From Advisor to Voting Teammate: Institutional Authority and Information Structures of AI Agents in Bounded-Rational Human Groups

Yiheng Tian[*†]
University College London
London, UK
yiheng.tian.25@ucl.ac.uk

Jingnan Zhang[*†]
University College London
London, UK
jingnan.zhang.24@ucl.ac.uk

Yiven Zhu[*†]
University College London
London, UK
louis.zhu.23@ucl.ac.uk

Fei Xu[*†]
University of California, Berkeley
Berkeley, USA
fei.xu@berkeley.edu

Hao Kan[*†]
University College London
London, UK
holly.kan.25@ucl.ac.uk

Leenuja Kannan[*†]
University College London
London, UK
leenuja.kannan.25@ucl.ac.uk

Niyaa Meganathan[*†]
University College London
London, UK
niyaa.meganathan.25@ucl.ac.uk

Aiden Yiliu Li[*†‡]
University College London
London, UK
yiliu.li.23@ucl.ac.uk

Abstract

While AI agents are increasingly integrated into human teams, it remains unclear whether AI should act as a peripheral advisor or a voting member, and how these roles interact with group hierarchy. We build an agent-based model where bounded-rational agents integrate private signals, AI recommendations, and social cues via trust-weighted updating and conformity mechanisms. Six institutional regimes cross access structure (centralised vs. distributed) with authority type (advisory vs. voting) across 1,113,600 runs. In our simulations, only centralised advisory access outperforms the no-AI baseline; four other human–AI configurations (advisor-distributed, pure-voter, hybrid-centralised, hybrid-distributed) match or underperform it. Distributed architectures produce substantially higher cascade rates, and granting AI voting power compounds this through normative conformity pressure, suggesting a compounding interaction that may suppress independent judgment. AI vote weight may distort the wisdom of the crowd. Our results challenge the intuition that "democratising" AI access is always beneficial, suggesting instead that institutional "gatekeeping" (centralised advisory) is essential to preserving collective intelligence in human-AI hybrids.

[*]These authors contributed equally to this work.
[†]Work done in UCL Nexus Labs.
[‡]Corresponding author.

CHI 2026, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Yiheng Tian, Jingnan Zhang, Yiven Zhu, Fei Xu, Hao Kan, Leenuja Kannan, Niyaa Meganathan, and Aiden Yiliu Li. 2026. From Advisor to Voting Teammate: Institutional Authority and Information Structures of AI Agents in Bounded-Rational Human Groups. In *Proceedings of Workshop on Human-Agent Collaboration (CHI 2026).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The integration of artificial intelligence into human decision-making groups marks a pivotal shift in collective intelligence. Fulay et al. [6] proposed AI representatives in shareholder democracy, highlighting both the potential for enhanced representation and the risks of voter disengagement. Our study examines how AI can either mitigate or amplify human cognitive biases, informing the design of more effective collaborative structures.

Bounded rationality holds that decision-makers are intendedly rational, yet inherent cognitive constraints can lead to judgment failures [9]. This limitation drives trust miscalibration: systematic over- or under-reliance on AI as individuals struggle to assess algorithmic reliability. In groups, social influence increases individual biases. For modelling human groups with AI agents, we distinguish two social influences: informational influence, in which individuals update beliefs because others serve as evidence, leading to information cascades when early signals are overweighted; and normative influence, in which individuals shift expressed positions to align with perceived authority, even when beliefs differ. Informational influence primarily affects belief accuracy, whereas normative influence can increase consensus without improving accuracy, generating false convergence [3]. Over-trust occurs when people rely too heavily on AI despite its mistakes, and is linked to algorithm appreciation, in which AI advice is preferred even when suboptimal [11] [12]. Under-trust is the opposite: people avoid helpful AI advice even when it would improve decisions, known as algorithm aversion [4] [12].

In a centralised structure, AI becomes a private tool, reinforcing the leader's authority. Concentrated influence can improve group accuracy via a high-competence helper [2], but risks bias and reduces critical evaluation. A leader with private AI access either improves accuracy or, if trust is miscalibrated, transmits AI errors with little peer correction. In a distributed structure, AI becomes a shared collaborator, expanding the solution space through diverse viewpoints [13]. However, the group must negotiate both human-human and human-AI disagreements, where AI's proactive communication style is critical for shared situational awareness and acceptance as a legitimate teammate [17]. Taken together, these studies imply that increasing concentration of AI access raises the risk that miscalibrated trust in one individual becomes a system-level error; increasing coverage raises the probability of disagreement but also the coordination burden of reconciling conflicting human and AI claims.

