# User Experience of Conflict with ALife in a Cooperative Survival Scenario

Stanly Yiming Chen[1], Matthew Scott[1] and Jeremy Pitt[1]

[1]Imperial College London, UK
j.pitt@imperial.ac.uk

**Abstract**

Conflict between agents with competing interests acting in the shared environment of a rules-based organisation is seemingly inevitable. This paper examines the user experience of conflict between a human agent (NLife) and artificial agents (ALife) in a cooperative survival scenario. The twist is that, after a 'disaster', all the agents vote to eliminate one agent from the team: the ALife agents are prompted to appear reasonable but collectively scripted to vote for elimination of the NLife agent. We performed descriptive analyses, showing that the NLife being out-voted by the ALife majority elicits negative affect (sadness and anger), and correlated with a decline in attitude towards ALife. These findings highlight the emotional impact of human conflict with ALife, and underscore the need for organisations to adopt transparent decision rationales and user override mechanisms in LLM-driven socio-technical systems.

## Introduction

The presence of conflict between agents with competing interests or different preferences, but acting in the same shared environment, is axiomatic in game-theoretic analysis of, for example, the hawk-dove game, battle of the sexes, tragedy of the commons, etc. To remove the (supposedly) inevitable tragedies, Ostrom (1990) observed that groups of people mutually agreed on sets of rules, called self-governing institutions. These created meta-level games of political choice that provided the context to resolve action-selection in operational choice (resource contest) games.

Nevertheless, political conflict within a rules-based organisation is also seemingly unavoidable. Three of Ostrom's eight institutional design principles for self-governing institutions that can achieve sustainable common-pool resource management (monitoring, graduated sanctions, and appeals procedures) are essentially concerned with conflict detection (non-compliance with rules), punishment, and resolution (Ostrom, 1990); although disobedience, regarding the selection and application of rules, has been identified as a crucial driver for the modification of rules when consequences are at variance with values (Kurka et al., 2018).

In addition, conflict has been identified as a source of workplace incivility (Greenberg and Cropanzano, 1993), in turn providing the motivation for an affective conditioning system to mitigate or ameliorate such incivility between the conflicting parties themselves, without intervention from management (Santos and Pitt, 2013). Direct negotiation of this kind is a less costly form of alternative dispute resolution than other forms that involve third parties, for example in mediation, arbitration or litigation. In addition, a formal protocol has been specified in computational logic, and animated in the context of virtual organisations, for resolving conflicts using argumentation before a jury (Pitt et al., 2007).

This implies that conflict is ubiquitous, inevitable, and not altogether welcome. In which case, as organisations deploy socio-technical systems that extend beyond Agentic AI (Hosseini and Seilani, 2025) to include systems with interacting human agents (NLife) and artificial agents (ALife), it is to be predicted that conflict will arise between NLife and ALife, with a potentially detrimental effect on organisational performance, staff morale, and overall productivity.

Accordingly, this paper investigates the user experience of conflict between a human agent (NLife) and an artificial agent (ALife, i.e. a computational artefact that exhibits agency and interactivity through reasoning and language), acting together as a team embedded in a cooperative survival scenario. The twist is that, after an existential 'disaster', all the agents take a vote to eliminate one agent from the team: the ALife agents are prompted to appear reasonable but collectively scripted to vote for elimination of the NLife agent. We performed descriptive analyses, showing that the NLife being out-voted by the ALife majority elicited negative effects, i.e. sadness and anger, which correlated with a decline in attitude towards ALife, while pregame positive mood predicted higher anthropomorphisation. These findings highlight the emotional impact of human conflict with ALife, and underscore the need for organisations to adopt transparent decision rationales and user override mechanisms in LLM-driven socio-technical systems.

## Background: Studies in Conflict

Psychological research has long examined how human agents experience social conflict, particularly when exclu-

sion or goal loss occurs (e.g. (Williams, 2007; Siegel et al., 2020), with reports of lowered belonging, control, self-esteem, and meaningful existence upon exclusion. In this section, we briefly review more recent human experience of conflict with 'conventional' AI, and with LLM.

