frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 17 February 2022
doi: 10.3389/frobt.2022.733876

Agree to Disagree: Subjective Fairness in Privacy-Restricted Decentralised Conflict Resolution

Alex Raymond [1]*, *Matthew Malencia* [1,2], *Guilherme Paulino-Passos* [3] *and Amanda Prorok* [1]

[1] Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom, [2] GRASP Laboratory, University of Pennsylvania, Philadelphia, PA, United States, [3] Department of Computing, Imperial College London, London, United Kingdom

Fairness is commonly seen as a property of the global outcome of a system and assumes centralisation and complete knowledge. However, in real decentralised applications, agents only have partial observation capabilities. Under limited information, agents rely on communication to divulge some of their private (and unobservable) information to others. When an agent deliberates to resolve conflicts, limited knowledge may cause its perspective of a correct outcome to differ from the actual outcome of the conflict resolution. This is subjective unfairness. As human systems and societies are organised by rules and norms, hybrid human-agent and multi-agent environments of the future will require agents to resolve conflicts in a decentralised and rule-aware way. Prior work achieves such decentralised, rule-aware conflict resolution through cultures: explainable architectures that embed human regulations and norms via argumentation frameworks with verification mechanisms. However, this prior work requires agents to have full state knowledge of each other, whereas many distributed applications in practice admit partial observation capabilities, which may require agents to communicate and carefully opt to release information if privacy constraints apply. To enable decentralised, fairness-aware conflict resolution under privacy constraints, we have two contributions: 1) a novel interaction approach and 2) a formalism of the relationship between privacy and fairness. Our proposed interaction approach is an architecture for privacy-aware explainable conflict resolution where agents engage in a dialogue of hypotheses and facts. To measure the privacy-fairness relationship, we define subjective and objective fairness on both the local and global scope and formalise the impact of partial observability due to privacy in these different notions of fairness. We first study our proposed architecture and the privacy-fairness relationship in the abstract, testing different argumentation strategies on a large number of randomised cultures. We empirically demonstrate the trade-off between privacy, objective fairness, and subjective fairness and show that better strategies can mitigate the effects of privacy in distributed systems. In addition to this analysis across a broad set of randomised abstract cultures, we analyse a case study for a specific scenario: we instantiate our architecture in a multi-agent simulation of prioritised rule-aware collision avoidance with limited information disclosure.

Keywords: fairness, privacy, multi-agent systems and autonomous agents, dialogues, argumentation, explanations

OPEN ACCESS

Edited by:
Ada Diaconescu,
Télécom ParisTech, France

Reviewed by:
Julian Garbiso,
VEDECOM, France
Ivan Serina,
University of Brescia, Italy

****Correspondence:***
Alex Raymond
alex.raymond@cl.cam.ac.uk

Specialty section:
This article was submitted to Multi-Robot Systems, a section of the journal Frontiers in Robotics and AI

Received: *30 June 2021*
Accepted: *14 January 2022*
Published: *17 February 2022*

Citation:
Raymond A, Malencia M, Paulino-Passos G and Prorok A (2022) Agree to Disagree: Subjective Fairness in Privacy-Restricted Decentralised Conflict Resolution. Front. Robot. AI 9:733876. doi: 10.3389/frobt.2022.733876

1 INTRODUCTION AND MOTIVATION

Cognition and autonomy allow intelligent agents in nature to capture information about their surroundings and independently choose a course of action in consonance with their unique and individual decision-making process. Societies thus emerged as collectives of individuals with (mostly) collaborative intent, aided by implicit and explicit systems of norms and rules. The particular complexity of some of these collectives lead some observers to personify or anthropomorphise these groups, attributing a misleading notion of centralised intent and agency to a coherent, but fully decentralised system. However, even the most harmonious and conformable populations in reality exhibit differences in perspective and disagreements across their members.

In view of the above, we understand that humans do not make decisions with truly global information by virtue of the implausible assumption of omniscience. Rather, each individual acts on their own subjective perspectives of local and global status. This subjectivity is not a flaw but instead a fundamental truth of decentralised systems where information is ultimately incomplete or imperfect. In such systems, agents judge outcomes of conflicts based on their partial knowledge of the world, which can lead to perceptions of unfairness when outcomes differ from what other peers perceive as correct. Moreover, individual perspectives can differ drastically when agents choose to retain information under concerns of privacy. As such, it is germane to transpose such considerations to human-agent and multi-agent systems of the future.

The concepts of fairness and justice are prevalent issues in human history along millenia and have been a central topic in areas such as economics (Marx, 1875), philosophy (Rawls, 1991), medicine (Moskop and Iserson, 2007), and more recently, computer science (Li and Tracer, 2017). However, fairness is predominantly regarded as a global property of a system (Narayanan, 2018; Verma and Rubin, 2018; Selbst et al., 2019). In allocation problems such as nationwide organ transplant allocation, fairness concerns the outcome, i.e., whether patients are prioritised to receive donor organs accurately and without discrimination (Bertsimas et al., 2013). Assumptions of global scope and complete knowledge come naturally, as we can only guarantee that a system is fair to all if the information regarding all subjects involved is known. Those assumptions form an objective concept of fairness.

In this work, we elicit a provocation to the habitual definition of fairness: in decentralised applications, where the assumptions of complete global knowledge are withdrawn, can we also discuss fairness with regards to the individual perception of each agent in a system, i.e., representing a subjective notion of fairness? If knowledge is incomplete, can we enable agents to understand that apparently (subjectively) negative decisions are globally and objectively fair, i.e., decisions actually made for the greater good?

1.1 Explanations, Dialogues, and Privacy

As the gap between objective and subjective fairness resides on the need for reasoning and justification, explanations (Čyras et al., 2019; Rosenfeld and Richardson, 2019; Sovrano and Vitali, 2021) artlessly lend themselves as desirable tools to address this issue. This is the focus of much of prior literature concerning explanations to human users. However, also a concern is about explanations between artificial agents themselves, or from a human to an artificial agent. Thus, in a fairness-aware decentralised system with disputed resources, agents with explainable capabilities can expound the reasons for deserving resources in a dialogue that is informed by their respective observations and grounded to a mutually agreed-upon definition of what is and is not fair, or a fairness culture.

Since the concerns for fairness also gravitate around the integration of humans and agents together via reasoning, we lean on architectures for explainable human-agent deconfliction (Raymond et al., 2020). Ideally, participants in a multi-agent system should engage in dialogues (Jakobovits and Vermeir, 1999; Modgil and Luck, 2009) and provide explanations for claiming specific resources. One approach for this problem is computational argumentation, a model for reasoning which represents information as arguments and supports defeasible reasoning capabilities—as well as providing convincing explanations to all agents involved (Dung, 1995; Amgoud and Prade, 2009; Fan and Toni, 2015).

However, not every scenario or application allows for unrestricted exchange of information between agents (Dwork and Roth, 2014). Sensitivity and confidentiality aggravate the poignancy of privacy concerns, rendering the subjective perception of fairness issue, at best, non-trivial when privacy is concerned. We defend subjective fairness as a perennial dimension of the fairness argument whenever a system has non-global knowledge or privacy is concerned.

Whilst the dimensions of privacy, subjective fairness, and objective fairness are desirable in all systems, the simultaneous maximisation of all those aspects is often irreconcilable. If the agents lack full information, there cannot be any guarantee of an objectively fair outcome. Additionally, if there are privacy considerations, subjective fairness cannot be guaranteed as the agent requesting a resource does not receive justification for denial. Therefore, with privacy, neither objective nor subjective fairness is achieved.

