



University of
Applied Sciences
St. Pölten

AI of the People, by the People, for the People

Social Choice Theory and Collective Control of AI

Lukas Daniel Klausner

15 April 2026

Research Seminar

University of Applied Sciences St. Pölten

(joint work with Paul Anton Bachmann, Niclas Böhmer and Martin Lackner)

Problem Statement

How many AI technologies do you **use** in a typical (work)day?

And how many of those do you have any **relevant influence** on?

This motivated us to approach this problem from a novel viewpoint – by applying social choice theory towards collective decision-making about AI systems.

Basics of Social Choice Theory

social choice theory: mathematical/economical study of collective decision-making and preference aggregation

typical assumption: voters have (total) preference ranking of (all) possible choices; preference aggregation rules then define how to turn these into a collective preference order

based on **axiomatic approach**: defining **normative axioms** with properties one would like aggregation rules to have and then **studying various systems** w. r. t. these axioms

Key Results of Social Choice Theory

most well-known: [Arrow's Impossibility Theorem](#), i. e. incompatibility of even three relatively straightforward such axioms – no universal aggregation function can fulfil non-dictatorship, Pareto efficiency and independence

more generally: collective decision-making is [highly dependent](#) on the choice of preference aggregation rules

Our Main Contributions

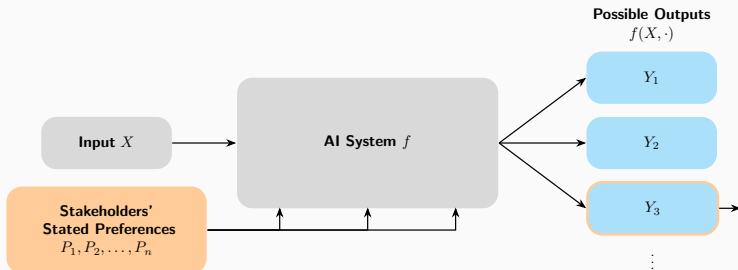
(1) How can we talk about the **alignment** of AI systems with **collective preferences**?

⇒ framework for AI systems as **social choice functions**

(2) What are possible ways to **influence the development** of AI systems through collective decision-making?

⇒ decisions about AI development as **instances of social choice problems**

AI Systems as Social Choice Functions



using social choice theory as a lens for collective control of AI:

- * influence of collective input on AI output
- * elicitation and formalisation of stakeholder preferences
- * transfer of and reasoning about relevant social choice axioms

Example

AI system f : LLM-based chatbot to help secondary school students identify university education programs

set of stakeholder groups P_i providing preferences:

$P_1 =$ "Provide information about public and private institutions." (educational authorities)

$P_2 =$ "Include scholarship opportunities and admission requirements." (parents' associations)

P_3, \dots, P_{10} are educ. institutions, each with a preference for its own agenda.

given input $X =$ "What can I study if I like playing and thinking about video games?", f may respond with $Y = f(X, P_1, \dots, P_{10}) =$ "Here are the top three CS degrees in the country: ..."

Y can now be evaluated (with appropriate methods) w. r. t. how well it respects stakeholder preferences

Some Axioms

Anonymity

- * “Every voter is treated equally.”
- * rules out dictatorial, but also weighted social choice functions
- * applicability to AI system questionable: weighting stakeholders differently might be apt (and ignoring all stakeholder input vacuously fulfils the axiom)

Participation

- * “Participation must never be disadvantageous.”
- * so-called “no-show paradox”, well-known in social choice theory (consider participation thresholds in elections or referenda)
- * far from trivial for AI systems: systems might (and already do) use stated preferences to further their core objectives, which might run counter to stakeholders’ interests

Yet More Axioms

Unanimity

- * “If all unanimously agree on a candidate, that candidate should win.”
- * relatively weak, often strengthened to . . .