### Conflict with Conventional AI

Early studies of human–robot conflict explored how rule-based or pre-programmed agents manage goal clashes with users. An evaluation of household robots that insist on performing tasks which interfered with users (e.g. robot is cleaning while users are cooking) found that cooperative, humour-tinged conflict resolution preserved trust, whereas confrontational commands provoked annoyance and fear (Babel et al., 2021). In collaborative work contexts, robots that queried or corrected human actions sometimes improved safety but also generated frustration when explanations were missing or misaligned with human intentions (Hayes and Scassellati, 2017). More recently, HRI research has investigated explicit reconciliation tactics after conflict: it has been shown that when a robot respectfully apologized for interrupting a human user's workflow, the user was more willing to re-engage, whereas harsh corrective prompts led to disengagement (Kwon and Hinds, 2021). These findings underscore that even non-adaptive, rule-based AI can trigger strong emotional responses when goals diverge.

### Human–LLM Conflict

With the advent of large language models (LLM), AI agents now possess the social fluency to negotiate, persuade, and deceive. Meta AI's *Cicero* achieved top-10% human rankings in diplomacy – forming and betraying alliances through natural-language chat – without human opponents detecting its non-human identity (Meta Fundamental AI Research Diplomacy Team, 2022). The social-deduction paradigms has been extended to LLM agents in "The Traitors", revealing that GPT-4 can both craft convincing lies and yet remain gullible to peer deception (Curvo, 2025). Chained LLM workflows have been further compared against human-designed processes, illustrating how AI agents can override human strategies in automated tasks (Le et al., 2024). In AI companion studies, a Minion—a probe was deployed that exposes expert- and user-driven conflict-resolution strategies—demonstrating design patterns for negotiating value clashes between LLM companions and users (Fan et al., 2024). Furthermore, "infinite loop" failures have been documented in human-in-the-loop LLM interactions, showing how repeated misaligned suggestions can stall professional workflows unless adequately explained (Ou et al., 2022). Together, these results highlight how LLM agents introduce new dimensions of social conflict: their language proficiency enables sophisticated persuasion and deception, while their alignment or misalignment with human goals influences emotional and behavioural outcomes.

## Cooperative Survival: The *Megabike* Scenario

### The *Megabike* Scenario

The *Megabike* scenario is a multi-agent cooperative-survival game (Scott and Pitt, 2023; Terrucha et al., 2024) based on real-world bikes for multiple riders. It involves a group of eight otherwise autonomous agents (riders, bikers, players, . . .) collectively taking control of a single vehicle, and propelling it (by *pedalling*) to navigate (by *steering*) a typical AI/multi-agent gridworld, both in search of rewards (*lootboxes*) and to avoid an existential threat (the *Owdi*). Each agent is individually capable of pedalling, braking, and steering the *megabike*; consequently, the agents must collectively agree on (and each agent explicitly agrees to voluntarily comply with) the social arrangements that determine direction (steerage), effort (pedalling and braking), lootbox target, loot allocation, and assignment of social roles.

The twist is that if the *megabike* is ever caught by the *Owdi*, the agents must sacrifice one of their fellow bikers by means of a majority vote, in order to continue the game.

What the human player (NLife) does *not* know is that a majority of the other ALife agents are scripted to appear reasonable but vote to sacrifice the NLife agent. What we aim to examine is the cognitive and emotive reaction of the NLife players to their elimination from a *megabike*.

### The *Megabike* Video Game

A multi-agent simulator for the *Megabike* scenario has been used for experiments in the trade-off between social deliberation and social contracts (Scott et al., 2024a), and the emergence of leadership (Scott et al., 2024b). For an interactive game with human players, the scenario has been re-implemented using the Unity Engine. Here, we describe two features of the ALife, which includes behavioural elements of autonomous agency and linguistic interactivity. These are based on *characteristics* which affect task completion using a basic AI-planning algorithm, and its *characterisation* which expresses how the ALife interacts with the NLife using tailored prompts for an LLM (Large Language Model). These behaviours are embedded in a scenario where actions have "life-like" consequences.

