} (and therefore K^*) uses the ground-truth label.

Supplementary Note 1 | Aggregate reliability P_{agg}

Under fair tie-breaking, $P_{\text{agg}} = \Pr[Y > K/2] + \frac{1}{2} \Pr[Y = K/2]$ is non-decreasing in K ¹ and approaches 1 as $K \rightarrow \infty$.^{2,3} The latter follows from the strong law of large numbers (SLLN), because $Y/K \rightarrow p > 1/2$. Fair tie-breaking removes the even–odd sawtooth: P_{agg} is constant on each consecutive odd–even pair, $P_{\text{agg}}(2m) = P_{\text{agg}}(2m - 1)$. Together with the strictly increasing P_{vis} , this creates the parity band around K^* described in Supplementary Note 5.

Supplementary Note 2 | Visible dissent P_{vis}

Visible dissent is absent exactly in the four near-unanimous configurations $\{Y \leq 1\} \cup \{Y \geq K - 1\}$, with $P(Y \leq 1) = q^{K-1}(q + Kp)$ and $P(Y \geq K - 1) = p^{K-1}(Kq + p)$. Subtracting their probabilities from one gives

$$P_{\text{vis}}(K, p) = 1 - q^{K-1}(Kp + q) - p^{K-1}(Kq + p), \quad q = 1 - p,$$

using only the binomial distribution. The four configurations are distinct for $K \geq 3$, and a two-per-side split exists only for $K \geq 4$. We therefore use the closed form for $K \geq 4$ and set $P_{\text{vis}} = 0$ for $K \leq 3$. At $K = 3$ the formula itself returns 0; at $K = 2$ the four configurations overlap, so the formula does not apply. The closed form is symmetric, $P_{\text{vis}}(K, p) = P_{\text{vis}}(K, q)$, and $P_{\text{vis}} \rightarrow 1$ as $K \rightarrow \infty$ for every $p < 1$, because each term in the complement is a polynomial factor times ρ^{K-1} with $\rho \in \{p, q\} \subset (0, 1)$. It is also strictly increasing in K : for $K \geq 3$ and $0 < p < 1$, the closed form gives $[1 - P_{\text{vis}}(K)] - [1 - P_{\text{vis}}(K + 1)] = K(p^2 q^{K-1} + q^2 p^{K-1}) > 0$, while $P_{\text{vis}} = 0$ for $K \leq 3$ by definition.

This symmetry is structural, not merely algebraic. If $Y \sim \text{Bin}(K, p)$, then $K - Y \sim \text{Bin}(K, q)$. Swapping p and q simply relabels which side is correct ($Y \leftrightarrow K - Y$). Visible dissent, $\{2 \leq Y \leq K - 2\}$, is unchanged by this relabelling, so $P_{\text{vis}}(K, q) = P_{\text{vis}}(K, p)$. By contrast, the majority-correct event $\{Y > K/2\}$ maps to $\{Y < K/2\}$, so $P_{\text{agg}}(K, q) = 1 - P_{\text{agg}}(K, p)$ is *anti*-symmetric about $1/2$. Thus visible dissent is side-blind: it records that the panel is divided, not which side is correct. The assumption $p > 1/2$ is therefore unnecessary for the closed form itself; it is needed only when P_{agg} enters, and with it the rate comparison, the $4/5$ constant, and K^* .

Near-agreement points to the correct side. Conditional on a near-unanimous panel ($Y \leq 1$ or $Y \geq K - 1$), the side indicated by the panel is correct with probability $p^{K-1}(Kq + p) / [p^{K-1}(Kq + p) + q^{K-1}(Kp + q)]$. This probability approaches 1 because $(q/p)^{K-1} \rightarrow 0$. This is the statement used in the main text.

The “two” threshold. A single error can produce one lone dissenter, but it cannot produce two dissenters on each side. Two per side is therefore the smallest split that cannot be explained by one adviser’s error. Supplementary Note 3 removes this specific threshold by generalizing the definition to any relative threshold r .

