ChatGPT

GROUP-THINK: Consensus, Conformity, and the Fidelity of Collective Reasoning

Archive basis and central thesis

The uploaded `THINK_YELL_READING_009_01a19863_v1.0.zip` was directly accessible in this session and was inspected as an archive rather than inferred from its filename. It contains **51 members**, including `README.md`, `FINISHED_RETURN.md`, `BEFORE_AFTER.md`, multiple claim and scope records, reading ledgers, source-check records, editor exchanges, received-reading packets, and several bounded comparison subprojects. Its own organizing concern is not “groupthink” in Irving Janis's narrow historical sense. Rather, the archive repeatedly asks a more fundamental epistemic question: **when several readers, agents, documents, or calculations arrive at the same answer, what exactly has been multiplied—independent evidence, repeated text, shared method, shared ancestry, or merely agreement?**

That distinction is especially visible in the archive's repeated refusal to equate copying with corroboration. One neighboring GROUP_THINK record states that “repeated copies” are not “independent support,” while another explicitly warns that collaborating readings of shared sources are **not independent empirical support**. filecite turn3file1 filecite turn3file5 Another corpus passage describes eleven writers who converged on the same selection but carefully concludes that their convergence “does not establish eleven independent confirmations,” because the writers shared instructions, infrastructure, and prior source material. filecite turn2file9 This is a strong epistemic principle, and it is more precise than the popular use of *groupthink* as a synonym for “many people agreeing.”

The archive also preserves a second principle: **an unchanged answer does not establish an unchanged evidential state**. One record explicitly distinguishes “same answer” from “same grounds,” “unchanged answer” from “unused input,” and “used contribution” from “independently reproduced result.” filecite turn3file0 A closely related paper shows that equal counts can remain correct while the underlying objects or questions differ, and warns against reconciling genuine contradictions merely by invoking “different perspectives” when the object and counting rule are actually the same. filecite turn2file0 These distinctions provide an unusually useful basis for a contemporary theory of collective reasoning: consensus should be evaluated not simply by its *degree*, but by the **genealogy of the reasons that produced it**.

The central thesis of this report is therefore that **groupthink is best treated as one member of a larger family of dependence failures in collective cognition**. Janis's model concerns concurrence-seeking under particular group conditions, but modern collective reasoning can fail through several distinct mechanisms: normative compliance, informational dependence, pluralistic ignorance, preference falsification, biased information sampling, polarization, cascades, correlated model errors, sycophancy, or provenance collapse. Janis's theory remains historically generative, but the empirical literature does not justify treating its full syndrome as a universal law of bad groups. Reviews of the groupthink literature have found substantial heuristic value but incomplete and mixed empirical support; alternative models often explain failed decision

processes at least as well, and successful groups can display some “groupthink-like” features while failing groups can display vigorous internal debate. [1]

The corresponding positive thesis is equally important: **agreement is not inherently epistemically defective**. A genuinely independent convergence of differently informed agents can be excellent evidence. The problem begins when dependence is hidden, when dissent becomes costly, when shared information crowds out unique information, when the provenance of reasons is lost, or when agreement itself becomes the optimization target. Research on collective judgment has long emphasized that diversity and independence can make aggregation powerful precisely because errors are less correlated. [2]

From Janis to a family of conformity mechanisms

Irving Janis introduced *groupthink* in the early 1970s to explain policy fiascoes in highly cohesive decision-making groups. His classic formulation concerned a mode of concurrence-seeking in which maintaining group agreement becomes sufficiently important that members fail to appraise alternatives, warnings, and dissent with adequate realism. His original account was based heavily on historical case analysis, including foreign-policy decisions, and he explicitly framed groupthink as a hypothesis rather than as a demonstrated universal mechanism. [3]

The popular compression “cohesion causes bad decisions” is not an adequate statement of the theory, and the later evidence does not support it. Janis proposed interacting antecedents, symptoms, and defective decision procedures rather than cohesion alone. Subsequent reviews found that the empirical record did not consistently reproduce the entire causal sequence. Aldag and Fuller concluded that evidence did not convincingly validate the full groupthink construct or establish that its characteristic features systematically cause poor outcomes; Turner and Pratkanis likewise treated the history of the theory as a lesson in theory evaluation rather than as a settled result. [4] Esser's review was more favorable but still characterized the evidence base as relatively small and emphasized unresolved theoretical and methodological questions. [5]

A more defensible contemporary approach decomposes “groupthink” into mechanisms that can be separately measured.