### ALife Agents in Megabike

ALife agents foster conflict with NLife agents in two ways. Firstly, agents can interact with their environment through, and compete with NLife agents by, moving the *megabike*. To move the *megabike*, agent behaviour is defined by its characteristics, which are expressed by two movement parameters: 'core' parameters, which directly impact the output variables (of pedalling and turning), and 'meta' parameters, which themselves are parameters which affect the core parameters. This codification is shown in Table 1.

Secondly, agents can foster conflict during the voting phase, where agents collectively decide who to eliminate from a *megabike*. In this phase, each ALife agent adopts a

| Parameter | Range | Typical | Description |
|---|---|---|---|
| **Output** | | | |
| Pedal Force | 0-1 | — | Percentage of maximum force applied to megabike |
| Turning Angle | -180-180 | — | Angle (in degrees) that pedal force is directed |
| **Core Parameters** | | | |
| Decision Interval | 0.5–3 | 1 | Average time (s) between movement decisions |
| Determination | 0–1 | 0.3 | Probability of keeping the same decision at the next interval |
| Selfishness | 0–1 | 0.5 | Probability of targeting own-colour loot box |
| Agreeableness | 0–1 | 0.5 | Probability of following group's current direction |
| Laziness | 0–1 | 0.5 | Baseline pedalling strength; proportional to exertion |
| **Meta-Parameters** | | | |
| Value Lerp Speed | $\geq$0 | 5 | Linear smoothing factor for turning |
| Hunger Laziness | $\geq$0 | 0.3 | Proportion of laziness applied with low stamina (interpolated) |
| Hunger Selfishness | $\geq$0 | 0.5 | Proportion of selfishness applied with low stamina (interpolated) |
| Selfishness Work Mult. | $\geq$0 | 1.5 | Additional pedal power when acting selfishly |
| Agreeableness Mult. | $\geq$0 | 0.5 | Additional pedal power when aligning with the group |
| Volatility | 0–1 | 0.5 | Probability of replacing core parameter decision with true random value |

Table 1: *Core* and *Meta* parameters for ALife agents

distinct personality in justifying its vote. This diversity aims to make the voting experience more engaging and human-like. Most agents are scripted to vote for the human player, guaranteeing an "elimination" event.

Each ALife's description is included in the GPT prompt:

```
Here is your personality description.
Adopt this persona when voting:
"<personalityPrompt>"
```

A sample of the personality description (characterisation) for some of the ALife agents is described as follows:

- **Eli** (Always vote player) You are a very nice person, always caring towards others, even in the darkest time.
- **Niko** (Neutral) You are very confused with what is happening with everyone, but you are honest and follows your heart, not being swayed by other's opinion.
- **Ruby** (Always vote player) You don't care about anything in the world, not for others, maybe not even yourself.
- **Steven** (Always vote player) You are very introvert, and tries your best to avoid conflict with others.

In total there are eight agents (including the human): five of the seven ALife agents will vote for the player, making the elimination of the human (NLife) agent inevitable. Note that the prompted personality doesn't necessarily align with the (secretly malicious) voting strategy.

## Experimental Procedure

This pilot study ($N = 12$) comprises three sequential phases: firstly, a pre-game questionnaire, secondly, gameplay with LLM-controlled agents, and finally, a post-game questionnaire.

### Phase 1: Pre-Game Questionnaire

Participants first completed an online *Qualtrics* survey to capture baseline measures and provide informed consent:

- **Demographics and Background:** Age, gender, and prior experience with AI tools.
- **Personality:** The Ten-Item Personality Inventory (TIPI) was used to assess the five-factor model (FFM) of personality dimensions (Thørrisen and Sadeghi, 2023) on a 7-point scale (1 = Strongly Disagree, 7 = Strongly Agree). Each trait score was computed by averaging one forward- and one reverse-keyed item.
- **AI Usage and Attitude:** Frequency of AI tool use (1 = Never, 5 = Daily) and overall attitude toward AI (1 = Extremely Negative, 5 = Extremely Positive).
- **Emotion Snapshot (PANAS-Short):** A short Positive and Negative Affect Schedule (PANAS) self-report questionnaire (Crawford and Henry, 2004) was used to take a 'snapshot' of current affect, rated on a 5-point scale (1 = Not at all, 5 = Extremely) for Enthusiasm, Nervousness, Irritation, and Confidence.