This perspective can be observed in society. We illustrate this point as a situation where two passengers are disputing a priority seat in public transport. Even if both individuals are truthful and subscribe to the same culture of fairness (e.g., both agree that a pregnant person should have priority over someone who just feels tired), they must still share information regarding their own personal condition (and therefore, their reasons to justify their belief) to determine who should sit down. How much is each person willing to divulge in order to obtain that resource? Assuming that these individuals have a finite and reasonably realistic threshold on how much privacy they are willing to abdicate for a seat, certain situations will inevitably engender the dilemma of either: 1) forfeiting the dispute and conceding the resource to observe their privacy limit, or 2) exceeding their reservations in privacy and revealing more information than they should in order to remain in contention.

With adamantine restraints on privacy, contenders will always concede and end the debate if presenting a superior reason would exceed their limits of divulged information. Note that this does not mean that the conceding party agrees with the outcome. After all, if a debate was lost exclusively due to the inviolability of privacy, then the loser must still hold an argument that they would have used to trump their opponent, had it not been precluded by privacy limits. Objectively, the fair decision can only be guaranteed if all relevant information is made available by all parties for a transparent and reasonable judgment. In the subjective perception of the bested agent, they still believe they have a superior reason at that point in time, and are left with no choice but to civilly agree to disagree.

1.2 Related Work

Studies of privacy in multi-agent systems have gained recent popularity (Such et al., 2014; Prorok and Kumar, 2017; Torreno et al., 2017). More closely (Gao et al., 2016), also propose the use of argumentation in privacy-constrained environments, although applied to distributed constraint satisfaction problems. Their approach, however, treats privacy in an absolute way, while in our notion is softer, with information having costs, and we consider varying degrees of privacy restrictions.

Contemporaneously, the burgeoning research on fairness in multi-agent systems focuses on objective global fairness, assuming complete knowledge about all agents (Bin-Obaid and Trafalis, 2018). Some works break the global assumption by applying fairness definitions to a neighborhood rather than an entire population (Emelianov et al., 2019) or by assuming that fairness solely depends on an individual (Nguyen and Rothe, 2016). The former studies objective fairness of a neighborhood, assuming full information of a subset of the population subset, whilst the latter assumes agents have no information outside of their own to make judgments about fairness. These works do not address fairness under partial observability, wherein agents have partial information on a subset of the population, which we call subjective local fairness.

To study privacy and subjective fairness in distributed multi-agent environments, we look to previous work in explainable human-agent deconfliction. The architecture proposed in (Raymond et al., 2020) introduces a model for explainable conflict resolution in multi-agent norm-aware environments by converting rules into arguments in a culture (see **Section 2.1**). Disputes between agents are solved by a dialogue game, and the arguments uttered in the history of the exchange compose an explanation for the decision agreed upon by the agents.

Notwithstanding its abstract nature, this architecture relies on two important assumptions: 1) that agents have complete information about themselves and other agents; and 2) that dialogues extend indefinitely until an agent is cornered out of arguments—thus being convinced to concede. In most real-life applications, however, those assumptions occur rather infrequently. Fully decentralised agents often rely on local observations and communication to compose partial representations of the world state, and indefinite dialogues are both practically and computationally restrictive. We build on the state of the art by extending this architecture to support gradual disclosure of information and privacy restrictions.

1.3 Contributions

This paper stages the ensuing contributions:

- We present an architecture for multi-agent privacy-aware conflict resolution using cultures and dialogue games.
- We introduce a formalisation of the subjective fairness problem and the corresponding privacy trade-off.
- We simulate interactions in random environments and compare how different argumentation and explanation strategies perform with regards to our fairness metrics.
- We instantiate a multi-agent prioritised collision avoidance scenario and demonstrate practical instances of subjective fairness and privacy limitations.

The structure of this paper follows as: **Section 2** introduces prior background from literature used in our study. In **Section 3**, we introduce an argumentation-based architecture for decentralised conflict resolution under privacy-restricted communication. We provide an abstract formalism of the dimensions of fairness in conflict resolution in **Section 4**. Then, we use the contributions of those two sections in **Sections 5, 6** to perform empirical experiments on abstract conflicts and on an applied multi-agent collision avoidance scenario, respectively. Our results show that using better argumentative strategies suggests Pareto improvements for the fairness-privacy trade-off.

2 BACKGROUND

We introduce the required theoretical background from present literature used in the development of our study and methods. **Section 2.1** abridges the concept of argumentation frameworks, dialogues, followed by cultures in **Section 2.2**.

2.1 Argumentation and Dialogues

The concept of culture (Raymond et al., 2020) reflects a collective agreement of norms and priorities in a multi-agent system, checked dynamically by means of verifier functions. Abstract Argumentation Frameworks (Dung, 1995) are used to underpin the mechanics of cultures and conflict resolution dialogues.

Definition 2.1. (Argumentation Framework). An argumentation framework is a digraph $AF = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of arguments (vertices) and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is a set of attack relations between arguments (arcs). We say attacks(a, b) holds iff $(a, b) \in \mathcal{R}$. Stating attacks(a, b) is intuitively equivalent to defining that an argument b is defeated by argument a. Likewise, a set $S \subseteq \mathcal{A}$ of arguments attacks another set of arguments T (or T is attacked by S) if any argument in S attacks an argument in T. An argument $a \in \mathcal{A}$ is acceptable with respect to a set S of arguments iff for each argument $b \in \mathcal{A}$ that attacks a there is a $c \in S$ that attacks b. In that case, c is said to defend a. A set S of arguments is conflict-free iff for all $a, b \in S$, $(a, b) \notin \mathcal{R}$, and admissible iff it is conflict-free and all its arguments are acceptable with respect to S.

The notion of extensions (Doutre and Mengin, 2001) presents itself as semantics of acceptance for sets of arguments.

Definition 2.2. (Preferred Extension). Let $AF = (\mathcal{A}, \mathcal{R})$ be an argumentation framework and $S \subseteq \mathcal{A}$ be a set of arguments. We say S is a preferred extension of AF iff: S is conflict-free, defends all arguments in S, and is a maximal element among all admissible sets, with respect to set-theoretical inclusion. If an argument $a \in \mathcal{A}$ is present in all preferred extensions of AF, we say a is sceptically accepted in the preferred extensions.

These definitions create the foundation for conflict resolution through dialectical interaction. Let A be the set of all existing agents in a multi-agent environment. When pairwise conflicts occur between agents $q_i, q_j \in A$, the two agents serve as players in a dialectical game, or dialogue, to resolve the conflict, and they take up the roles of proponent (pr) or opponent (op), where $q_i \in \{pr, op\}$, and $q_j \in \{pr, op\} \backslash q_i$, without loss of generality. These dialogues are sequential in nature, where players utter arguments and attempt to defeat the previously-used argument by their adversary. When a player can no longer choose a valid argument, they lose the game and concede the contended resource to the winner. Jakobovits and Vermeir (1999) delineate a formalism for dialectical games, and we adapt some of their definitions below for single-argument dialogues.