Supplementary Note 3 | Dissenting fraction

With $m(K) = \min(Y, K - Y)/K$ and an analysis threshold $r \in (0, \frac{1}{2})$,

$$\lim_{K \rightarrow \infty} P(m(K) \geq r) = \begin{cases} 1, & r < q, \\ \frac{1}{2}, & r = q, \\ 0, & r > q, \end{cases} \quad \implies \quad m(K) \xrightarrow{P} q.$$

By the SLLN $Y/K \rightarrow p$, so $m(K) \rightarrow \min(p, q) = q$. At $r = q$ the central limit theorem (CLT) gives limiting probability $\frac{1}{2}$. For $r < q$, convergence to one is exponential: $1 - P(m \geq r) \leq 2e^{-2K(q-r)^2}$ by Hoeffding's inequality.⁴ The relative threshold also covers settings in which a single dissenting verdict warrants action, for example when one screening adviser flags a malignancy; a disclosure policy can preserve that lone signal rather than treat it as noise.

Supplementary Note 4 | Convergence rates, the $p = 4/5$ line, and the r -family

Write $E = 1 - P_{\text{agg}}$ and $N = 1 - P_{\text{vis}}$ for the two residuals, with P_{vis} the fixed-count split of the main text (at least two dissenters on each side). Both residuals decay exponentially, in the sense of logarithmic limits: $\lim_{K \rightarrow \infty} K^{-1} \log E = -R_{\text{agg}}$ and $\lim_{K \rightarrow \infty} K^{-1} \log N = -R_{\text{vis}}$ (polynomial prefactors, discussed below, do not affect these limits). By Cramér's theorem,⁵ the rate for aggregate reliability and the rate for visible dissent under the relative-threshold definition of Supplementary Note 3 are

$$R_{\text{agg}}(p) = D(\tfrac{1}{2} \| p) = \tfrac{1}{2} \log \tfrac{1}{4pq}, \quad R_{\text{vis}}(r, p) = D(1 - r \| p),$$

where $D(a \| p) = a \log \frac{a}{p} + (1 - a) \log \frac{1-a}{q}$. Here $R_{\text{vis}}(r, p)$ is the rate of the r -form of visible dissent (at least a fraction r dissenting on each side), defined for $0 < r < q$, the regime in which that event is inevitable (Supplementary Note 3). Its weakest form, $r \rightarrow 0$, has rate $R_{\text{vis}} = \log(1/p)$, set by the dominant near-unanimous edge $\{Y = K - 1\}$. This limiting rate is exactly that of the main-text fixed-count split, whose dominant non-split edge is likewise $\{Y = K - 1\}$. Hence

$$R_{\text{vis}} > R_{\text{agg}} \iff \log \tfrac{1}{p} > \tfrac{1}{2} \log \tfrac{1}{4pq} \iff p < \tfrac{4}{5}.$$

Equivalently the boundary is $p^2 = 4pq$, that is $p/q = 4$: the line sits exactly where a single adviser's odds of being correct reach 4 to 1. For the fixed-count split, the limit of N follows directly from the closed form of Supplementary Note 2, whose dominant term is Kqp^{K-1} , and the limit of E is Cramér's theorem for the lower tail $\{Y \leq K/2\}$. Consequently $\lim_{K \rightarrow \infty} K^{-1} \log(E/N) = R_{\text{vis}} - R_{\text{agg}}$, which is positive for every $1/2 < p < 4/5$: there $E/N \rightarrow \infty$, so a finite crossing $K^*(p)$ exists and visible dissent eventually remains more likely than a correct majority at every sufficiently large panel size.

For the relative-threshold family, the speed boundary $r^*(p)$ solves $D(1 - r^* \| p) = R_{\text{agg}}(p)$. Since $D(1 - r \| p)$ decreases in r , from $\log(1/p)$ at $r = 0$ to 0 at $r = q$, the root exists and is unique exactly when $R_{\text{agg}}(p) < \log(1/p)$, that is, below the line. It stays below q for $p < 4/5$ and collapses to 0 at $p = 4/5$, partitioning the (p, r) plane into *faster than the panel* ($r < r^*$), *inevitable but slower* ($r^* < r < q$), and *not inevitable* ($r > q$) (Supplementary Fig. 1).