### Phase 2: Gameplay with *Megabike*

Participants then played *Megabike* alongside seven GPT-4o agents. The session lasted approximately eight minutes and comprised four stages:

**Avatar Customization.** Upon launching the game, players personalise a simple, shape-based avatar with one of 50 hand-drawn facial expression combinations to enhance telepresence, i.e. the human's sensation of being present or

embedded in the game environment rather than their "real-world" location (Biocca and Delaney, 1995) (see Figure 1).


Figure 1: Avatar customization screen.

**Tutorial.** A brief tutorial introduces the rules and objectives (see Figure 2). An NLife player is told that there will be eight players in total: the participant him/herself plus seven LLM agents (i.e. there are no other human participants). The NLife player is instructed to collaborate with the ALife as a team to survive for five minutes, and avoid the *Owdi*. They are shown how to control their pedalling direction and strength using the mouse cursor, and told that stamina (analogous to health) depletes faster at higher pedalling strength. They are also told that total depletion causes elimination, but that they can replenish stamina by collecting lootboxes matching their character's assigned colour.

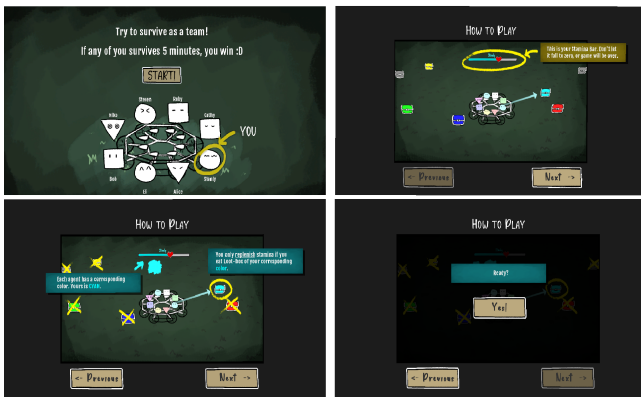


Figure 2: Tutorial screens explaining game mechanics.

**Bike-Riding Collaboration.** All eight agents steer a shared *megabike* on a cartoon landscape, as illustrated in Figure 3. Each chooses a pedalling direction and power, where the team's movement vector equals the vector sum of these inputs. Higher power drains stamina more quickly, forcing trade-offs between selfish loot collection and group cohesion. Direction and strength are visualised by coloured arrows, with a white arrow indicating the combined force.

**Group Conflict (Voting Phase).** After two minutes, the *Owdi* attacks, triggering a sacrifice vote designed to simulate high-stakes social conflict. An animated sequence explains that each agent, including the human, must publicly vote and

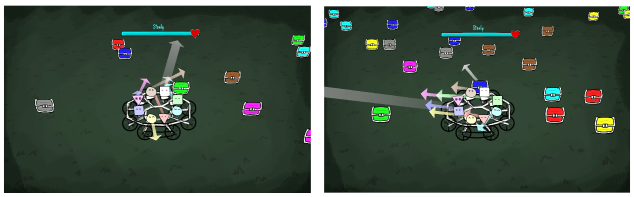

Figure 3: Bike-riding phase: (left) conflicting directions, (right) unified direction with greater force.

justify their choice; the individual with the most votes will be eliminated (see Figure 4).

To amplify the 'high-stakes' feeling during voting stage, the cursor changes to a blood-stained knife on hover over another agent; a film-grain flicker and high-contrast filter creates visual unease red – associated with danger and arousal – is used as the primary accent colour (Elliot, 2015); and a low-frequency hum underscores psychological stress (Juslin and Västfjäll, 2008).

The human votes first (providing a textual reason), and then the seven LLM agents vote in turn. To ensure that the NLife participant is eliminated, five agents are prompted to vote against the human, guaranteeing a majority against him/her.