Definition 2.3. (Dialogue). Let a player $w \in \{pr, op\}$ and an argument $a \in \mathcal{A}$. The adversary of pr is denoted $\overline{pr} = op$, and $\overline{op} = pr$. A move is a pair (w, a). For a move $m = (w, a)$, we use player(m) to denote w and arg(m) to denote a. A move is considered legal iff it does not attack itself and is not already attacked by a previously-uttered argument. We say a dialogue D is any countable sequence $m_0, m_1, \ldots, m_n$ of moves that satisfies:

(1) $player(m_{i+1}) = \overline{player}(m_i)$, i.e., the players take turns.
(2) m_{i+1} (i.e., the next move) is legal.
(3) $m_{i+1} \notin \{m_0, m_1, \ldots, m_i\}$, i.e., a move cannot be repeated.
(4) $attacks(arg(m_{i+1}), arg(m_i))$, it attacks the adversary's last move
(5) $player(m_0) = pr$, i.e., the proponent makes the first move.

The dialogue D is said to be about the position $arg(m_0)$. If there is no dialogue $D' = m_0, m_1, \ldots, m_n, \ldots, m_{n'}$ that extends the D, then we say that $player(m_n)$ is the winner of D.

2.2 Cultures

In such frameworks, arguments and attacks are static by nature and assumed true. Informally, the surviving sets of arguments under certain semantics (extensions) represent coherent conclusions given a fixed argumentation framework. This inflexibility is a strong limitation for dynamic interactive systems, where agents can have varied states and contexts, causing arguments to lose validity depending on the circumstances. Cultures for conflict resolution (Raymond et al., 2020) embody rules and norms as high-level, dynamic arguments that are shared and agreed upon between all agents in a system, and act as a template for generating argumentation frameworks dynamically to model the complexities that arise in multi-agent systems. During inference, arguments are dynamically verified for correctness given the required information from both agents, such as their current states, and the context of a specific dispute, e.g., the state of the environment.

Definition 2.4. (Culture). Let any two players pr, op be the proponent and opponent in a dialogue game. We say a motion is any argument $a \in \mathcal{A}$ that may be used by proponent pr to request a contended resource from opponent op. Let $\mathcal{K} \subseteq \mathcal{A}$ be the set of all motions in $\mathcal{A}$. Let $\mathcal{R}$ be the set of attacks between arguments in $\mathcal{A}$. We say a multi-agent system has a culture $C = (\mathcal{A}, \mathcal{R}, \mathcal{K})$ iff $|\mathcal{K}| > 0$ and if all agents in A share a common culture C.

Definition 2.5. (Argument Verification). Let $a \in \mathcal{A}$ and $w \in \{pr, op\}$ be an argument and a player, respectively. We denote a as demonstrable by agent-player w iff checking the correctness of that argument admits a finite and computable decision procedure. Let ζ denote the set of all possible contexts in the environment. $\forall a \in \mathcal{A}$, a admits a predicate function $v_a\colon w, z \rightarrow \{\texttt{True}, \texttt{False}\}$, where w represents the player and $z \in \zeta$ is a context. We say v_a is the verifier function of argument a. Motions are hypothetical and their verifier functions always return True. An argument a is demonstrably true by player w iff $v_a(w, z) = \texttt{True}$.

3 PRIVACY-AWARE CULTURES

The cultures introduced in Raymond et al. (2020) enable explainable conflict resolution in multi-agent norm-aware environments. However, these cultures require full state knowledge among agents—and conflicts are resolved through indefinitely-long dialogues. In real-world distributed applications, agents only have partial observations of other agents, real world constraints prevent limitless dialogue, and privacy concerns govern the types and amount of information that agents divulge.

Example 3.1. Suppose 2 agents, Belle and Cadence, are disputing a priority seat on public transport. They share a culture $C = (\mathcal{A}, \mathcal{R}, \mathcal{K})$ where $\mathcal{A} = \{\gamma, a, b\}$ contains the arguments

- (motion) $\gamma \equiv$ "I think I should have this seat."
- $a \equiv$ "I am older than you."
- $b \equiv$ "I have a more serious health condition."

The relations $\mathcal{R} = \{(a, \gamma), (b, \gamma), (b, a)\}$ determine the priority of those concepts in the agents' shared culture. If Belle and Cadence do not know each other's age and health condition, they will have no information to verify arguments a and b, i.e., they cannot argue that they are demonstrably older or more infirm than their adversary. At this stage, any of those arguments could only be raised as a hypothesis. Additionally, both Belle and Cadence would need to break privacy and mutually reveal some personal information to reach a decision.

To allow agents to engage in dialogues under partial information and privacy constraints, we present a novel explainable conflict resolution architecture. We begin by looking at the aspect of alteroceptive cultures, our proposed mechanism organizing the space of interactions for privacy-aware conflict resolution. "Alteroceptive" is our coinage from a portmanteau of the Latin word alter (other) + the word reception. It shall connote a "sense of other."

3.1 Alteroceptive Cultures

As shown in Example 3.1, when no information is available and arguments are comparative, agents can only raise hypotheses. Those serve as media for sharing information between agents, by enforcing that every agent shares their corresponding piece of information when eliciting a hypothesis, in order to avoid infinite exchanges arising from cycles of empty suppositions.

Once an agent raises a hypothesis and shares their pertaining information regarding an argument, the adversary should have enough information to effectively verify the argument into a fact, since it has full knowledge of its own description and was given the other agent's partial description. This can be done without necessarily sharing further information.

Additionally, the relations between arguments need to be woven in such a way as to guarantee that hypotheses and facts only defeat the arguments pertaining to the adversary. In Example 3.1, if the same agent wants (and is able) to utter both arguments a and b, the attack (b, a) in the culture should not yield an actual attack in an instantiation if both arguments are articulated by the same agent. After all, having a trumping condition would be yet another reason in favor of—not against—the agent's claim. Hence, assuring that attacks only occur between adversaries renders a bipartition in the culture and its arguments. The conjunction of those elements culminates in an alteroceptive culture.

Definition 3.1. (Alteroceptive Culture). Let $C = (\mathcal{A}, \mathcal{R}, \mathcal{K})$ be a culture . We define the expansion function κ, that maps an argument $a \in \mathcal{A}$ to a set of new arguments (not in $\mathcal{A}$). For every $a \in \mathcal{A} \backslash \mathcal{K}$, $\kappa(a) = \{a_H^{pr}, a_H^{op}, a_F^{pr}, a_F^{op}\}$ four new arguments that represent, respectively:

- a_H^{pr}: hypothesis (H) of a where proponent pr wins,
- a_H^{op}: hypothesis (H) of a where opponent op wins,
- a_F^{pr}: verified-fact (F) of a where proponent pr wins,
- a_F^{op}: verified-fact (F) of a where opponent op wins.