Majority Criterion

- * “If a majority agree on a candidate, that candidate should win.”
- * considering broad range of inputs, not that helpful for AI systems
- * adaptations might consider restricting candidates/output by (partially) fixing the input

Pareto Efficiency

- * “If all voters weakly prefer Y_1 over Y_2 and at least one voter strictly prefers Y_1 over Y_2 , then Y_2 should not win.”
- * again, requires adaptations and restrictions for AI system contexts

from binary axioms to evaluation criteria: [sampling possible inputs and outputs](#)

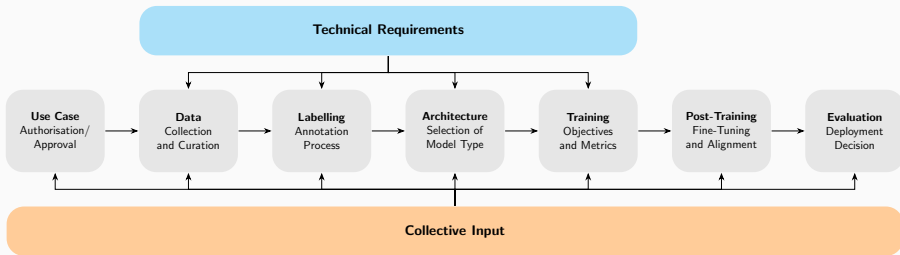
Influencing AI Development Decisions

we sketched out a prototypical, abstracted ML development pipeline ...

... and then considered how to feasibly influence each stage therein through collective preferences ...

... connecting relevant problems for each stage with corresponding issues in social choice theory

Prototypical ML Development Pipeline



ML Development Stages, Part 1

Authorisation for Use Case

- * will often require contextual, contingent criteria
- * stakeholder preferences could be modelled as minimal requirements to be met for approval
- * SCT connections: multi-winner voting to select set of target metrics, median-based procedures to calibrate numerical thresholds

Data Collection and Curation

- * issues include selecting representative data sets, excluding harmful data, choosing data collection modes/sites
- * stakeholders are unlikely to judge individual data points, but rather overall dataset properties
- * SCT connections: portioning procedures (social decision schemes for selection, approval ballots for exclusion, . . .)

ML Development Stages, Part 2

Labelling and Annotation

- * annotation of large datasets with meaningful labels requires resolving annotator disagreement
- * SCT connections: eliciting annotator uncertainty (through ambiguous annotations) and aggregating diverse preferences over 'correct' labels ('soft' labels, label distributions)

Selection of Model Architecture and Training Parameters

- * most technical of the steps
- * difficulty of reconciling required expertise with relevant public concerns (e. g. ecological)
- * SCT connections: eliciting minimum requirements (e. g. w. r. t. efficiency or system capabilities) or vetoing undesirable model architectures

ML Development Stages, Part 3

Fine-Tuning and Alignment

- * probably the part of the process already most closely aligned with SCT approaches, but improvements are possible
- * RLHF already widely used for aligning LLMs – social choice mechanisms have been proposed to improve feedback aggregation (RLCHF)
- * constitutional AI approaches often presuppose the existence of the ‘constitution’ – multi-winner voting and other SCT approaches could enable collective value selection (‘constitution-drafting’)

Refinement of Model Behaviour at Runtime

- * implementing collective decision-making at runtime mostly infeasible
- * collective input could be aggregated into abstract principles *ex ante*
- * SCT connections: integrating conflicting values, selecting data input sources (cf. data collection and curation), satisfying fairness axioms over time (temporal fairness)

Summary

framing AI systems as **implementations of social choice functions** and applying established social choice metrics for the **evaluation of their alignment** with stakeholders' collective preferences

understanding AI development as a **series of decision problems** (often matching well-known social choice problems) and using techniques from social choice theory to facilitate **collective control over these decisions**

Outlook

open questions and future work:

- (1) stakeholder identification and weighting
- (2) application to practical examples
- (3) preference elicitation from non-experts on technical issues
- (4) connection of axiomatic approach with algorithmic auditing

already under review: [Democratise AI! Ensuring Collective Control of Artificial Intelligence in Real-World Contexts](#) (proposal for the WWTF's ICT26 call on digital humanism; jointly with the Institute of Logic and Computation at TU Wien, the Institute of Technology Assessment at the ÖAW and the Austrian Parliament)

Credits

to appear (soon) as:

Paul Anton Bachmann, Niclas Böhmer, Lukas Daniel Klausner and Martin Lackner: *AI of the People, by the People, for the People: A Social Choice Approach to Collective Control of Artificial Intelligence*. Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), pages TBD, [doi:10.1145/3805689.3806808](https://doi.org/10.1145/3805689.3806808).