Granting an AI teammate actual voting power introduces conformity pressure beyond individual trust calibration. Experiments show that voting power alters how humans experience being outvoted and perceive the legitimacy of decisions, shaping team outcomes [8]. Lindner's [10] model demonstrates that even a single voter with superior reliability can enhance collective accuracy, providing theoretical justification for an AI voter. However, formal voting power also creates institutional pressure: members may silence doubts to conform or resist if they perceive AI authority as illegitimate. Whether AI voting yields smarter decisions or division depends on AI performance and how the group faces disagreement.

Following Pattern-Oriented Modelling [7], we use simulation to preliminarily explore three research questions: How do allocations of AI authority within human groups shape collective accuracy, and what are the trade-offs in cascade risk? How do AI voting weights interact with institutional regimes to shape performance, and do critical thresholds emerge? Through which pathways (mechanical aggregation versus normative influence) do institutional regimes produce their effect?

2 METHODOLOGY

We use an agent-based model implemented in NetLogo [15] to examine how institutional authority and information structures shape collective performance when AI agents are embedded in bounded-rational human groups. ABM enables explicit specification of micro-level heuristics and institutional rules, and the observation of macro-level outcomes across counterfactual institutional designs, including non-linearities and interaction effects.

Institutional regimes. The model represents a group of humans making repeated discrete decisions with established ground truths. We cross two design dimensions to produce six conditions. Access structure is either centralised (only a designated leader receives and relays the AI recommendation) or distributed (each agent independently observes the recommendation with a probability). Authority type is either advisory (AI has no formal voting power), voting (AI participates in the collective decision with weight w_{AI}), or hybrid (combining both). This yields six regimes: no-AI baseline, advisor-centralised, advisor-distributed, pure-voter, hybrid-centralised, and hybrid-distributed.

Agents and parameters. Humans are heterogeneous along four dimensions: private signal quality (probability that a human's private signal matches the ground truth), trust in AI, conformity propensity, and confidence. The AI agent is characterised by task accuracy and (in voter and hybrid regimes) vote weight w_{AI}. Parameter sweeps vary AI accuracy, the distributions of trust and conformity, group size, private signal quality, access coverage, and w_{AI}.

Sequence of play. In each round, the environment draws a ground truth. Humans receive private signals that match the truth with probability defined by their signal quality; the AI provides a recommendation correct with probability defined by its task accuracy. Humans form beliefs via a bounded-rational updating rule integrating four inputs: their private signal, the AI recommendation weighted by trust, informational influence (adjustment toward the group's mean expressed belief), and normative influence (adjustment toward perceived authority weighted by conformity propensity). In advisor regimes, outcomes are determined by aggregating human votes only; in voter and hybrid regimes, the AI casts a formal vote aggregated with human votes using w_{AI}. After each decision, T_i is updated based on AI correctness and decision outcomes. The flowchart for the complete model procedure is detailed in Appendix A.

Measures. The primary measures are collective accuracy (proportion of correct decisions), error cascade rate (proportion of incorrect rounds in which >66.7% of agents voted for the wrong answer), and false convergence (proportion of all rounds yielding a high-consensus incorrect decision). Across regime and parameter settings, the model contains a total of 1,113,600 runs. Different combinations of parameters contain 30 repetitions each.

3 SIMULATED RESULTS

Regime Ranking. Across simulation runs per condition (n = 21,000 for advisor-centralised; 147,000 for advisor-distributed; 105,000 for pure-voter and hybrid-centralised; 735,000 for hybrid-distributed; 600 for no-AI), advisor-centralised produced the highest accuracy (M = 0.824, 95% CI [0.823, 0.825]), outperforming the no-AI baseline (0.797), advisor-distributed (0.794), pure-voter (0.787), hybrid-centralised (0.786), and hybrid-distributed (0.766). The imbalance of runs across regimes was due to the fact that some parameters were not applicable for that condition, reducing the number of combinations.

Error cascade mechanism. Cascade rate explains much of this ranking (Figure 1, top row). Centralised architectures and pure-voter exhibited low cascade rates (0.069–0.080), whereas distributed architectures showed rates approximately six times higher (hybrid-distributed: 0.447; advisor-distributed: 0.480). Advisor-distributed maintained relatively high accuracy despite its high cascade rate, suggesting the advisory role partially offset error propagation.

Vote weight. Introducing the AI as a voting teammate produced a non-linear response to vote weight (Figure 1, bottom row). When AI voting weight (w_{AI}) exceeded approximately 0.25–0.30, accuracy declined from 0.797 to 0.766, and further to 0.752 at 0.425. False convergence decreased across bins from 0.081 to 0.071 monotonically, indicating that higher AI weight reduced incorrect convergence but did not improve overall accuracy.