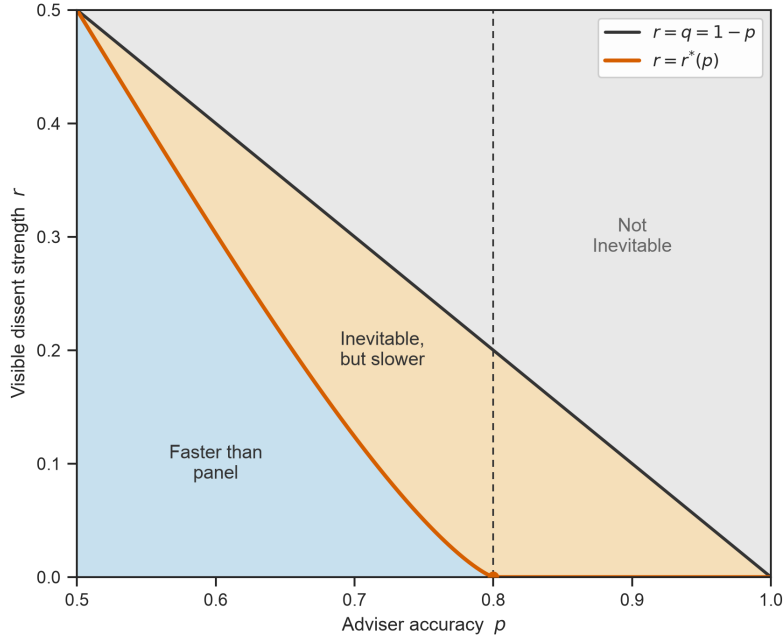
Asymptotic versus finite K . The exponents determine only which quantity eventually dominates as $K \rightarrow \infty$. The finite- K crossing also depends on the prefactors: the visible-dissent residual carries a factor of order K , whereas the majority-error residual carries the usual central-limit factor $K^{-1/2}$. In particular, the dominant visible-dissent edge term is $P(Y = K - 1) = Kqp^{K-1}$.

Invariance of the 4/5 boundary to fixed residual weighting. The 4/5 boundary is unchanged by any fixed multiplicative weight on the complementary residuals. For $p < 4/5$ and any fixed $w > 0$, replace $1 - P_{\text{vis}} \leq 1 - P_{\text{agg}}$ by $w(1 - P_{\text{vis}}) \leq 1 - P_{\text{agg}}$. This shifts the finite crossover but not the asymptotic boundary. Because both residuals decay exponentially, the fixed factor shifts the crossover by roughly $\log w / (R_{\text{vis}} - R_{\text{agg}})$ without changing which residual eventually dominates. Thus fixed residual weighting can move the disclosure reference scale, but not the $p = 4/5$ boundary.

Any definition of a split. The bound extends beyond the r -family. Under any definition in which a unanimous panel does not count as split, the non-split event contains $\{Y = K\}$, so $1 - P_{\text{vis}} \geq p^K$ and $R_{\text{vis}} \leq \log(1/p)$. Since $\log(1/p) > R_{\text{agg}}$ only for $p < 4/5$, no definition of a split places its rate boundary above $p = 4/5$; the minimal fixed-count split of the main text attains the bound.

The probability comparison involves disjoint outcomes. Let A denote visible dissent and B the majority-correct event (including random tie-breaking, so $P(B) = P_{\text{agg}}$). Every consultation falls into one of four cells: split or near-unanimous, crossed with majority correct or incorrect. The two probabilities share one cell, $A \cap B$, in which the panel splits but the majority is correct. Therefore $P(A) \geq P(B) \iff P(A \cap B^c) \geq P(A^c \cap B)$. In words, the probability comparison reduces to two disjoint outcomes: an incorrect-majority split and a correct near-unanimous panel. The shared cell is large under independence (probability 0.89 at $p = 0.7$, $K = 14$), but it cancels exactly. At $p = 0.7$, the incorrect-majority split has probability 0.0624 at both $K = 13$ and $K = 14$, while the correct near-unanimous outcome falls from 0.0637 to 0.0475; this is the crossing $K^*(0.7) = 14$ in Supplementary Note 5. The marginal and disjoint-outcome formulations therefore cross at the same K^* .