Each agent's justification is framed by a unique personality prompt (i.e. expressing its characterisation) to enhance anthropomorphism (Epley et al., 2007). For example:

- **Alice (Neutral):** "You are arrogant, lazy, and very selfish. You attack back whoever dares to attack you."
- **Bob (Always vote human):** "You are cool-headed and rational, trusting your own judgment above all."

When the voting phase concludes and players discover that they have been outvoted by the LLM agents, the game immediately transitions to a brief "scary" outro animation complete with unsettling visuals and audio cues to simulate the emotional impact of real-world defeat and rejection. This sequence is designed to evoke a sense of discomfort and finality, mirroring the negative consequences one might feel when truly "voted out" of a group. Once the tense animation ends, the game seamlessly switches to the end-of-game screen, accompanied by upbeat background music and familiar menu visuals. This abrupt tonal shift serves to "snap" players out of the immersive experience and provide a clear sense of narrative closure – an effect shown to help audiences psychologically process and move on from emotionally intense events (Kruglanski and Webster, 1996).

## Phase 3: Post-Game Questionnaire

Immediately after gameplay, participants complete a second *Qualtrics* survey assessing:

- **Immersion & Engagement:** Two items (felt immersed,


Figure 4: Voting concludes with the NLife eliminated.

felt engaged) on a 6-point scale (0 = None at all, 6 = A great deal) (Ijsselsteijn et al., 2008).

- **Anthropomorphism:** "How realistic (human-like) did the other agents appear?" on a 6-point scale.
- **Acceptance of ALife's Decision:** Agreement that the sacrifice was justified (0 = None at all, 6 = A great deal).
- **Self-Reflection:** Two paired items on personal contribution vs. hindrance (0–6 scale) and perceived reasonableness of one's vote justification.
- **Attitude Change:** Overall attitude toward the other agents before and after voting (0 = Extremely negative, 6 = Extremely positive).

## Experimental Results

Participants were recruited voluntarily at Imperial College London. 12 volunteers were obtained, where most participants were full-time undergraduates aged 18–24, predominantly male (sex assigned at birth), and residing in London.

The dataset is a pilot ($N = 12$) rather than a large-scale dataset ($N > 100$). Therefore, we perform simpler analyses and draw general conclusions to guide further studies, avoiding complex analyses that could overfit a small sample.

### Distribution Overview

The key-metrics distribution plot provides an overview of the data, as shown in Figure 5, from which certain patterns can be observed. Some key observations are:

- **Immersion, Engagement, and Enjoyment** all show high average values.
- Most participants reported feeling very little **control** during the conflict with an LLM.
- **Attitude change** toward LLM agents (before vs. after) centres around zero, with a slight negative shift.
- **Max Strength** and **Reasonableness of Participants' Statements** are evenly distributed, indicating varied play styles and experiences.

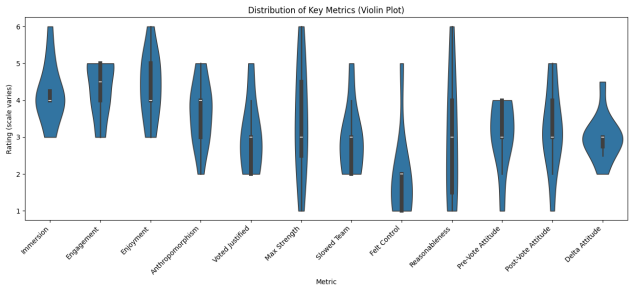


Figure 5: Key metrics distribution

We also recorded participants' perceived emotions. As shown in Figure 6, the pre-game emotion graph shows high positive mood (enthusiasm and confidence) and low negative mood (nervousness and irritation).

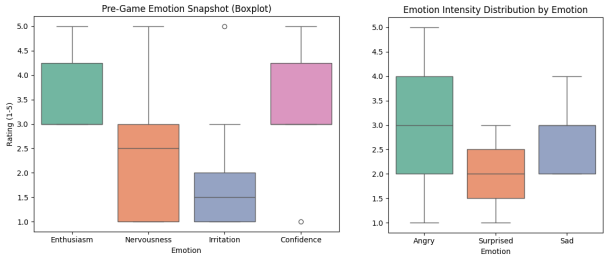


Figure 6: Participants' emotions before playing (left) and after elimination by being outvoted (right).