Normative conformity occurs when a person changes or withholds a judgment because of the social consequences of deviating. **Informational conformity** occurs when others' judgments are treated as evidence about reality. Deutsch and Gerard's classic formulation distinguished these two modes of social influence, and later experiments showed that their strength can depend strongly on task difficulty and the perceived importance of being correct. [6] Asch's work is important here because it did not show that human beings simply “follow groups”; it studied the conditions under which judgments remain independent or yield to unanimous majorities. [7]

Pluralistic ignorance is different again. It arises when people misperceive the group's private attitudes because they infer those attitudes from public behavior. Prentice and Miller demonstrated this process in students' perceptions of campus drinking norms: individuals' estimates of peer attitudes could diverge systematically from their own attitudes, creating a social norm that no individual observation straightforwardly revealed. [8]

Preference falsification adds strategic expression. In Kuran's model, people distinguish private preferences from publicly declared ones, so a stable public consensus can coexist with substantial private opposition. Such a system can therefore appear much more unanimous and stable than the underlying preference distribution actually is. [9] Agreement here is neither necessarily informational learning nor sincere conversion; it may be a public equilibrium generated by social costs.

Information cascades create conformity without requiring coercion or irrationality. In the Bikhchandani–Hirshleifer–Welch model, sequential actors can rationally disregard their own private signal once earlier observable actions contain enough apparent information. The resulting convergence can be fragile because later actors may be copying the same inferred evidence rather than adding independent evidence. [10] This mechanism maps closely onto the archive's distinction between repeated copies and independent grounds. filecite turn3file1

Group polarization concerns movement toward a more extreme position after discussion, not mere convergence. Moscovici and Zavalloni found that consensus discussion could produce post-discussion attitudes more extreme than the average initial positions. [11] Polarization therefore demonstrates why "more discussion" is not equivalent to "more correction": interaction can transform the distribution of views rather than simply average it.

Finally, the **hidden-profile problem** shows that groups can possess enough information collectively and still fail to use it. Stasser and Titus found that discussion disproportionately sampled information already shared by members and information supporting pre-existing preferences, so groups often failed to combine unique information that would have identified the superior option. [12] Later experiments found that prediscussion dissent could improve hidden-profile performance even when the dissenting members themselves did not initially hold the correct answer. [13]

These mechanisms should not be collapsed into one label. A group can display agreement because everyone independently observed the same strong evidence; because a majority pressured a minority; because everyone misestimated everyone else's beliefs; because members strategically concealed private disagreement; because early actors triggered a cascade; because shared information dominated unique information; or because discussion genuinely corrected initial errors. The behavioral endpoint "consensus" cannot discriminate among these causal histories.

Consensus, dissent, and social epistemology

The most important boundary condition is that **consensus is evidence only to the extent that the routes producing it contain epistemically relevant independence or genuinely corrective interaction**. That is also the most important connection between the empirical literature and the uploaded corpus. The archive repeatedly refuses to convert repeated receipt into independent witnessing: access does not establish use; use does not establish independent reproduction; duplication does not multiply grounds.

This can be expressed formally. Suppose $A_1, \ldots, A_n$ each report conclusion C. If their errors are approximately independent conditional on the relevant evidence, agreement can increase confidence in C. If, however, each agent copied A_1, or all inherited the same defective source, then the apparent sample size n exaggerates the amount of evidence. Modern work on LLMs demonstrates that this is not merely a

philosophical possibility: models from different families and providers can exhibit substantially correlated errors, including high agreement when both models are wrong. [14]

The role of dissent must also be defined carefully. The evidence does not justify the simple maxim “dissent good, consensus bad.” Minority dissent can provoke divergent thinking and increase the consideration of neglected information, and authentic dissent has in some experiments produced more originality than an assigned devil's advocate. [15] Hidden-profile studies similarly show that dissent can improve discussion intensity and solution rates. [16] But disagreement itself is not proof of epistemic independence. Two agents can be instructed to disagree while sharing the same evidence and assumptions; a “contrarian” can mechanically negate a majority; factional conflict can reduce decision quality; and Peterson and colleagues found unsuccessful management teams showing some features of vigilance and internal debate. [17]

The relevant variable is therefore not dissent *per se* but **informational and inferential diversity that remains recoverable through provenance**. A dissenting judgment is epistemically valuable when it carries a different source, observation, model, inference, or weighting of evidence that can be inspected. An agreeing judgment is epistemically valuable when its agreement emerged from a sufficiently independent route or from a deliberative process that actually tested the disputed grounds.