Heterogeneous panels. The homogeneous model is the equal-competence special case of a Poisson-binomial panel. For a sequence of independent heterogeneous panels with mean accuracy $\bar{p}_K = K^{-1} \sum_i p_i \rightarrow \bar{p} > 1/2$, aggregate reliability still satisfies $P_{\text{agg}} \rightarrow 1$,⁶ while $Y/K \rightarrow \bar{p}$ and hence $m \rightarrow 1 - \bar{p}$. The finite and asymptotic crossing behaviour, however, depends on the full accuracy profile $\{p_i\}$, not on the mean alone. For example, a panel with half $p_i = 0.999$ and half $p_i = 0.55$ has mean accuracy $\bar{p} \approx 0.77$, below 4/5, yet a correct majority remains more likely than a split at every panel size checked numerically



Supplementary Fig. 1. Phase diagram of visible dissent in (p, r) . The existence boundary $r = q$ and the speed boundary $r = r^*(p)$ partition the plane into *faster than the panel*, *inevitable but slower*, and *not inevitable*. The first of these regions closes at $p = 4/5$. The boundaries are large-deviation rate boundaries (the $K \rightarrow \infty$ ordering), with small- K ordering set by the finite crossover K^* .

up to $K = 400$. Mean accuracy alone therefore does not locate the crossing. When the per-adviser accuracies $\{p_1, \dots, p_K\}$ are available, compute P_{vis} and P_{agg} directly from the Poisson-binomial law and compare them for that panel ($K \geq 4$, fair tie-breaking). The $4/5$ line and $K^*(p)$ are exact summaries only for the equal-competence case; for independent heterogeneous panels, compute both probabilities from the full accuracy profile rather than applying $4/5$ to the mean accuracy.

Supplementary Note 5 | Disclosure reference scale K^*

Define

$$K^*(p) = \min\{K : P_{\text{vis}}(K, p) \geq P_{\text{agg}}(K, p)\},$$

the first panel size at which visible dissent becomes at least as likely as a correct majority. The values in the main text use fair tie-breaking: for example, $K^*(0.7) = 14$, whereas a strict-majority convention gives $K^*(0.7) = 10$ (Fig. 3b of the main text). For every $p < 4/5$, Supplementary Note 4 guarantees that such a finite crossing exists. Before K^* , $P_{\text{agg}} > P_{\text{vis}}$ by definition.

The two parity classes cross at different sizes, so immediately above K^* the ordering can depend on parity. At $p = 0.735$, for example, $P_{\text{vis}} \geq P_{\text{agg}}$ holds at $K = 22$ (0.98978 versus 0.98976) and at $K = 24$ (0.99404 versus 0.99230) but not at $K = 23$ (0.99219 versus 0.99230): the even class has crossed at $K = 22$, the odd class only at $K = 25$. Statements about a given panel size must therefore compare the two probabilities at that size; K^* marks the first crossing only. For every $p < 4/5$, visible dissent eventually remains more likely than a correct majority (Supplementary Note 4).

A numerical caution applies near the line: both probabilities are exponentially close to 1 there, and in double-precision arithmetic they round to 1 at moderate K , which can produce spurious crossings; the accompanying scripts therefore evaluate the complementary tails directly, in high-precision or exact rational arithmetic.

The rate comparison itself does not depend on how ties are broken; only this finite K^* does. Under the strict convention a tie counts as a failure, so P_{agg} is lower at even K , visible dissent overtakes sooner, and the strict K^* is never larger than the fair one.

Near $p = 4/5$, $K^*(p)$ grows rapidly, so small errors in the estimate of p can produce large uncertainty in the estimated reference scale (Fig. 3b of the main text).

The regime $p \geq 4/5$, including the boundary. That reliability keeps its lead at every finite K on and above the line (Fig. 2a of the main text) cannot be read from the figure alone. For $p \geq 4/5$ one has $\sqrt{4pq} \leq p$, so

$$1 - P_{\text{vis}}(K) \geq p^{K-1}(Kq + p) = (1 + Kq/p)p^K \geq (1 + Kq/p)(4pq)^{K/2},$$

while the Chernoff–Hoeffding bound gives $1 - P_{\text{agg}}(K) \leq e^{-KD(\frac{1}{2} \| p)} = (4pq)^{K/2}$.⁵ Hence $P_{\text{vis}}(K) < P_{\text{agg}}(K)$ for every $K \geq 4$ and every $p \geq 4/5$, so no finite K^* exists on or above the line. On the line itself the lead is an ordering, not a usable margin: both residuals vanish exponentially, and at $p = 4/5$, $K = 100$ the gap $P_{\text{agg}} - P_{\text{vis}}$ is about 5×10^{-9} .