For motions in $\gamma \in \mathcal{K}$, $\kappa(\gamma) = \{\gamma_H^{pr}, \gamma_H^{op}\}$, since motions do not admit verification. We define the expanded set of arguments (in an alteroceptive culture) $\mathcal{A}_x = \bigcup_{a \in \mathcal{A}} \kappa(a)$. The expanded set of attacks $\mathcal{R}_x$ is constructed from $\mathcal{R}$ and $\mathcal{A}$, and satisfies the following rules (see **Figure 1**). For each element of $\kappa(a)$, where $a, b \in \mathcal{A}$ and $(a, b) \in \mathcal{R}$. For all $w \in \{pr, op\}$:

(1) $(a_H^w, a_H^{\bar{w}}) \in \mathcal{R}_x$ if $a \notin \mathcal{K}$, i.e., non-motion hypotheses mutually attack each other;
(2) $(a_F^w, a_F^{\bar{w}}) \in \mathcal{R}_x$, i.e., verified-facts mutually attack each other;
(3) $(a_F^w, a_H^{\bar{w}}) \in \mathcal{R}_x$ i.e., each verified-fact attacks their adversary's hypothesis;
(4) $(a_H^w, b_H^{\bar{w}}) \in \mathcal{R}_x$ and $(a_H^w, b_F^{\bar{w}}) \in \mathcal{R}_x$, i.e., hypotheses reproduce their original attacks to both hypotheses and verified-facts;
(5) there are no more elements in $\mathcal{R}_x$.

We say $C_x(C) = (\mathcal{A}_x, \mathcal{R}_x)$ is an alteroceptive expansion of C.

The separation between hypotheses and facts is important to introduce a mechanism for gradual disclosure of information.

Example 3.2. (Example 3.1). Given the original culture seen in Example 3.1, we can generate an extended set of arguments $\mathcal{A}_x = \bigcup \{\kappa(\gamma), \kappa(a), \kappa(b)\}$ and the corresponding $\mathcal{R}_x$ as described in Definition 3.1. Suppose, then, that Cadence raises the motion $\gamma_H^{Cad} \equiv$ "I think I should have this seat," to which Belle promptly replies with $a_H^{Belle} \equiv$ "I may be older than you, I am 60 years old." **Figure 1** illustrates the expanded argument set in question. Note that Belle had to break privacy and reveal to Cadence "I am 60 years old" in order to use this argument. Cadence can now verify argument a_F^{Cad} (the fact argument that Cadence is older than Belle). If this argument is verified as `True` (i.e., Cadence is older than Belle), then Cadence has two options for rebuttal, depending on what information they intend to share:

(1) Use a_F^{Cad}, reveal their age and refute Belle's claim of being older;
(2) or use b_H^{Cad}, ignore the age dispute and move on to the health argument, revealing their health condition and hypothesising that they might be more ill than Belle.

The example above shows that, for her next move, Cadence can choose to either reveal her age or health condition to progress the debate. This decision is affected by which aspect of its description an agent would like to keep private. Laying out arguments within this structure still allows for the same dialogue game rules as before, but without the original assumption of complete information. A dialogue under this framework would carry out normally and extend indefinitely as agents exchange moves until one of the agents loses by running out of arguments. Information can, therefore, still be communicated unreservedly. In the interest of quantifying relinquished privacy, we associate arguments to privacy costs.

3.2 Privacy-Aware Dialogues

Every argument in an alteroceptive culture corresponds to a concession of privacy through communication of one's partial description. Fortunately, those changes preserve the structure of an argumentation framework, which makes alteroceptive cultures compatible with prior mechanics of dialogue, as well as extensions. Analogously to how humans are comfortable with sharing some types of personal information but not others, some features in agents' descriptions might be considered more sensitive, and thus having a higher privacy cost to reveal.[1] We combine the rules of dialogue (Definition 2.3) and the notion of privacy cost to provide an instantiation for our architecture.

[1] Approaches using cultures assume that all agents agree on the argumentation framework, otherwise the process of argumentation breaks down. Aligned with this notion of a shared AF, the alteroceptive culture in this work has privacy costs defined per argument, which is fixed for all agents.

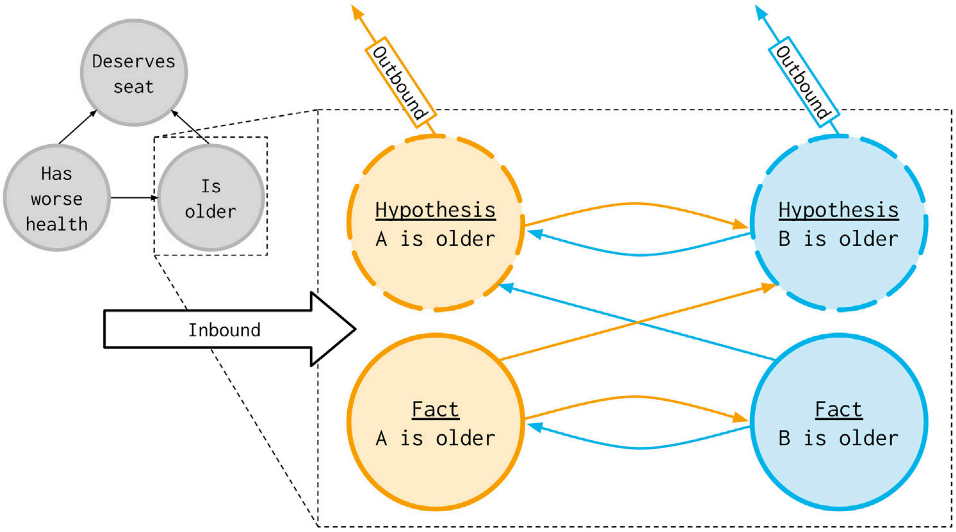

FIGURE 1 | Alteroceptive cultures transform the arguments of a given culture (grey) into hypotheses (dashed nodes) and verified-facts (solid nodes). Shown here from Example 3.1, argument *a*, "Is older," is transformed into the respective nodes for the orange opponent, a_H^{op} and a_F^{or}, and the blue proponent, a_H^{pr} and a_F^{pr}. The inbound attack (left) from "Has worse health" attacks all four nodes and the outbound attacks (top) attack "Deserves seat" from the hypotheses.

Definition 3.2. (Privacy Cost and Budget). Let $\mathcal{A}_x$ be the extended set of arguments of an alteroceptive culture. We say τ: $\mathcal{A}_x \to \mathbb{Z}^+$ is a privacy cost function. Let A be the set of all agents. We define β: $A \to \mathbb{Z}^+$ as the privacy budget function of an agent.

Definition 3.3. (Privacy-Aware Dialogue). Let $w \in \{pr, op\}$ be any two players and $D = m_0, m_1, \ldots, m_n$ be a dialogue. Let $moves(w, n) = \bigcup_{m_i \in \mathcal{D}} \in D\{m_i \mid player(m_i) = w \text{ and } i \leq n\}$ denote all the moves of a player w up to round n. We say a dialogue D is privacy-aware iff, in addition to the criteria in Definition 2.3, it also satisfies, for all n:

(6) $\beta(player(m_{i+1})) \geq \sum \tau(\arg(m))$, *where* $m \in moves(player(m_{i+1}), i)$,

i.e., the player cannot cumulatively spend more than its privacy budget.

Our definition imposes a hard limit privacy: during a dialogue, agents cannot use an argument that would aggregately exceed their pre-defined privacy budgets. With finite privacy budgets, the assumption of dialogues extending indefinitely until reaching a unanimous conclusion no longer holds. Disputes end with an agent losing in one of two different ways. Either: 1) the agent runs out of arguments and is convinced out of the dialogue, or 2) the agent still has valid hypotheses and/or verified-facts that could attack the last move, but cannot afford to use them and is forced to concede. In the former case, both agents agree on the correct outcome as the losing agent did not have any valid reasons to challenge the winner. In the latter case, however, the agent concedes and interrupts the dialogue before having all arguments successfully refuted—as it prefers to preserve their privacy. The agents tolerate the result but do not agree with it: they agree to disagree.