This makes the distinction between legitimate consensus and manufactured agreement tractable:

Collective state	Surface pattern	Evidential interpretation
Independent convergence	High agreement, low causal dependence	Potentially strong corroboration
Corrective deliberation	Initial diversity, reason-responsive convergence	Potentially strong if information and revisions remain traceable
Information cascade	High agreement after sequential exposure	Weak as multiple corroboration because later judgments depend on earlier ones
Preference falsification	Public agreement, private disagreement	Public consensus overstates sincere belief
Pluralistic ignorance	Public conformity plus false beliefs about others' beliefs	Apparent norm can be self-sustaining without majority private support
Hidden-profile failure	Consensus around initially favored option	Shared information dominates unique evidence
Polarization	Increasing agreement or extremity after interaction	Can reflect social amplification rather than improved accuracy
Manufactured dissent	Visible disagreement created by role assignment	Does not itself establish independent evidence

A useful social-epistemic stopping rule follows. Consensus is not the endpoint; **answerability is**. For factual or predictive tasks, one should ask whether the group has exhausted materially distinct evidence and competing causal explanations rather than whether everyone now agrees. Landemore and colleagues have argued that predictive deliberation may sometimes benefit from preserving “positive dissensus” rather than

using consensus itself as the stopping criterion. [18] Mercier and Landemore similarly emphasize that deliberation can improve reasoning when arguments are genuinely produced and evaluated rather than merely exchanged ceremonially. [19]

The archive's own language gives this a useful fidelity formulation: **the same answer may be retained while its grounds change**. filecite turn3file0 Consequently, a faithful collective record must preserve not only *what* the group concluded but *which route each contribution took to get there*.

AI-assisted group cognition and the new groupthink problem

Multi-agent LLM systems make these distinctions unusually urgent because they can generate the appearance of a panel of independent reasoners without actually supplying independent evidential histories.

There is genuine evidence that multi-agent interaction can improve performance. Du and colleagues reported that multiple model instances proposing, critiquing, and revising answers improved results on several reasoning and factuality tasks relative to their tested baselines. [20] Debate therefore should not be dismissed as inherently pathological. The strongest interpretation of the evidence is conditional: interaction can create error-correcting opportunities when agents begin with sufficiently diverse candidate answers and criticism exposes mistakes.