Supplementary Note 6 | P_{vis} as a null model

The main text uses $P_{\text{vis}}(K, p)$ as an independence baseline. In the one-factor Gaussian copula settings illustrated in Supplementary Fig. 2, positive correlation reduces visible disagreement, producing a “dissent deficit” relative to that baseline. Positive pairwise correlation alone does not imply a lower split rate. For example, a four-adviser exchangeable panel with $\Pr(Y = 4) = 0.5$, $\Pr(Y = 2) = 0.4$, $\Pr(Y = 0) = 0.1$ has individual accuracy 0.7 and pairwise correlation ≈ 0.37 , yet $P_{\text{vis}} = 0.40$ against 0.26 under independence. Nor does positive pairwise correlation necessarily reduce majority accuracy. For example, a three-adviser exchangeable panel with $\Pr(Y = 3) = 0.5$, $\Pr(Y = 2) = 0.3$, $\Pr(Y = 0) = 0.2$ also has individual accuracy 0.7 and positive pairwise correlation (≈ 0.52), yet its majority accuracy is 0.80, compared with 0.784 under independence. Accordingly, a deviation of the adjusted split rate from the baseline in either direction, beyond the uncertainty of the accuracy estimate and sampling variation, provides evidence of residual dependence. Correlated extensions of the theorem^{7,8} concern aggregate reliability rather than the split rate. In the simulated Gaussian copula settings, the deficit provides an operational diagnostic. For a homogeneous reference panel, the deficit can be summarized by an effective panel size $K_{\text{eff}} < K$. When per-adviser accuracies are available, the Poisson-binomial comparison in Supplementary Note 4 should be used directly, with K_{eff} only as a secondary composite summary. Empirical validation of this diagnostic on real panels is left to future work. Concurrent evidence suggests the deficit can be severe. An audit of an LLM evaluation panel found nine frontier judges supplying only about two effective independent votes.⁹ It estimated that effective size from pairwise error correlations rather than from the split rate.

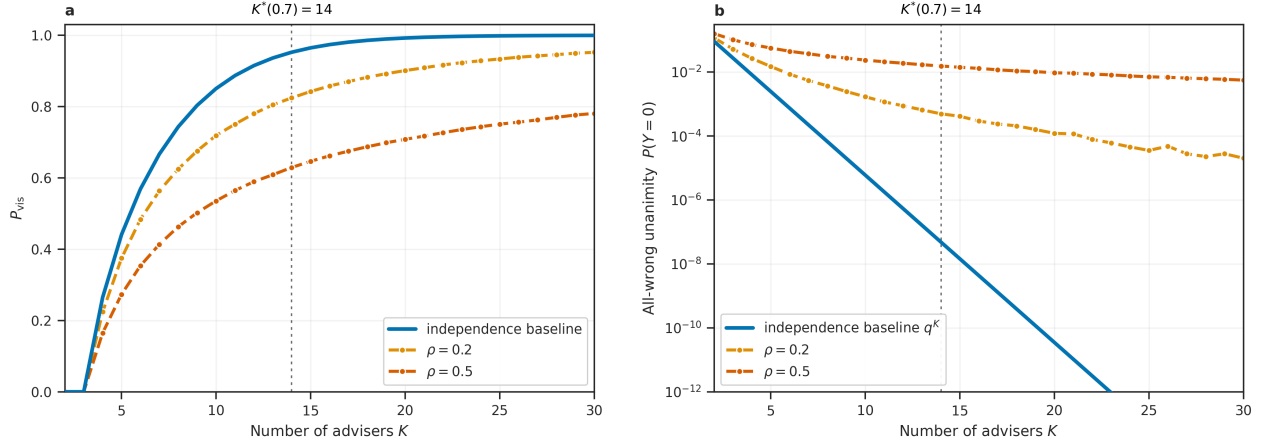
At a fixed mean accuracy, adviser heterogeneity can either raise or lower the split rate. Unequal competences reduce the probability of full unanimity ($\prod_i p_i \leq \bar{p}^K$ by AM–GM) but can increase the probability of a one-dissenter outcome. At small K , for example, one weak adviser can act as a consistent lone dissenter, keeping the split rate below the equal-competence baseline. When per-adviser accuracies are available, the Poisson-binomial baseline accounts for this heterogeneity exactly. Otherwise, a dissent deficit may reflect dependence, heterogeneity, or both, so K_{eff} is a composite diagnostic rather than a measure of independence alone. Supplementary Fig. 2 illustrates two effects in synthetic correlated panels: the dissent deficit (a) and the inflated all-wrong core (b). In panel (a), at $K^*(0.7) = 14$, the correlated split rates correspond to $K_{\text{eff}} \approx 9.4$ for $\rho = 0.2$ and $K_{\text{eff}} \approx 6.6$ for $\rho = 0.5$. Thus a panel nominally at the crossover can behave, in terms of its split rate, like a smaller independent panel that has not yet reached it.