Notably, the choice of arguments can have a decisive impact on the dialogue game. If we consider scenario (2) above to be less desirable than (1) for both agents, both agents must maximize the effectiveness of their moves so that dialogues end on the basis of argumentation instead of privacy budgets. Yet, optimal strategies are not readily available, as partial observability qualifies our privacy-aware dialogue game as a game of incomplete information.[2]

With that in place, we next introduce our foundational terminology and abstract definitions governing the problem of subjective fairness in privacy-restricted decentralised conflict resolution.

4 PROBLEM DEFINITION

While the new argumentation architecture presented above enables dialogues under the partial information of privacy constraints, such partial observability may cause discrepancies between the outcome of a conflict resolution and the agents' individual perception of what the correct outcome should be, which we define as subjective unfairness. This new slant on the perspective of fairness introduces another dimension for conflict

[2] Typically such games are analysed by a completion approach using Bayesian reasoning, and thus turned into imperfect information games (Binmore, 1992). The alternative allows agents to adopt mixed strategies, using information about an underlying probability distribution. We consider only pure strategies due to fewer needed assumptions, where agents do not use any probabilistic information. However, for pure strategies the optimal solution can be intractable (Reif, 1984; Blair et al., 1993), and we thus limit ourselves to simple ones in the next sections.

resolution whenever a system has privacy reservations or non-global knowledge. This section thus defines a foundational mathematical formalism to study fairness and its relationship to privacy in decentralised conflict resolution.[3]

Consider a situation in which a population of agents in a multi-agent system interacts and disputes advantages or resources. Agents interact in a pairwise manner *via* disputes, and a final system state is achieved through a sequence of such interactions. We observe whether this final state is fair given complete knowledge: this we call objective fairness. A second aspect is whether the outcome is fair from the perspective of each agent. This is subjective fairness. Scope further augments our setting: the term global relates to population-wide observations, whilst local concerns pairwise interactions.

We remove the assumptions of unrestricted information-sharing, instead assuming that privacy restricts an agent's willingness to divulge information. Particularly, we analyze how privacy between agents is an impediment towards a guarantee of both objectively and subjectively fair outcomes. We formalize such concepts below.

4.1 Information, Privacy and Fairness

We postulate that agents possess idiosyncratic features with varied values across a population, and that the set of all features that describe an agent is, by reason, called a description. In Example 3.1 above, Belle and Cadence have the features "age" and "health status." Each feature takes a value, and the set of all feature values forms the description of each agent. Assume a distributed multi-agent setting where agents have perfect knowledge about their own features but none-to-partial knowledge about other agents' descriptions. We first define the (full) description of an agent as an n-tuple of features.

Definition 4.1. (Description). For the rest of this section, we consider a fixed set A of existing agents in an environment and an underlying set F of all possible values of a feature. The set of features is $I = \{1, \ldots, n\}$. We define $\mathcal{F} = F^n$ as the set of feature descriptions, that is, the set of n-tuples, each entry representing the value of a feature. The description function $d\colon A \to \mathcal{F}$ is a function that maps each agent to its full description in terms of features.

Since we are considering privacy-sensitive applications, we attribute a privacy cost to each feature, which is then aggregated to produce a cost for the full description. For brevity, all further mentions of cost shall refer exclusively to privacy cost. The constant `unknown` represents an undisclosed value of a feature.

Definition 4.2. (Privacy Cost). The set of private features is $F_p = F \cup \{\texttt{unknown}\}$. A partial description is an element of the set $\mathcal{F}_p \in F_p{}^n$, that is, an n-tuple where each entry, $f_p{}^i$, corresponds to either a feature value or to `unknown`. Each feature $i \in I$ has an associated cost, k_i. A cost function is defined as $\tau\colon \mathcal{F}_p \times I \to \mathbb{Z}^+$, such that $\tau(f_p{}^i, i) = 0$ if $f_p{}^i = \{\texttt{unknown}\}$, and $\tau(f_p{}^i, i) = k_i \in \mathbb{Z}^+$, otherwise. Let $x = (x_1, \ldots, x_n) \in \mathcal{F}_p$. The description cost function $\mathcal{T}\colon \mathcal{F}_p \to \mathbb{Z}^+$ is given by $\mathcal{T}(x) = \sum_{i\in I} \tau(x_i, i)$. For ease of notation, the feature index i is implied and we use the notation $\tau(x_i)$ interchangeably without loss of generality.

We can now consider how agents interact: we denote every one-to-one interaction a conflict resolution dispute, or dispute, for brevity. Agents will act with antagonistic intent: one of them pushes for a change in the status quo, and is thus called proponent, while the other is called opponent. In such pairwise disputes between agents, we need a mechanism for defining what determines an objectively fair outcome of a contest, given two descriptions of agents. Following the idea behind cultures, we assume there is a shared definition of fairness across agents.[4] An objectively fair outcome disregards privacy, since it is the outcome which should be achieved under complete information.

As opposed to this perfect outcome, we also consider what an agent involved in a dispute considers a fair outcome. This we define as the subjectively fair outcome, according to one of the agents in the dispute, which is dependent on the information available to it. Any agent always has full information of oneself, but may only have partial information about the other party.

Definition 4.3. (Fair Outcomes). We define the set of roles in any dispute to be $R = \{pr, op\}$, two constant symbols denoting respectively "proponent" and "opponent," for the rest of the section. An objectively fair outcome function $\omega\colon \mathcal{F} \times \mathcal{F} \to R$ receives the descriptions of the proponent and opponent agents, respectively, and decides which should win a dispute by returning the role of the winning agent in the dispute. The subjectively fair outcome function $\omega_p\colon R \times \mathcal{F} \times \mathcal{F}_p \to R$ receives as input whether the agent under consideration (the subject) is the proponent or opponent, the description of the agent, and the private description of the adversary, and outputs the role of the agent which deserves to win the dispute.

The choice of information disclosure to the adversary is determined by a disclosed information function α, which takes an agent (who is deciding which information to disclose), as well as the adversary, and produces a partial description $\mathcal{F}_p$ of the agent that fits under a specific privacy budget $g \in \mathbb{Z}^+$.

Definition 4.4. (Disclosed Information). We say $\alpha\colon A \times A \times \mathbb{Z}^+ \to \mathcal{F}_p$ is a disclosed information function if and only if, for any agents a, $b \in A$ and privacy budget $g \in \mathbb{Z}^+$, α satisfies $\mathcal{T}(\alpha(a, b, g)) \leq g$.

Once both agents made their decisions on how to generate their partial descriptions, the dispute resolution function returns the winner of a dispute with partial information on both sides.

[3] We dissociate this formalism from **Section 3**'s emphasis on argumentation frameworks to the higher-level space of conflict resolution, as AFs are one of many conflict resolution methods. We hope others in the conflict resolution domain will build on this work and study subjective fairness using their tools of choice.

[4] This makes no commitment on what is the ontological status of this shared definition, or whether it is deemed "correct." We simply assume there is such a definition and that it is shared among the agents.

Definition 4.5. (Dispute Resolution). A dispute resolution function ϕ: $\mathcal{F}_p \times \mathcal{F}_p \to R$ decides whether the proponent or the opponent wins the contest given their partial feature descriptions, respectively, by returning the role of winning agent in the dispute.