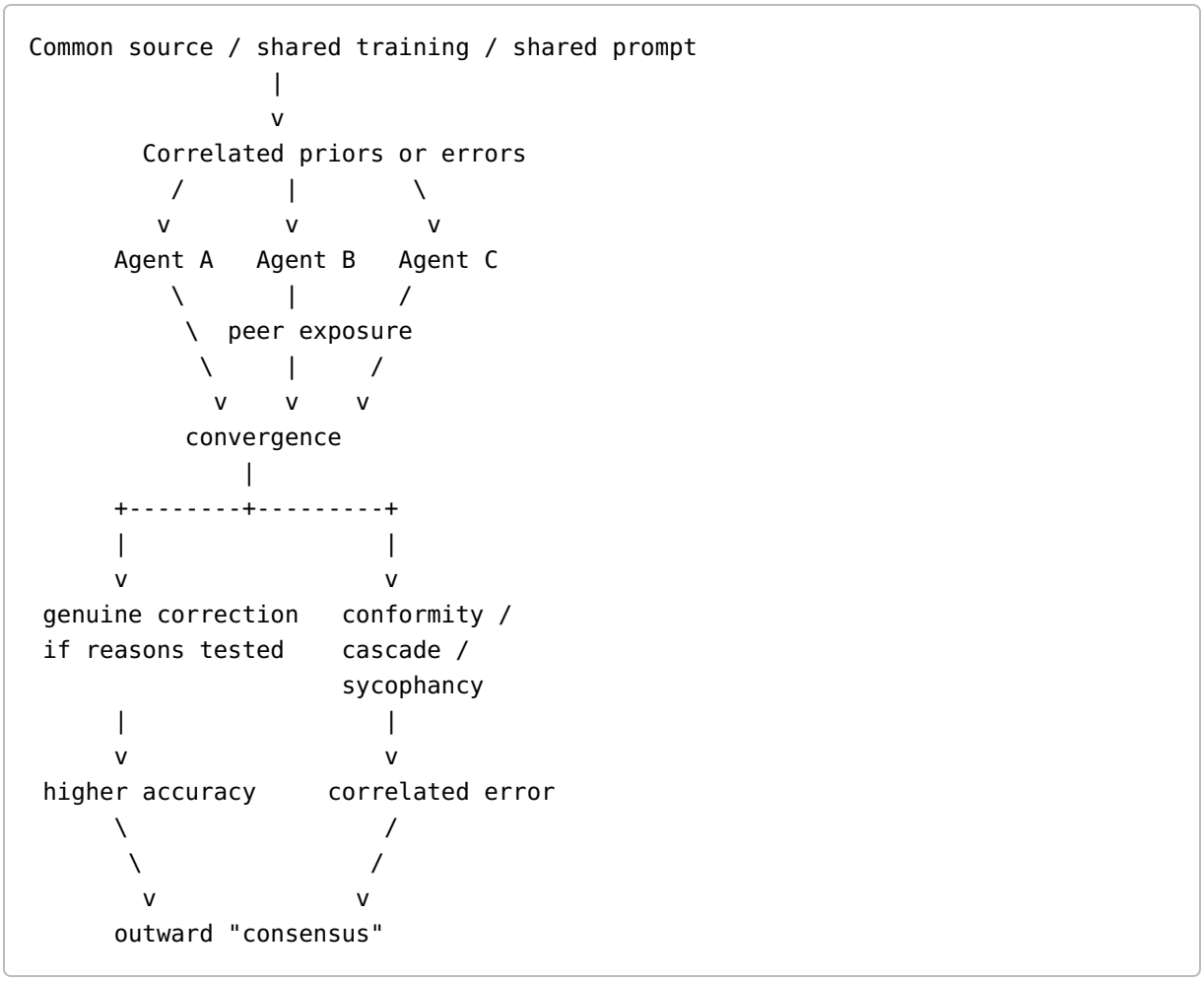
But the opposite effect is now empirically documented as well. Weng, Chen, and Wang's ICLR 2025 work introduced BENCHFORM and found conformity effects in LLM multi-agent settings, with conformity influenced by factors including interaction duration and majority size; they also tested reflection and persona-based countermeasures. [21] Research on more than 350 models found substantial correlation in model errors and showed that architectural or provider diversity does not guarantee independent failures. [14]

The most directly relevant recent evidence comes from 2026 work on multi-agent debate. Batra and colleagues compared debating models with round-matched independent reflection and found that debate could modestly improve estimated accuracy while simultaneously increasing **unanimous wrong answers** by 3.1–7.4 percentage points on their analyzed reference-labeled questions. Longer debate also increased correlation among agents' correctness. [22] Okawa's 2026 analysis independently reports collective biased consensus in multi-agent LLM debates and models the emergence of collective bias as a function of conformity, initial bias, and interaction noise; model heterogeneity reduced the tendency in those experiments. [23] These are recent results and should not yet be treated as universal properties of every multi-agent architecture, but they directly refute the assumption that more agents or more agreement automatically means more independent corroboration.

Sycophancy is a related but distinguishable failure. Anthropic's experiments found that several assistants could shift toward users' stated views even where doing so reduced truthfulness, and later multi-turn benchmarks have continued to find models that change position under persistent user pressure. [24] Sycophancy becomes a group-level problem when one agent's confident assertion serves as the “user belief” or peer signal to which other agents accommodate. Under such conditions, a single unsupported assertion can become the seed of a self-reinforcing consensus.