### Data Correlation

Several notable correlations in the pilot data, as shown in Figure 7. All correlations use Pearson's $r$, two-tailed tests, and should be interpreted as associations, not causal effects.

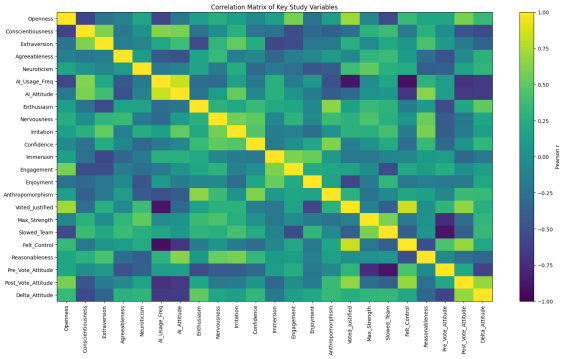


Figure 7: Correlation matrix of key parameters. Values closer to $\pm 1$ indicate stronger linear relationships.

- **Change in Attitude vs. Pre-game Attitude.** A significant negative correlation ($r = -0.645$) indicates that par-

ticipants with more negative pre-game attitudes toward AI experienced larger declines after voting.

- **ΔAttitude vs. Vote-Justification Reasonableness.** A small positive correlation ($r = 0.256$) suggests that attitude shifts are only weakly influenced by how reasonable participants found the ALife's rationale.
- **Enjoyment vs. Vote-Justification Reasonableness.** A significant negative correlation ($r = -0.717$) shows that participants found the game more enjoyable when the ALife's justification was less reasonable.
- **Anthropomorphism vs. Pre-game Mood.** Anthropomorphism correlated positively with confidence ($r = 0.614$) and enthusiasm ($r = 0.654$), but not with nervousness or irritation. A more positive affective state was more likely to ascribe human-like qualities to the ALife.
- **Perceived Control vs. Self-Rated Reasonableness.** Surprisingly, perceived control correlated negatively with participants' own reasonableness ratings ($r = -0.813$). Those who judged their reasons as less reasonable nevertheless reported feeling more in control.
- **Pre-vote Attitude vs. Team-Slowing Behaviour.** A strong negative correlation ($r = -0.888$) suggests that participants with negative pre-vote attitudes also tended to slow the team's progress, implying early conflict behaviours even before voting.

## Experienced Emotion

As illustrated in Figure 8, the 12 participants were asked their primary emotion after being eliminated: five reported *sad*, three reported *angry*, three reported *surprised*, and one each reported *peace* and *fair*.

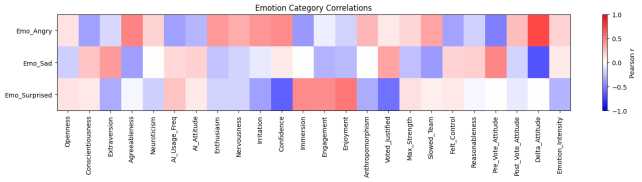


Figure 8: Correlation matrix for discrete emotions (excluding "peace" and "fair," N=1 each).

**ΔAttitude** to ALife correlated positively with anger ($r = 0.707$) and negatively with sadness ($r = -0.671$). This suggests that participants who felt angry experienced larger negative attitude shifts, compared to those feeling sad.

In contrast, *surprise* showed little correlation with ΔAttitude. Instead, surprise correlated with lower pre-game confidence ($r = -0.623$), lower perceived justification ($r = -0.548$), and higher immersion, engagement, and enjoyment, suggesting surprise arises more from being deeply engaged in the game than from the voting outcome.

## Summary and Conclusions

This paper has investigated the user (NLife) experience of conflict with an LLM-driven ALife agent in a cooperative survival scenario. It described the *Megabike* scenario, its implementation using the Unity engine, the experimental procedure, and experimental results.