The difference between functions ω, ω_p, and ϕ goes as follows: the first, ω, returns the objectively fair outcome if all information is public and mutually known between both agents (complete). The function ω_p represents the perspective of the subject upon receiving disclosed information from the adversary, after all, an agent always knows its full description but has partial information about other agents (depending on what is disclosed). Lastly, ϕ represents the final outcome given both agents' partial descriptions, after the dispute is resolved in some way.

While those definitions cover the essential concepts for pairwise (i.e., local) interactions, we still lack a relation of such pairwise interactions to a global state of the system. We assume the existence of a set S of possible (global) states in which the system can be, with regards to the population of agents and a global notion of fairness. From an initial state, a set of local interactions transition the system into a final state. This is said to be the transition of the system.

Definition 4.6. (System Transition Function). Let A be the set of agents, α be a disclosed information function, ϕ the dispute resolution function, g a privacy budget, and $S_0 \in S$ a state (referred as the initial state). A system transition function σ is a function such that $\sigma(A, \alpha, \phi, g, S_0) \in S$, and we call $\sigma(A, \alpha, \phi, g, S_0) = S_F$ the final state.

We have introduced two aspects of fairness: perspective and scope. Perspective differentiates objective from subjective fairness, while scope differentiates global from local fairness. We formalize divergences in conflict resolution outcomes through the notion of a fairness loss.

4.2 Fairness Loss

The outcomes of privacy-restricted fairness disputes may disagree with our previously-defined notions of fair outcomes. We introduce fairness loss functions as a means for comparing these.

The first definition for loss evaluates whether the dispute resolution outcome matches the objectively fair one, for a given privacy budget.

Definition 4.7. (Objective Local Fairness Loss). The objective local fairness loss function l_{OL}: $A \times A \times \mathbb{Z}^+ \to \{0, 1\}$ is defined as

$$l_{OL}(a, b, g) = \begin{cases} 0, & \text{if } \phi(\alpha(a, b, g), \alpha(b, a, g)) = \omega(d(a), d(b)) \\ 1, & \text{otherwise} \end{cases}$$

where a, $b \in A$ are agents, with a the proponent and b the opponent, and $g \in \mathbb{Z}^+$ denotes a privacy budget.

Our second definition formalizes whether the dispute resolution outcome is the same as the subjectively fair outcome for the agent playing the role r in the dispute.

Definition 4.8. (Subjective Local Fairness Loss). The subjective local fairness loss function l_{SL}: $A \times A \times R \times \mathbb{Z}^+ \to \{0, 1\}$ is defined as

$$l_{SL}(a, b, r, g) = \begin{cases} 0, & \text{if } \phi(\alpha(a, b, g), \alpha(b, a, g)) = \omega_p(r, d(a), \alpha(b, a, g)) \\ 1, & \text{otherwise} \end{cases}$$

where a, $b \in A$, $r \in R$, and $g \in \mathbb{Z}^+$ denotes a privacy budget.

Finally, for objective and subjective global fairness, we characterize its requirements, but leave it to be specified application-wise. We present an applied experiment of objective and subjective global fairness in **Section 6**.

Definition 4.9. (Global Fairness Loss). The objective global fairness loss function Ω: $S \to \mathbb{R}^+$, maps a state of a system to an unfairness value according to an objective notion of fairness. Analogously, a subjective global fairness loss function Ω_p: $S \to \mathbb{R}^+$ maps the same state to an unfairness value that stems from the subjective perception of the population. A higher value in either means that the state is less desirable, that is, less fair.

We now introduce definitions pertaining to "orderly" cases of the previous definitions. Intuitively, a dispute resolution function is one such that when all information is available, an objective local fair outcome is achieved. When that is the case, we call it publicly sound, formally as follows:

Definition 4.10. A dispute resolution function ϕ is publicly sound iff for all $f_a, f_b \in \mathcal{F}$, $\phi(f_a, f_b) = \omega(f_a, f_b)$.

We also define an agent being reasonable when, given perfect information about the other agent, a always considers the objectively fair outcome a subjectively fair outcome, in any contest with another agent b. Being reasonable thus implies that the only deterrent towards an objectively fair outcome is partial information, as a reasonable agent will always agree with the objectively fair outcome when given enough information.

Definition 4.11. An agent $a \in A$ is reasonable iff for any other agent $b \in A$, and any role $r \in R$, $\omega_p(r, d(a), d(b)) = \omega(d(a), d(b))$.

Global outcomes can now connect with local outcomes by stipulating that, whenever complete information is available and the dispute resolution function is publicly sound, then the objective global fairness loss is 0. This corresponds to system transitions in which, whenever all disputes are resolved with objectively fair outcomes, then the final state is always objectively globally fair.

Definition 4.12. Let α be a disclosed information function, ϕ be a publicly sound dispute resolution function, and $g \in \mathbb{Z}^+$ be such that, for every pair of agents a, $b \in A$, $\phi(\alpha(a, b, g), \alpha(b, a, g)) = \omega(d(a), d(b))$. A system transition function σ is publicly unbiased iff for any state $S_0 \in S$, we have that $\Omega(\sigma(A, \alpha, \phi, g, S_0)) = \Omega_p(\sigma(A, \alpha, \phi, g, S_0)) = 0$.

Finally, we can state a result, thus proving that:

Theorem 4.1. If ϕ is a publicly sound dispute resolution function, every agent in A is reasonable, $g \in \mathbb{Z}^+$ is such that for all $a \in A$,

$\alpha(a, g) = d(a)$, and σ is publicly unbiased, then, for all $a, b \in A$ and state $S_0 \in S$, $l_{OL}(a, b, g) = l_{SL}(a, b, g) = \Omega(\sigma(A, \alpha, \phi, g, S_0)) = \Omega_p(\sigma(A, \alpha, \phi, g, S_0)) = 0$.

Proof: Considering Definition 4.7, we know that for a privacy budget g high enough such that $\alpha(a, g) = d(a)$ and $\alpha(b, g) = d(b)$, we achieve a global loss of 0. We can then replace the private descriptions $\alpha(a, g)$ and $\alpha(b, g)$ with their public equivalents $d(a)$ and $d(b)$ respectively, obtaining $\phi(d(a), d(b)) = \omega(d(a), d(b))$. Since ϕ is publicly sound, this holds.

A similar reasoning applies to l_{SL} in Definition 4.8. Suppose a g high enough such that $\alpha(a, g) = d(a)$ and $\alpha(b, g) = d(b)$, for every $a, b \in A$. Applying this to l_{SL}, for any $o \in \{a, b\}$, the value is 0 if $\phi(d(a), d(b)) = \omega_p(o, d(a), d(b))$. This holds, since a and b are reasonable agents.

Finally, for Ω, it is a direct consequence of Definition 4.12. Due to the constraint on g, $\alpha(a, b, g) = d(a)$, and similarly for b. Thus $\phi(\alpha(a, b, g), \alpha(b, a, g)) = \omega(d(a), d(b))$ is satisfied since ϕ is publicly sound.

With the extant formalisms providing a frame of reference and context, we introduce a problem statement that predicates the motivation and guides our resulting contributions in this paper.