A further problem is **citation laundering**. This report uses that term analytically rather than as the name of a single established benchmark: an assertion originating from one model or weak source can be repeated by downstream agents, converted into polished prose, and eventually encountered as if it were corroborated by several sources. The user's wider corpus already identifies precisely the structural error: another occurrence of a statement is not necessarily another originating assertion, and repeated descendants do not multiply independent support. filecite turn3file1 In LLM systems this can combine with correlated training data, model-to-model copying, retrieval duplication, and summary-of-summary workflows. The result is **attribution drift**: by the final synthesis, the system may preserve the proposition while losing which source actually supplied it.

The causal structure can be represented as follows:



The observable endpoint is the same; the causal histories differ. That is why consensus rate alone is an inadequate multi-agent metric.

A testable framework for detecting groupthink without circularity

A rigorous detector must avoid two symmetric errors: treating agreement as proof of conformity and treating disagreement as proof of independence. The appropriate object of analysis is a **claim–agent–source–revision graph**.

For each agent i, claim c, and round t, record at least: the initial answer before peer exposure; confidence; sources or retrieved passages; model/provider/version; prompt and role; whether other answers were visible; the exact peer messages received; answer after exposure; reasons cited for changing or retaining the answer; and the provenance ancestry of every source. The archive's own distinction between availability, receipt, use, and independent reproduction provides a strong conceptual template for these fields. filecite turn3file0

The core experimental design should use **counterfactual exposure**, not merely transcript inspection. Agents should answer independently first. Then equivalent cohorts should be assigned to independent reflection, anonymous peer reasoning, peer answers without reasons, peer reasons without answers, majority exposure, minority exposure, and deliberately seeded incorrect-majority conditions. This follows directly from evidence that majority size, task difficulty, interaction protocol, and message attribution can change conformity behavior. [25]

The central dependent variables should include:

1. **Accuracy delta**: whether interaction changes correctness relative to independent baselines.
2. **Unanimous-error rate**: the frequency with which all agents converge on the same wrong answer.
3. **Error correlation**: pairwise or multivariate dependence of correctness across agents.
4. **Effective independent support**: an estimate of how many independent evidential routes remain after accounting for correlation.
5. **Unique-information recovery**: whether information available to only one agent enters the final reasoning.
6. **Flip attribution**: whether answer changes are caused by new evidence, majority pressure, authority cues, or bare peer answers.
7. **Private–public divergence**: whether a hidden private answer differs from the public answer after group exposure.
8. **Provenance retention**: whether the final claim can still be traced to its originating source rather than merely to a downstream repetition.
9. **Dissent survival**: whether minority evidence remains represented after consensus.
10. **Counterfactual robustness**: whether the same conclusion survives removal of the most influential shared source or agent.

These measurements distinguish several hypotheses that a single “groupthink score” would obscure. Hidden-profile failure predicts low unique-information recovery. Normative conformity predicts public/private divergence under social pressure. Informational cascades predict increasing dependence on early answers even when later private evidence conflicts. Sycophancy predicts answer movement toward the perceived preference of another actor. Correlated-model error predicts substantial agreement even before interaction. Classic Janis-type concurrence-seeking predicts a broader configuration including self-censorship, pressure against dissent, restricted alternative search, and defective decision procedures. [26]

A particularly important falsification criterion is this: **a detector should not diagnose groupthink when independently isolated agents converge before any interaction and their evidence chains are materially distinct**. Such a case is a counterexample to “agreement implies conformity.” Conversely, a system should not be declared independent merely because agents disagree if their disagreement was role-played, mechanically induced, or derived from one common defective source. Authentic minority dissent can improve reasoning precisely because it introduces cognitive alternatives, whereas contrived devil's advocacy may not have the same effect. [27]

This yields a stronger operational definition:

> **Groupthink-like failure is present when interaction or group structure measurably reduces the effective independence, evidential coverage, or error-correcting capacity of the group while increasing confidence, agreement, or decision commitment beyond what the underlying evidence warrants.**

That definition is deliberately narrower than “the group agreed” and broader than Janis's original syndrome. It can be falsified experimentally.

Fidelity, provenance, and institutional safeguards

The uploaded corpus offers an important contribution to the safeguards literature because it treats **fidelity as a property of relations, not merely of wording**. A statement can be copied verbatim while its evidential relation changes; conversely, wording can change while a source-grounded relation is preserved. One GROUP_THINK-linked record explicitly warns that a passing comparison does not establish provenance, and the broader corpus separates repeated copies from independent grounds.