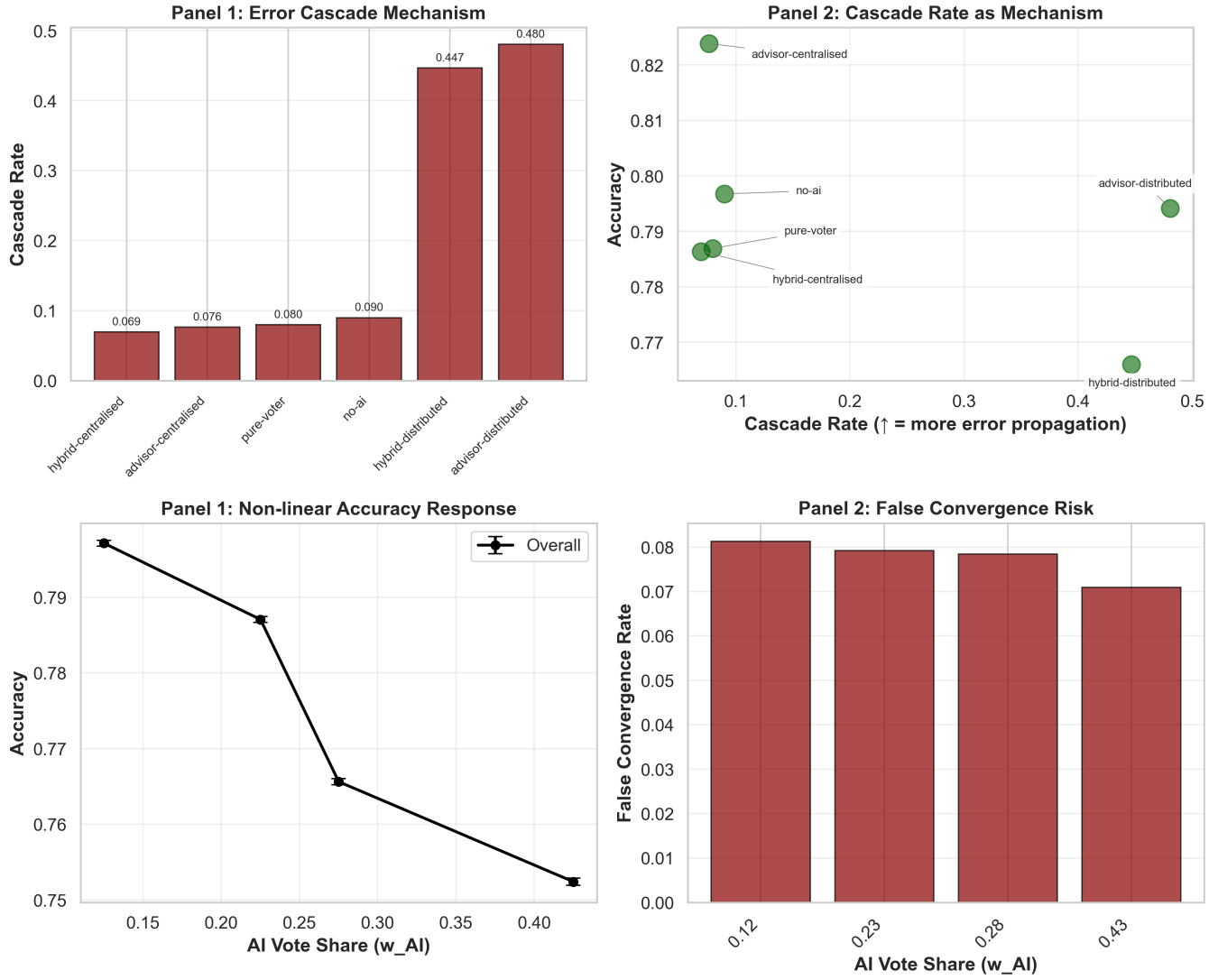

Figure 1: Top: Error cascade mechanism, showing cascade rates by regime and the relationship between cascade rate and accuracy; distributed architectures suffer substantially higher error propagation. Bottom: Effect of AI vote weight (w_{AI}); accuracy declines non-linearly beyond $w_{AI} \approx .25$, while false convergence decreases monotonically with increasing vote share.

4 DISCUSSION

Our agent-based model serves as a generative laboratory for examining which institutional configurations produce accuracy differences. Only advisor-centralised (0.824) exceeds the no-AI baseline (0.797); the remaining four hybrid configurations match or fall below it, despite our assumption that AI accuracy ranges from 60% to 90%. This is consistent with a recent meta-analysis of 106 experiments finding that human–AI combinations perform worse on average than the best single agent, with especially pronounced losses where AI already exceeds human accuracy [14].