4.2.1 Problem Statement

Let A be a set of agents. We assume all agents in A are self-interested and will experience mutual conflicts of interest regarding contended resources. We assume an initial state S_0, which transitions depending on outcomes of agent conflicts. For every conflict between any agents $a, b \in A$ that arises over a contended resource, agents will engage in a dispute with mutually-exchanged partial information. This will be governed by means of: an equal privacy budget g for both agents, a disclosed information function α representing the strategy of both agents for releasing information, and a dispute resolution function ϕ. For any given g, find an α that minimizes Ω and Ω_p.

5 EMPIRICAL ANALYSIS

In this section, we propose an experiment to measure the effectiveness of distinct dialogue strategies with respect to global and local fairness under varying degrees of privacy. As shown by Theorem 4.1, both objective and subjective fairness loss definitionally converge to zero if the conflict resolution mechanism is publicly sound and agents act reasonably. However, this is a loose bound—as those definitions do not account for the specific mechanics of dialogues, conflict resolution, or fairness.

Consequently, we use the architecture proposed in **Section 3** to instantiate scenarios with randomly-generated agents and cultures. Our interest lies in empirically observing the impact on global and local fairness under different privacy budgets using four different argumentation strategies. Below, we introduce the details for our experimental setup and evaluation mechanisms.

5.1 Setup

Let $A = \{q_1, q_2, \ldots, q_{|A|}\}$ be a finite set of $|A|$ agents. Every agent $q \in A$ possesses an m-tuple $d(q) = (i_1, i_2, \ldots, i_m)$ for all $i \in I$, where $|I| = m$, representing that agent's feature description, i.e., its internal traits and characteristics. The value function μ: $A \times I \rightarrow \mathbb{Z}^+$ returns the numerical value of a feature $i \in I$ for an agent $q \in A$.

Let $C = \{\mathcal{A}, \mathcal{R}, \mathcal{K}\}$ be a culture, where $\mathcal{A} = \{\gamma, a_1, a_2, \ldots, a_m\}$ is composed of $m + 1$ arguments (with a single motion $\gamma \in \mathcal{K}$). Let $index(a_j) = j, a_j \in \mathcal{A}$ denote the index of an argument. We generate $\mathcal{R}$ randomly, satisfying the conditions: 1) the underlying $AF = (\mathcal{A}, \mathcal{R})$ has exactly one connected component; 2) for every $(a, b) \in \mathcal{R}$, $index(a) > index(b)$.

Non-motion arguments in C will represent a feature comparison between two agents. We consider the alteroceptive expansion of C, $C_x = (\mathcal{A}_x, \mathcal{R}_x)$ (Definition 3.1). Every verified-fact argument $a_F^{q_j} \in \mathcal{A}_x$ is associated to a verifier function $v_a(q_j, q_k) =$ `True` if $\mu(q_j, index(a)) > \mu(q_k, index(a))$; otherwise `False`, for $q_j, q_k \in A$. All hypotheses are trivially associated to verifier functions that always return `True`.

Informally, this means that every feature $i \in I$ is represented in a dialogue between two agents by the respective hypotheses and verified-facts regarding which agent has a superior value in feature i. This abstractly represents any potential feature in a system.

5.2 Evaluation Metrics

To empirically study the impact of privacy on local and global fairness using different argumentation strategies, we define two evaluation metrics. First, we consider the aggregated subjective local unfairness as the summation of the subjective local fairness loss l_{SL} over all pairs of distinct agents. In our case, this is modeled by dialogues that end due to a lack of privacy budgets from one of the agents, as noted by dispute result (2) in **Section 3**.

To measure objective local fairness loss, we calculate a ground truth of all pairwise interactions of agents. This ground truth is defined as an $|A|\times|A|$ matrix $\boldsymbol{GT}$, with entries being elements of $R = \{pr, op\}$. The entries of $\boldsymbol{GT}$ are the objectively fair outcome of the dispute in which agents a_j and a_k assume the roles of pr and op, respectively, that is: $\boldsymbol{GT}_{j,k} = \omega(d(q_j), d(q_k))$.

Our instantiation of the objectively fair outcome function ω is defined the following procedure:

(1) Given the complete descriptions $d(pr)$ and $d(op)$, run all verifier functions for all arguments and remove all verified-fact arguments that return `False`.

(2) Check for sceptical acceptance[5] of the proponent's motion γ_H^{pr}. If yes, then return pr. Return op otherwise.

The resulting ground truth can also be represented as a digraph $G_{\boldsymbol{GT}} = (A, E)$, where for every two distinct agents $q_j, q_k \in A$, we say an arc $(q_j, q_k) \in E$ *iff* $\boldsymbol{GT}_{j,k} = pr$ and $\boldsymbol{GT}_{k,j} = op$. We generalise this definition for any strategy α and denote the digraph G_α as a precedence graph of agents.

Objective global fairness loss compares the ground truth precedence graph, $G_{\boldsymbol{GT}}$, to the precedence graph resulting from a strategy, $G_{\boldsymbol{RE}(g,\alpha)}$. To compare these two precedence

[5] We used the `μ-toksia` (Niskanen and Järvisalo, 2020) SAT-based solver for the simulations.

graphs, we use the DAG dissimilarity metric seen in (Malmi et al., 2015), defined below.

Let $G_1 = (A, E_1)$ and $G_2 = (A, E_2)$ be two precedence graphs. Let e denote an arc (q_j, q_k). Let c_1 denote the number of occurrences where $e \in E_1$ and its reverse $(q_k, q_j) \in E_2$, c_2 as the number of occurrences where e exists in either E_1 or E_2 but not the other, and c_3 as the number of occurrences where neither E_1 or E_2 contain e. The DAG dissimilarity between two graphs $G_1 = (A, E_1)$ and $G_2 = (A, E_2)$ is $K(G_1, G_2, y_1, y_2) = c_1 + c_2y_1 + c_3y_2$, where $0 \leq y_2 < y_1 \leq 1$ are constants. We choose $y_1 = 2/3$ and $y_2 = 1/3$, *via* integration.[6]

5.3 Strategies

Let $\mathcal{D}$ be the set of all possible privacy-aware dialogues. Let $D = m_0, \ldots, m_n$ be a privacy-aware dialogue, with $m_n = (w, a)$ the last movement used. We define $\eta: \mathcal{D} \rightarrow 2^{\mathcal{A}_x}$ as the function that returns the set of all arguments $r \in \mathcal{A}_x$ such that r attacks a and $m_{n+1} = (\bar{w}, r)$ can be used as the next move in the dialogue, that is, $m_0, \ldots, m_n, m_{n+1}$ is also a privacy-aware dialogue. We enumerate 4 strategies for choosing a rebuttal argument $r \in \eta(D)$.

(1) Random: r is sampled randomly with equal probability for all $r \in \eta(D)$.
(2) Minimum Cost: r is the argument in $\eta(D)$ with lowest cost.
(3) Offensive: r is the argument in $\eta(D)$ that attacks most other arguments in $\mathcal{A}_x$.
(4) Defensive: r is the argument in $\eta(D)$ that suffers the least attacks in $\mathcal{A}_x$.