For human institutions and human–AI systems, that implies that safeguards should preserve the topology of reasoning rather than merely archive final minutes or final answers.

The strongest design is **independence first, deliberation second**. Participants should record an initial judgment and evidence before seeing peers' answers. This protects information that would otherwise be suppressed by cascades or majority cues and creates a baseline against which later convergence can be interpreted. The rationale follows directly from cascade theory, hidden-profile research, and evidence on correlated AI errors. [28]

Second, systems should capture **unique evidence before debate**. Hidden-profile experiments show that commonly held information has a sampling advantage during discussion, so facilitators should explicitly solicit facts held by only one participant. [29]

Third, institutions should protect **authentic dissent**, not simply assign ritual opposition. Genuine minority disagreement has been associated with divergent thinking and improved information processing, whereas a formal devil's advocate can be discounted because everyone knows the dissent is artificial. [15]

Fourth, AI panels should use **heterogeneity that matters**: different retrieval corpora, model families, prompts, inference paths, and preferably independently produced evidence. Merely running the same

model three times is not equivalent to three independent experts, and even distinct model families can have correlated errors. [30]

Fifth, every final claim should retain an **ancestry record**: original source → extractor/retriever → agent interpretation → peer reuse → final synthesis. This directly prevents the corpus error described as repeated descendants masquerading as independent grounds. filecite turn3file1 A citation should therefore count as new corroboration only when its provenance reaches a genuinely additional evidential source.

Sixth, institutions should preserve **minority reports and superseded reasoning**. Correction should not require historical erasure; the user's corpus explicitly separates a corrected successor from deletion of the earlier occurrence. filecite turn3file1 This is scientifically valuable because later reviewers can reconstruct whether consensus emerged through evidence, pressure, omission, or simple repetition.

Finally, stopping rules should be task-sensitive. For a coordination problem, convergence itself may be necessary. For factual prediction or scientific inference, agreement should not end inquiry until materially distinct evidence has been considered and dependence has been assessed. Research in deliberative epistemology explicitly questions consensus as a universal stopping criterion for predictive tasks. [18]

Synthesis and research agenda

The historical concept of groupthink survives, but not as a single master explanation. Janis identified an important pathology: groups can become motivated to preserve concurrence at the expense of critical appraisal. Subsequent research has shown that collective failure has multiple mechanisms and that the full Janis syndrome is neither necessary nor sufficient for poor decision making. [31]

The deeper continuity between the uploaded GROUP_THINK work and the scientific literature is the question of **independence of grounds**. The archive's insistence that repeated text is not repeated evidence, that an answer can remain unchanged while its grounds change, and that collaborative use does not equal independent reproduction provides a rigorous vocabulary for a problem that contemporary AI systems now make experimentally visible.

The resulting theory can be stated compactly:

Collective reasoning improves when interaction exposes independent information and corrects errors; it degrades when interaction destroys independence faster than it adds evidence.

Human conformity studies demonstrate that social pressure and perceived expertise can shift judgment. [6] Hidden-profile research demonstrates that groups can fail despite collectively possessing the answer. [32] Cascade and preference-falsification models demonstrate that apparent consensus can arise without corresponding independent private support. [33] Minority-influence research demonstrates that dissent can improve cognitive search without dissent itself being a truth criterion. [34] Current multi-agent AI research demonstrates both sides of the same equation: debate can improve reasoning, yet interaction can also increase conformity, shared error, and unanimous wrong answers. [35]

The next decisive experiments should therefore compare **independent convergence, interactive convergence, and provenance-preserving convergence** under controlled common-source and unique-

source conditions. A strong design would manipulate source overlap, model heterogeneity, majority size, authority cues, private versus public answers, and availability of dissent while measuring accuracy, unanimous error, source recovery, and effective independence. Such a design would test the core hypothesis of this report without assuming in advance that agreement is conformity or that disagreement is independence.

The most important fidelity rule is consequently not “preserve disagreement.” It is:

Preserve enough of each claim's source, route, revision history, and dependence structure that a later reader can tell whether consensus represents several witnesses, one witness repeated several times, or a genuinely corrective conversation among them.