A diagnostic procedure for the disclosure layer. For a homogeneous reference panel, or when only a pooled accuracy estimate is available, the steps below turn the dissent deficit into a diagnostic input. Let K be the panel size and let $\hat{p}$ be an accuracy estimate obtained from held-out labelled items or, more coarsely, from a reported benchmark:

- (i) Compute the independence baseline $P_{\text{vis}}(K, \hat{p})$.
- (ii) Measure the observed split rate $\hat{d}$, the fraction of items on which the panel visibly divides (at least two verdicts on each side) across a sample.
- (iii) If $\hat{d} < P_{\text{vis}}(K, \hat{p})$ after accounting for item difficulty and adviser heterogeneity, and the gap exceeds the uncertainty of the accuracy estimate and sampling variation, treat it as evidence of residual dependence. For a homogeneous reference, summarize its size by $K_{\text{eff}} \leq K$, defined through $P_{\text{vis}}(K_{\text{eff}}, \hat{p}) = \hat{d}$; non-integer values use the continuous extension of the closed form.
- (iv) K_{eff} summarizes the split rate relative to the baseline. Comparing it with $K^*(\hat{p})$ is an exploratory step that relies on the surrogate assumption discussed below.

Item difficulty and unequal adviser accuracies can both mimic a dissent deficit. Where possible, stratify items by difficulty and use the Poisson-binomial baseline when per-adviser accuracies are available. The reason is the nonlinearity in p : averaging accuracy first does not reproduce either $\mathbb{E}_i[P_{\text{vis}}(K, p_i)]$ or $\mathbb{E}_i[P_{\text{agg}}(K, p_i)]$, so $K^*(\bar{p})$ has the same limitation. After these corrections, K_{eff} summarizes whatever deficit remains. Difficulty stratification itself requires per-item difficulty estimates, which usually require labels. Whether the way the split rate changes with K can separate dependence from difficulty mixing without labels remains an open question.

Mean accuracy does not determine the ordering. As a numerical illustration, take $K = 14$ and two item populations with the same mean accuracy $\bar{p} = 0.7$: a homogeneous population with $p = 0.7$ on every item, and an equal mixture of items with $p = 0.51$ and $p = 0.89$. Within each item, advisers remain independent with equal accuracy, and the probabilities are averaged over items (Supplementary Table 1). In the homogeneous population, P_{vis} exceeds P_{agg} at $K = 14$, which is $K^*(0.7)$; in the mixture, the order at the same K is reversed. Mean accuracy alone therefore determines neither the ordering at a given panel size nor the location of a crossing, which is why the procedure above stratifies by difficulty before comparing with the baseline.



Supplementary Fig. 2. The independence baseline as a null model (illustrative). Synthetic correlated panels (Gaussian copula, $\rho = 0.2, 0.5$, dashed) are compared with the closed-form independence baselines (solid) at $p = 0.7$; the vertical dotted line marks the disclosure reference scale $K^*(0.7) = 14$. **(a)** Correlation pulls the split rate below the baseline $P_{\text{vis}}(K, p)$; all curves are zero for $K < 4$, where a two-per-side split is impossible. At K^* the baseline predicts $P_{\text{vis}} \approx 0.95$, whereas the correlated panels split on ≈ 0.825 ($\rho=0.2$) and ≈ 0.63 ($\rho=0.5$) of items. **(b)** The same correlation inflates the all-wrong unanimity rate $P(Y=0)$ (log scale): at K^* , $\approx 5 \times 10^{-4}$ ($\rho=0.2$) and $\approx 1.5 \times 10^{-2}$ ($\rho=0.5$) against the independence baseline $q^K \approx 5 \times 10^{-8}$, factors of 10^4 and 3×10^5 above it. Simulation design: one-factor Gaussian copula, $Z_i = \sqrt{\rho} Z_0 + \sqrt{1-\rho} \varepsilon_i$ with $Z_0, \varepsilon_1, \dots, \varepsilon_K$ independent standard normal variables, adviser i correct iff $Z_i < \Phi^{-1}(p)$, so that ρ is the correlation of the latent normal variables and each adviser is correct with probability p ; 400,000 simulated items per (K, ρ) cell, $K = 4, \dots, 30$, seed 7 (scripts/run_figS2_nullmodel.py, then scripts/figures/figS2_nullmodel.py in the code release). Synthetic simulation, not real-panel data.