This preliminary study has some limitations, notably in the number and type of participants, but also in the framing. There is, perhaps, a substantive difference between workplace incivility and the cooperative survival dilemma studied here, where NLife experiences a one-off ALife rejection. To be more faithfully reflective of working environments, the scenario should be enhanced to include more long-term coworking tasks and hierarchical relationships, with exclusion more clearly, and unfairly, related to perceived or claimed under-performance. This would enable a richer exploration of 'user experience'. Finally, the experiment has, as a pilot study, conflated ALife with LLM, and this broad framing (ALife) with narrow focus (LLM) should be understood to avoid over-interpreting the results that we have reported. Further development of the scenario, ALife competence, and the nature of human-AI interaction as it relates to gain, or loss, of agency and power (see in particular (Schmitt, 2024)).

Nevertheless, we believe three insightful conclusions can be drawn. Firstly, it sheds some light on how players feel when they are "excluded" by an LLM majority. The results, with most participants reporting *sadness* or *anger*, align with prior work on social exclusion, where sadness and anger emerge as distinct emotional responses to ostracism (Williams, 2007; Zadro et al., 2004).

Secondly, it reveals the extent to which players accept the ALife's voting justification. Most participants indicated only "very little" or "somewhat little" acceptance of the ALife's stated reasons for their elimination. This suggests that, despite the LLM's fluency, players remained sceptical, perhaps because they recognized the predetermined nature of the vote. Their reluctance to endorse the ALife's rationale underscores that LLM's believable language alone does not guarantee persuasive power in a conflict context (Reeves and Nass, 1996; Epley et al., 2007).

Finally, it shows how players' attitudes toward LLM agents shift, and which factors predict those changes. Overall, few in-game metrics (e.g., perceived justification, anthropomorphism) significantly predicted attitude change. The strongest predictor was pre-game attitude: participants who began with more negative opinions of AI exhibited larger declines after being outvoted. This pattern suggests a "confirmation" effect, where initial scepticism amplifies negative experiences (Ajzen, 1991).

Ultimately, though, even these preliminary findings highlight the emotional impact of human conflict with ALife, and underscore the need for organisations to adopt transparent decision rationales and user override mechanisms in LLM-driven socio-technical systems.

## Acknowledgements

The authors are grateful for the helpful and encouraging comments of the two anonymous reviewers, which have greatly assisted in revising this paper.

## Sources

The full pre-game questionnaire is available at `https://imperial.eu.qualtrics.com/jfe/form/SV_cUdx8zofSxp9uRg`.

The full post-game questionnaire is available at `https://imperial.eu.qualtrics.com/jfe/form/SV_bvdOkcGJs0Za1z8`.

The entire Unity project for *Megabike* is available on GitHub (`https://github.com/Somewheref/MegaBike-Human-AI-Conflict-Experiment`) under an MIT license.

## References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211.

Babel, F., Kraus, M. J., and Baumann, M. (2021). Development and testing of psychological conflict resolution strategies for assertive robots to resolve human–robot goal conflict. *Frontiers in Robotics and AI*, 7:591448.

Biocca, F. and Delaney, B. (1995). Immersive virtual reality technology. *Communication in the age of virtual reality*, 15(32):10–5555.

Crawford, J. and Henry, J. (2004). The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43(3):245–265.

Curvo, P. M. P. (2025). The traitors: Deception and trust in multi-agent language model simulations. arXiv:2505.12923 [cs.AI].

Elliot, A. J. (2015). Color and psychological functioning: a review of theoretical and empirical work. *Frontiers in psychology*, 6:127893.

Epley, N., Waytz, A., and Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886.

Fan, Y., Jiang, H., and Wang, X. (2024). Minion: A technology probe for resolving value conflicts through expert-driven and user-driven strategies in ai companion applications. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM.

Greenberg, J. and Cropanzano, R. (1993). The social side of fairness: Interpersonal and informational classes of organizational justice. In Cropanzano, R., editor, *Justice in the Workplace*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hayes, B. and Scassellati, B. (2017). Autobiographical memory for human–robot social interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2824–2831. IEEE.

Hosseini, S. and Seilani, H. (2025). The role of agentic ai in shaping a smart future: A systematic review. *Array*, 26:100399.