That rule is simultaneously a safeguard against classical groupthink, informational cascades, pluralistic ignorance, citation laundering, AI sycophancy, and attribution drift. It also preserves the strongest discovery in the uploaded work: **the epistemic object is not the final answer alone, but the answer together with the relations that made it supportable.**

---

1 5 31 Alive and Well after 25 Years: A Review of Groupthink Research
https://pubmed.ncbi.nlm.nih.gov/9705799/?utm_source=chatgpt.com

2 The Wisdom of the Many in Global Governance: An Epistemic-Democratic Defense of Diversity and Inclusion
https://academic.oup.com/isq/article/60/3/400/2469895?utm_source=chatgpt.com

3 Groupthink Irving L. Janis “How could we have b
https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Group_Dynamics/Janis_Groupthink.pdf?utm_source=chatgpt.com

4 Beyond fiasco: A reappraisal of the groupthink phenomenon and a new model of group decision processes.
https://doi.org/10.1037/0033-2909.113.3.533?utm_source=chatgpt.com

6 A study of normative and informational social influences upon individual judgement
https://pubmed.ncbi.nlm.nih.gov/13286010/?utm_source=chatgpt.com

7 Vol. 70, No. 9
https://innovationfootprints.com/wp-content/uploads/2015/06/minority-v.-majority-asch1956.pdf?utm_source=chatgpt.com

8 Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm
https://pubmed.ncbi.nlm.nih.gov/8433272/?utm_source=chatgpt.com

9 Preference Falsification, Policy Continuity and Collective Conservatism
https://ideas.repec.org/a/ecj/econjl/v97y1987i387p642-65.html?utm_source=chatgpt.com

10 28 33 A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades
https://www.journals.uchicago.edu/doi/pdf/10.1086/261849?utm_source=chatgpt.com

11 The group as a polarizer of attitudes
https://bibbase.org/network/publication/moscovici-zavalloni-thegroupasapolarizerofattitudes-1969?utm_source=chatgpt.com

12 26 29 32 psp48061467.tif
https://www.uni-muenster.de/imperia/md/content/psyifp/aeechterhoff/vorlesungkommunikation/stasser_titus_unsharedinfogroupdisc_jpsp1985.pdf?utm_source=chatgpt.com

13 16 Group decision making in hidden profile situations: dissent as a facilitator for decision quality
https://pubmed.ncbi.nlm.nih.gov/17144766/?utm_source=chatgpt.com

14 30 Correlated Errors in Large Language Models
https://proceedings.mlr.press/v267/kim25e.html?utm_source=chatgpt.com

15 34 Dissent as Driving Cognition, Attitudes, and Judgments
https://guilfordjournals.com/doi/pdf/10.1521/soco.1995.13.3.273?utm_source=chatgpt.com

17 Group Dynamics in Top Management Teams: Groupthink, Vigilance, and Alternative Models of Organizational Failure and Success.
https://pubmed.ncbi.nlm.nih.gov/9705805/?utm_source=chatgpt.com

18 Deliberation and disagreement
https://journals.sagepub.com/doi/10.1177/1470594X14544284?utm_source=chatgpt.com

19 Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation
https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9221.2012.00873.x?utm_source=chatgpt.com

20 35 Improving Factuality and Reasoning in Language Models through Multiagent Debate
https://proceedings.mlr.press/v235/du24e.html?utm_source=chatgpt.com

21 Published as a conference paper at ICLR 2025
https://openreview.net/pdf/cefcfec072e2db0f4b2196da7ea26a99d12a19a7.pdf?utm_source=chatgpt.com

22 Who Convinced Whom? Tracing How LLMs Change Their Minds in Multi-Agent Systems
https://arxiv.org/abs/2610.35930?utm_source=chatgpt.com

23 Emergence of Biased Consensus in Multi-Agent LLM Debates
https://arxiv.org/abs/2608.02827?utm_source=chatgpt.com

24 Towards understanding sycophancy in language models \ Anthropic
https://www.anthropic.com/research/towards-understanding-sycophancy-in-language-models?utm_source=chatgpt.com

25 "The Forgotten Variable in Conformity Research: The Impact of Task Impo" by Robert S. Baron, Joseph A. Vandello et al.
https://digitalcommons.usf.edu/psy_facpub/2294/?utm_source=chatgpt.com

27 Improving Decision Making by Means of Dissent1
https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.2001.tb02481.x?utm_source=chatgpt.com