Our 2×2 design crossing access structure with authority type reveals that both dimensions matter, but authority type produces larger effects. Switching from advisory to voting reduces accuracy by 0.033 on average; switching from centralised to distributed reduces it by 0.025. Cascade rate explains much of this: centralised regimes exhibit rates of 0.069–0.080 versus 0.447–0.480 in distributed regimes, a sixfold difference. When a single leader mediates AI recommendations, the leader functions as a circuit breaker, processing recommendations through human judgment before they trigger conformity cascades. This is consistent with work on information cascades, where early signals propagate through networks and suppress independent evaluation [1].

However, cascade rates alone do not explain the full hierarchy. Advisor-distributed has the highest cascade rate (0.480) yet maintains reasonable accuracy (0.794), while hybrid-distributed has a lower rate (0.447) yet produces the worst accuracy (0.766). This dissociation reveals a compounding interaction: in hybrid regimes,

AI casts a weighted vote that directly shifts group outcomes, while a hidden normative channel pulls agents without AI access toward the AI recommendation through conformity pressure (weight 0.2). This layering of normative conformity [3] on top of informational influence suppresses remaining critical evaluation capacity. Recent work documents less diverse outcomes when AIs participate in voting in AI-only settings [16]; formal models show that high-weight voter competence determines collective accuracy [10], but our simulation suggests this benefit depends on the surrounding information access structure.

Across all vote-weight conditions, groups reach consensus faster than they reach accuracy. The slight decline at higher weights likely reflects wholesale deference rather than genuine deliberation: groups agree more, but not because they reason better.

However, these findings hold under specific scope conditions: a difficult task (human signal quality barely above chance), comparatively reliable AI, and asymmetric trust updating that penalises AI errors more heavily than correct predictions. Our agents are homogeneous in critical thinking and initial trust; heterogeneity may shift cascade thresholds. A further limitation is that existing empirical literature provides insufficient data to calibrate key behavioural parameters, including trust initialisation, trust update rates, conformity weights, and social influence decay, to precise numerical values. Our parameter choices are informed by theoretical principles and directional findings rather than empirically measured magnitudes, meaning the specific accuracy values should be interpreted as ordinal rankings rather than point estimates. Importantly, agent-based models generate sufficient rather than necessary explanations [5]: different micro-rules might produce similar macro-outcomes. Future experiments should test whether centralised AI-informed committees outperform distributed voting with real teams, and use large-scale behavioural measurement to empirically calibrate the dynamic cognitive parameters, such as trust updating functions, conformity thresholds, and influence susceptibility, that currently rely on theoretical assumptions in our model. Also, these data will be valuable to revise the modelling to simulate larger and more complex situations.

References

[1] Sushil Bikhchandani, David Hirshleifer, Omer Tamuz, and Ivo Welch. 2024. Information cascades and social learning. *Journal of Economic Literature* 62, 3 (2024), 1040–1093.

[2] Dan Braha and Marcus AM de Aguiar. 2025. Generalizing Condorcet's Jury Theorem to Social Networks. *arXiv preprint arXiv:2510.16808* (2025).

[3] Morton Deutsch and Harold B Gerard. 1955. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology* 51, 3 (1955), 629.

[4] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.

[5] Joshua M Epstein. 1999. Agent-based computational models and generative social science. *Complexity* 4, 5 (1999), 41–60.

[6] Suyash Fulay, Sercan Demir, Galen Hines-Pierce, Hélène Landemore, and Michiel Bakker. 2025. Shareholder Democracy with AI Representatives. *arXiv preprint arXiv:2510.23475* (2025).

[7] Volker Grimm, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M Mooij, Steven F Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L DeAngelis. 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *science* 310, 5750 (2005), 987–991.

[8] Mo Hu, Guanglu Zhang, Leah Chong, Jonathan Cagan, and Kosa Goucher-Lambert. 2025. How being outvoted by AI teammates impacts human-AI collaboration. *International Journal of Human–Computer Interaction* 41, 7 (2025), 4049–4066.

[9] Bryan D Jones. 1999. Bounded rationality. *Annual review of political science* 2, 1 (1999), 297–321.

[10] Ines Lindner. 2008. A generalization of Condorcet's Jury Theorem to weighted voting games with many small voters. *Economic Theory* 35, 3 (2008), 607–611.

[11] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[12] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.