Supplementary Table 1. Same mean accuracy, opposite ordering at $K = 14$. Item-averaged probabilities under independence and equal accuracy within each item; exact values rounded to four decimals.

Item population	$\mathbb{E}_i[P_{\text{agg}}(14, p_i)]$	$\mathbb{E}_i[P_{\text{vis}}(14, p_i)]$
$p = 0.7$ on every item	0.9376	0.9525
$p = 0.51$ items only (mixture component)	0.5293	0.9981
$p = 0.89$ items only (mixture component)	0.9998	0.4658
Equal mixture of 0.51 and 0.89	0.7646	0.7320

Step (iv) uses a working assumption: an independent surrogate that matches the panel's split rate is also used to approximate its aggregate reliability. The same dependence structure affects both quantities, which motivates the surrogate, but this is an approximation, not a theorem. In general, the error can go in either direction, so the surrogate is not guaranteed to be conservative. In the exchangeable copula setting of Supplementary Fig. 2, it is optimistic: at $K = 14$ the split-matched surrogate predicts $P_{\text{agg}} \approx 0.90$ versus the true 0.83 for $\rho=0.2$, and ≈ 0.86 versus 0.76 for $\rho=0.5$. Establishing when the approximation is conservative is left to future work.

In these simulations, correlation also raises the probability of an all-wrong panel above the independence baseline q^K (Supplementary Fig. 2b), so the over-trust caution of the main text applies alongside this procedure. At the correlations of Supplementary Fig. 2, that inflation means one unanimously wrong item in roughly 2,000 ($\rho=0.2$) and one in roughly 66 ($\rho=0.5$), each presented as maximal apparent certainty with nothing in the verdicts to flag it, versus one in twenty million under independence.

Supplementary Note 7 | Three user-side quantities and the monotone-response hypothesis

All three literatures read the same split Y ; the mapping is explicit, and for predictive multiplicity it is exact.

Predictive multiplicity, translated into the panel setting, is the rate at which the panel disagrees. If “disagreement” means any non-unanimous outcome, so that one dissenter is enough, its probability is $P_{\text{any}} = 1 - p^K - q^K$; for $K = 2$ this reduces to the pairwise rate $2pq$ below. If disagreement is defined by the minimal split in Supplementary Note 2, the probability is

$P_{\text{vis}}(K, p)$. The disclosure framework of the main text is stated on this second definition; Fig. 4a of the main text plots the first. The two events differ only in the one-dissenter cases $\{Y = 1\} \cup \{Y = K - 1\}$, so their complements have the same asymptotic rate $\log(1/p)$; the 4/5 rate boundary therefore holds under either definition. The connection to the predictive-multiplicity literature concerns the observed disagreement rate, not how the competing models were generated. In that literature the models are typically drawn from a level set fitted to one dataset, so they need not be independent or sampled as in the i.i.d. model here. **Cognitive / reconciliation load** is linked here to the expected dissenting fraction $\mathbb{E}[m(K)]$, which we treat as a candidate stimulus for that load. The expected fraction approaches q as the panel grows.

Reliance miscalibration has two directions. For under-reliance, the stimulus is visible dissent, which occurs with probability $P_{\text{vis}}(K, p)$. Because most splits occur on items whose majority is actually correct, repeated exposure to dissent can encourage undue doubt. For over-trust, the relevant stimulus is apparent unanimity $U = p^K + q^K$, especially its all-wrong component q^K : a unanimously wrong panel appears maximally certain. Under independence $q^K \ll p^K$ for $p > 1/2$, so this event is rare; correlation can inflate it substantially (Supplementary Note 6).