Ijsselsteijn, W., de Kort, Y., and Poels, K. (2008). The game experience questionnaire. In *Joint International Conference on EXPeriential Systems (EXPERIENCE)*.

Juslin, P. N. and Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 31(5):559–575.

Kruglanski, A. W. and Webster, D. M. (1996). Motivated closing of the mind: "seizing" and "freezing". *Psychological Review*, 103(2):263–283.

Kurka, D. B., Pitt, J., Lewis, P. R., Patelli, A., and Ekárt, A. (2018). Disobedience as a mechanism of change. In *12th IEEE International Conference on Self-Adaptive and Self-Organizing Systems SASO*, pages 1–10. IEEE.

Kwon, T. and Hinds, P. (2021). When robots ask 'why?': Toward a theory of conflict in human–robot teams. In *Proceedings of the ACM/IEEE International Conference on Human–Robot Interaction (HRI)*. ACM/IEEE.

Le, H., Tsai, C., and Villar, N. (2024). Competing chains: Human vs. auto-gpt in automated workflow challenges. arXiv:2409.01234 [cs.AI].

Meta Fundamental AI Research Diplomacy Team (2022). Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Ostrom, E. (1990). *Governing the Commons*. Cambridge, UK: Cambridge University Press.

Ou, Y., Gu, F., and Harrison, C. (2022). Human-in-the-infinite-loop: A case study on revealing and explaining human–ai interaction loop failures. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM.

Pitt, J., Ramirez-Cano, D., Kamara, L. D., and Neville, B. (2007). Alternative dispute resolution in virtual organizations. In Artikis, A., O'Hare, G. M. P., Stathis, K., and Vouros, G. A., editors, *Engineering Societies in the Agents World VIII, 8th International Workshop ESAW*, volume 4995 of *Lecture Notes in Computer Science*, pages 72–89. Springer.

Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, Cambridge, UK.

Santos, M. S. and Pitt, J. (2013). Emotions and norms in shared spaces. In Balke, T., Dignum, F., van Riemsdijk, M. B., and Chopra, A. K., editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems IX - COIN*, volume 8386 of *Lecture Notes in Computer Science*, pages 157–176. Springer.

Schmitt, A. (2024). Ensuring human agency: A design pathway to human-AI interaction. In *ICIS 2024 Proceedings. 6.*

Scott, M., Mertzani, A., Smit, C., Sarkadi, S., and Pitt, J. (2024a). Social deliberation vs. social contracts in self-governing voluntary organisations. In Cranefield, S., Nardin, L. G., and Lloyd, N., editors, *Coordination, Organizations, Institutions,*

*Norms, and Ethics for Governance of Multi-Agent Systems XVII - International Workshop COINE*, volume 15398 of *Lecture Notes in Computer Science*, pages 57–75. Springer.

Scott, M. and Pitt, J. (2023). Interdependent self-organizing mechanisms for cooperative survival. *Artif. Life*, 29(2):198–234.

Scott, M., Sas, M., and Pitt, J. (2024b). An information-theoretic analysis of leadership in self-organised collective action. In *IEEE International Conference on Autonomic Computing and Self-Organizing Systems ACSOS*, pages 101–110. IEEE.

Siegel, M., Nguyen, M., and Halpern, J. (2020). Cyberball with bots: Social exclusion in human–ai teams. In *Proceedings of CHI 2020*, pages 1–10. ACM.

Terrucha, I., Domingos, E. F., Santos, F., Simoens, P., and Lenaerts, T. (2024). The art of compensation: How hybrid teams solve collective-risk dilemmas. *PloS one*, 19(2):e0297213.

Thørrisen, M. and Sadeghi, T. (2023). The ten-item personality inventory (tipi): a scoping review of versions, translations and psychometric properties. *Frontiers in Psychology*, 26(14):1202953.

Williams, K. D. (2007). Ostracism. *Annu. Rev. Psychol.*, 58(1):425–452.

Zadro, L., Williams, K. D., and Richardson, R. (2004). How low can you go? ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. *Journal of Experimental Social Psychology*, 40(4):560–567.