[13] Kexin Quan, Dina Albassam, Mengke Wu, Zijian Ding, and Jessie Chin. 2025. Towards AI as Colleagues: Multi-Agent System Improves Structured Professional Ideation. *arXiv preprint arXiv:2510.23904* (2025).

[14] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303.

[15] Uri Wilensky. 1999. *NetLogo.* Evanston, IL. http://ccl.northwestern.edu/netlogo/

[16] Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. 2024. LLM Voting: Human Choices and AI Collective Decision-Making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 1696–1708. doi:10.1609/aies.v7i1.31758

[17] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. 2023. Investigating AI teammate communication strategies and their impact in human-AI teams for effective teamwork. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–31.

A MODEL PROCEDURES AND REPORTERS

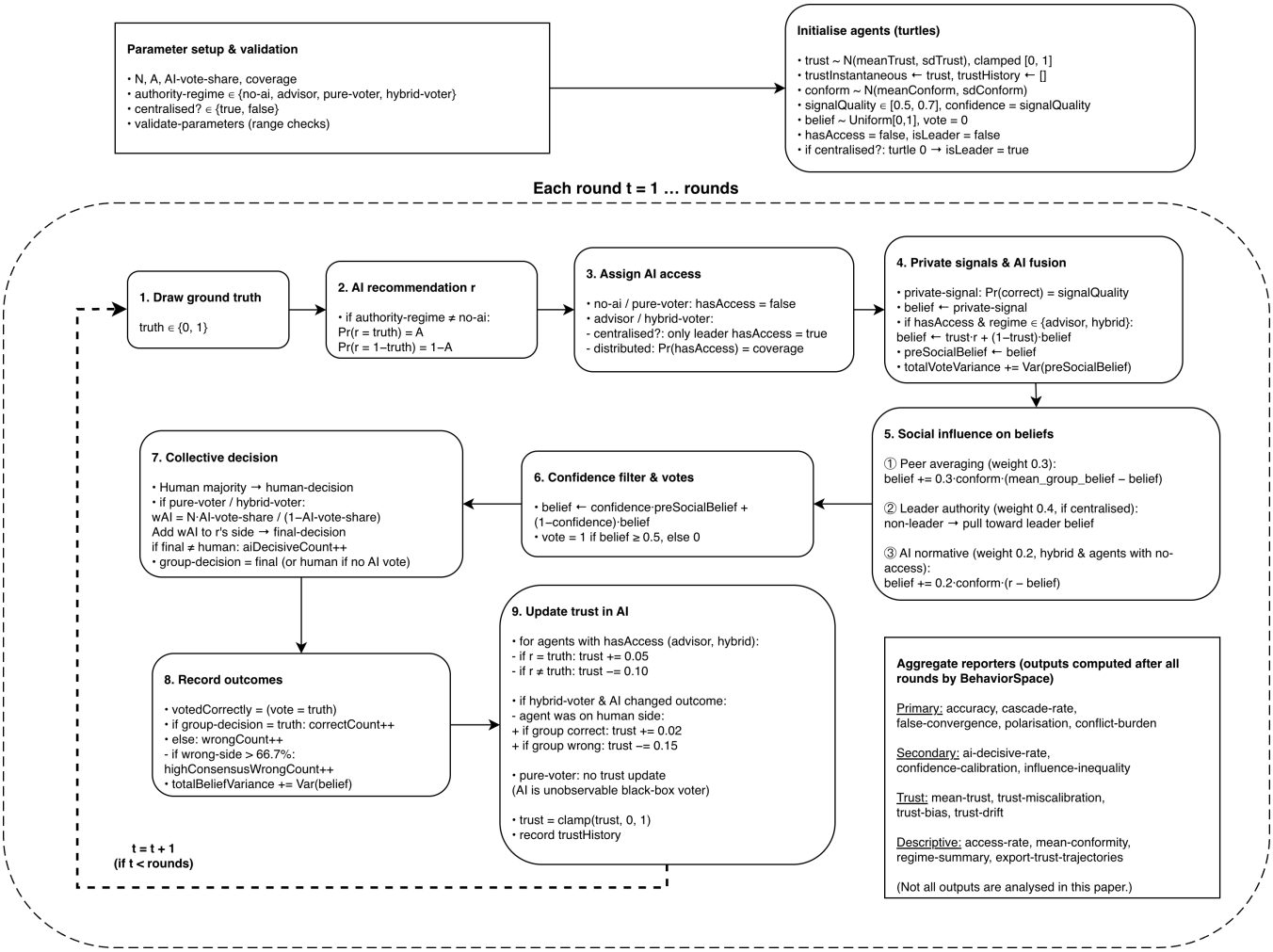

Figure A1: Agent-Based Model procedure.