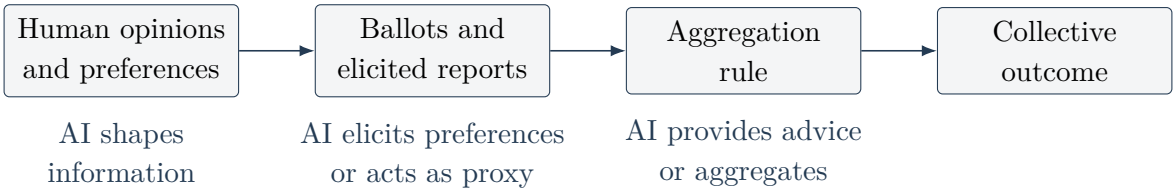
Foundations of Collective Decision Making in a World of Artificial Agents

From elections and public budgeting to school placements and resource allocation, collective decisions translate diverse interests into shared outcomes. Social choice provides a mathematical language for asking when this translation can be fair and efficient, and for proving when it cannot. Its axioms and quantitative guarantees raise fundamental questions about algorithms, approximation, incentives, communication and computational complexity.

Classically, people form opinions and preferences; elicitation converts them into ballots such as top choices, rankings or approvals; and an aggregation rule maps those reports to an outcome. Artificial intelligence can now intervene throughout this process: it can shape the information people receive, elicit preferences, act as a proxy, or advise an aggregation rule. Such systems can reduce the burden of expressing complex preferences, but they may misunderstand users, introduce bias, behave strategically, or make correlated mistakes when many people rely on a shared model. Classical assumptions therefore need to be re-examined under artificial mediation.



Artificial agents are therefore not merely new voters. In auto-bidding, advertisers state high-level objectives while automated systems choose bids across auctions, changing equilibrium and welfare analysis relative to classical models [1]. The Bank of England likewise warns that shared AI models in trading may induce correlated behaviour and herding [2]. These examples illustrate how individually useful artificial agents can alter collective outcomes.

Recent work studies LLM proxies [3], learning-augmented social choice [4], and social-choice approaches to alignment [5, 6]. My programme asks a distinct, compositional question: what happens to classical guarantees when artificial agents intervene between human preferences and collective outcomes? I will trace how error, bias, strategic behaviour and computational limits at successive stages propagate to welfare, fairness and representation. For bounded adversarial perturbations, I will seek worst-case composition bounds. For shared-model errors, I will specify stochastic dependence models and derive expected and high-probability guarantees. Together, these analyses will identify when losses add across stages and when dependence amplifies them.

Research Agenda

My work on distortion and incomplete preference information provides tools for measuring outcome loss, while my work on autobidding and public-spirited voting analyses how departures from classical behavioural assumptions change welfare and equilibrium. The programme focuses these tools on three connected directions.

Robust foundations for mediated choice. I will determine which classical guarantees survive when preferences or recommendations are incomplete, perturbed, biased or computationally restricted. The first objective is a composition theory: if mediator S introduces error ε_S under an appropriate preference metric, when does final loss grow smoothly with the ε_S, and when is any such bound impossible? I will seek algorithms with tight degradation guarantees and matching lower bounds, beginning with voting and participatory budgeting. I will then study rules that receive both reported preferences and an AI recommendation. The objective is an optimal trade-off between consistency, the

gain when advice is accurate, and robustness, the worst-case loss when it is unreliable. I will extend these guarantees to graph-structured objectives that capture complementarities among projects or committee members, and to repeated decisions in which advice is drawn from a shared distribution.

Preference formation, elicitation and delegation. A citizen choosing among many public projects might describe priorities in everyday language and answer a few questions before an LLM proxy submits a ballot. I will first treat people as having fixed latent preferences and ask how accurately a proxy can recover the information needed for a good collective decision. In structured domains such as metric voting, I will characterise when a full ranking over m alternatives can be replaced by $O(\text{polylog}\, m)$ adaptive questions while retaining a constant-factor welfare or representation guarantee, and prove lower bounds when this is impossible. I will then allow AI-provided information to change preferences rather than merely report them. Axioms for neutral advice, equal treatment and bounded influence will distinguish legitimate assistance from systematic manipulation. The resulting models will quantify how unequal model quality, strategic answers and shared bias affect query complexity and collective welfare.

Social choice for AI alignment. When an AI system affects several people, their conflicting judgments about its behaviour must be aggregated. Existing work develops axiomatic and linear social-choice models for this problem [5, 6]. I will study alignment when feedback is limited, noisy or unevenly distributed and the output space is too large to enumerate. How many adaptive comparisons are required to achieve welfare close to a full-information benchmark while satisfying unanimity or proportional representation? I will characterise communication–distortion and axioms–welfare trade-offs using approximation algorithms and lower bounds, and connect them to the composition results from the first direction.

Outcomes, feasibility and impact

Years 1–2 will develop the compositional framework and tight robustness bounds; years 2–3 will address proxy elicitation and bounded influence; and years 3–5 will extend the theory to alignment and repeated decisions. Each project has a positive algorithmic or axiomatic target and a corresponding lower-bound or impossibility target, so informative results remain possible when the strongest guarantee fails. The core programme is theorem-driven and models artificial agents through general properties rather than assumptions tied to present-day LLMs. Where useful, simulations and real-world data will test modelling assumptions rather than replace mathematical guarantees. The resulting theory will provide principled ways to evaluate AI-mediated systems through the quality of their collective outcomes.

Selected references

[1] Yuan Deng et al. Towards efficient auctions in an auto-bidding world. In *The Web Conference*, 2021.

[2] Bank of England. Financial stability report: July 2026. Bank of England, 2026.

[3] David Huang et al. Accelerated preference elicitation with LLM-based proxies. arXiv:2501.14625, 2025.

[4] Ben Berger et al. Learning-augmented metric distortion via (p, q)-veto core. In *ACM EC*, 2024.

[5] Luise Ge et al. Axioms for AI alignment from human feedback. In *NeurIPS*, 2024.

[6] Zachary Wojtowicz et al. Algorithmic impact reveals the hidden social choice structure of alignment. arXiv:2608.24046, 2026.