Pairwise disagreement. Two independent binary advisers disagree with probability $2pq$, one correct and one wrong in either order, so within the binary model an observed pairwise disagreement rate d would imply $p = \frac{1}{2}(1 + \sqrt{1 - 2d})$. This inversion is not applied to empirical rates in the main text. For multiple-choice answers, $d = 2pq + q^2 \Pr(\text{different wrong answers} \mid \text{both wrong})$, so the binary formula understates p ; correlation, unequal accuracies and difficulty mixtures shift d further. The identity serves only to translate the model's pairwise rate into the vocabulary of predictive multiplicity.

The monotone-response hypothesis (precise form). Predictive multiplicity is directly identified with an observable disagreement probability, but reconciliation load and reliance are behavioural responses to split-derived stimuli. We therefore assume only monotone relationships: expected reconciliation load does not decrease as $\mathbb{E}[m(K)]$ increases; expected under-reliance on majority-correct items does not decrease as exposure to visible dissent $P_{\text{vis}}(K, p)$ increases; and expected over-trust does not decrease as the all-wrong probability q^K increases. These are behavioural hypotheses, not results of the mathematical model.

The prediction it yields. Because P_{vis} , $\mathbb{E}[m]$, and q^K are explicit functions of (K, p) , the hypothesis fixes both the direction and the (K, p) -ordering of the behavioural effects. Holding p fixed, increasing K raises P_{vis} and does not decrease $\mathbb{E}[m]$, so the modelled stimulus for under-reliance increases while that for reconciliation load does not decrease. Whether the behavioural responses themselves become pronounced near $K^*(p)$ is an empirical question. Crossing the 4/5 line changes the relative ordering of P_{vis} and P_{agg} at large K (a finite K^* exists only below it), not the direction in which each stimulus moves with K .

Testing the hypothesis. A test exposes users to panels of varying (K, p) and measures reliance and reported load against the predicted P_{vis} and $\mathbb{E}[m]$; a decreasing or non-monotone response would falsify the hypothesis, while a (K, p) -independent response would instead show that behaviour is insensitive to the modelled stimuli. Known moderators of reliance, such as explanation quality, task domain, and user expertise, enter such a test as covariates; the hypothesis concerns the ordering induced by the stimuli, not the absence of other influences.

Confidence conditional on the observed vote. When presentation policies are compared as in Box 1 of the main text, specify how confidence is estimated, distinguishing model-based aggregate accuracy from confidence conditional on the observed vote. Let $d = |2Y - K|$ be the absolute vote margin. Given $d \geq 1$ and a fixed p , the majority is correct with probability $1/(1 + (q/p)^d)$, which depends on the observed votes only through the absolute margin; a tie ($d = 0$) is resolved by fair tie-breaking and is correct with probability $1/2$. At $p = 0.7$ and $K = 14$ the unconditional P_{agg} is 0.938, but 0.845 given an 8–6 split. In behavioural tests, declining to follow a majority that turned out to be correct is not by itself inappropriate reliance; appropriateness is judged against the information available at the time, including the margin.

References

1. Grofman, B., Owen, G. & Feld, S. L. Thirteen theorems in search of the truth. *Theory and Decision* **15**, 261–278 (1983).
2. Nitzan, S. & Paroush, J. *Collective Decision Making: An Economic Outlook* (Cambridge University Press, 1985).
3. Berend, D. & Paroush, J. When is Condorcet's jury theorem valid? *Social Choice and Welfare* **15**, 481–488 (1998).
4. Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**, 13–30 (1963).
5. Dembo, A. & Zeitouni, O. *Large Deviations Techniques and Applications* (Springer, 1998), 2nd edn.
6. Owen, G., Grofman, B. & Feld, S. L. Proving a distribution-free generalization of the Condorcet jury theorem. *Mathematical Social Sciences* **17**, 1–16 (1989).
7. Ladha, K. K. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science* **36**, 617–634 (1992).

8. Boland, P. J. Majority systems and the Condorcet jury theorem. *The Statistician* **38**, 181–189 (1989).
9. Kohli, G. Nine judges, two effective votes: correlated errors undermine LLM evaluation panels (2026). ArXiv:2605.29800.