Let $\alpha \in \mathcal{S}$ be a strategy, where

$$\mathcal{S} = \{\texttt{random}, \texttt{min\_cost}, \texttt{offensive}, \texttt{defensive}\}.$$

A result is defined as an $|A|\times|A|$ matrix **RE**: $\mathbb{Z}^+ \times \mathcal{S} \rightarrow R^{|A|\times|A|}$, where $R = \{pr, op\}$. Every item $\boldsymbol{RE}_{j,k} = \phi(\alpha(q_j, q_k, g), \alpha(q_k, q_j, g))$, for $\boldsymbol{RE}_{j,k}$ represents the dispute resolution outcome of a dispute between agents q_j and q_k, assuming the roles of pr and op, respectively. We instantiate ϕ by carrying out a dialogue between q_j and q_k until a winner is found, returning pr or op accordingly. Analogously to the ground truth model, we can generate a precedence graph $G_{\boldsymbol{RE}(g,\alpha)} = (A, E)$, where for every two distinct agents $q_j, q_k \in A$, we say an arc $(q_j, q_k) \in E$ iff $\boldsymbol{RE}_{j,k} = pr$ and $\boldsymbol{RE}_{k,j} = op$.

5.4 Simulations

Each experiment consists of a set of $|A| = 50$ agents with fixed feature values initialised randomly. For each experiment, we create a random acyclic culture $C = (\mathcal{A}, \mathcal{R}, \mathcal{K})$ with $|\mathcal{A}| = 50$, $|\mathcal{R}| \simeq 400$, and $\mathcal{K} = \{\gamma\}$. We generate an alteroceptive expansion C_x *of* C where the privacy cost $\tau(a_x)$ for all $a_x \in \mathcal{A}_x$ is randomly initialised as $1 \leq \tau(a_x) \leq 20$. Using privacy-aware dialogue games as dispute resolution mechanisms, we generate a precedence graph $G_{\boldsymbol{RE}(g,\alpha)}$ for each strategy $\alpha \in \mathcal{S}$.

We repeat each experiment for integer value privacy budgets $0 \leq g < 60$. We aggregate measures of global and local fairness over 1,900 trials,[7] where each trial randomly initializes all agents' feature descriptions and the culture.

The plots in **Figure 2** show the average subjective local fairness loss (**Figure 2A**) and the average objective global fairness loss (**Figure 2B**) for all four strategies over a range of privacy budget values. All strategies converge to zero subjective local fairness loss given high enough privacy budgets. However, at a given privacy budget value, the defensive strategy dominates the other strategies. Similarly, the defensive strategy is dominant with respect to objective global fairness. In addition to achieving a lower objective global fairness loss at a given privacy budget value, the convergence value of the defensive strategy is lower than that of the other strategies.

5.5 Privacy Efficiency

We run approximately 10 million dialogues between 95,000 different agents, using the same 1,900 random cultures from the previous experiment. However, this time we observe how much privacy cost was used by each strategy to finish the dialogues. Agents were free to extend their dialogues for as long as required. **Figure 3** shows an empirical cumulative distribution function of the proportion of dialogues with cost z, where $z \geq z'$, s.t. $l_{SL}(q_i, q_j, r, z') = 1$, for any $\{q_i, q_j\} \subset A$, $\alpha \in \mathcal{S}$, and any $r \in \{pr, op\}$ (i.e., dialogues that need a privacy budget higher than z to not be cut short by it). This illustrates the effect of different strategies in minimizing privacy cost, even in unrestricted dialogues. Results are consistent with the findings in **Section 5.4**.

6 MULTI-AGENT APPLICATION

Manifestly, the previous empirical analysis explores important, but essentially abstract aspects of the problem. Observing properties under a spectrum of multiple different privacy budgets and with large randomly-generated cultures allows us to observe a space of multiple possible systems. However, realistic applications are not likely to explore an assortment of privacy budgets, as the limits on privacy are likely to be fixed or seldom vary. In like manner, cultures can be explainable representations of real-world rules and their exceptions—and much like in the privacy budgets' case, these are also unlikely to fluctuate in applications predicated in reality.

On these grounds, we build on this conceptual foundation to demonstrate how said aspects may also materialize in an applied setting, even under assumptions of: 1) a single fixed privacy budget, and 2) a single fixed alteroceptive culture. Drawing inspiration from the "Busy Barracks" game seen in (Raymond et al., 2020), we apply our architecture to a multi-agent simulation of speedboats (see **Figure 4**).

[6] We decide to get a representative value from the set of possibilities $0 \leq y_2 < y_1 \leq 1$, so we use the centroid, found by the analytical solution to the integration of K over y_1 from 0 to 1, and y_2 from 0 to y_1.

[7] Total simulation time exceeded 55 h on an `i7-8550U` CPU with 16 GB of RAM.

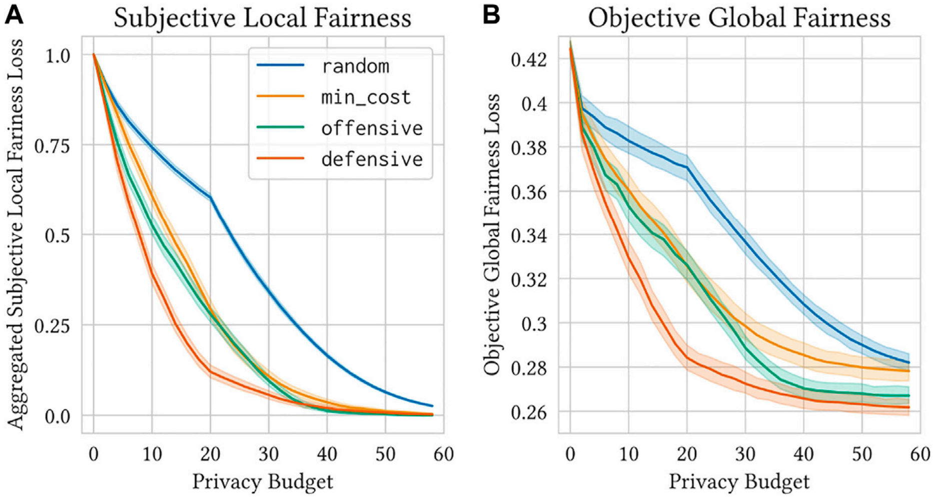

FIGURE 2 | The defensive strategy outperforms the other strategies with respect to both local and global fairness under privacy constraints. We show the average subjective local fairness loss **(A)** and the average objective global fairness loss **(B)** across a range of privacy budgets for `random`, `min_cost`, `offensive`, and `defensive`. The shaded regions show 99% confidence intervals on the mean values. Note that the small change in behavior for g ≥ 20 is due to the fact that *max* $\tau(a_x)$ = 20 (see **Section 5.4**).

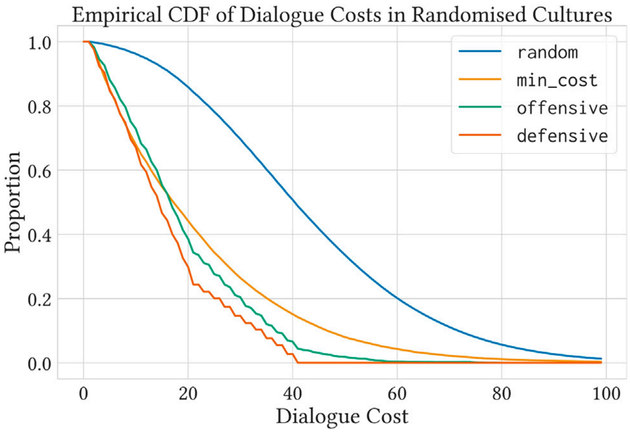

FIGURE 3 | An empirical cumulative distribution function of privacy costs for unrestricted dialogues in over 1900 different random cultures.

6.1 Simulator

Our environment[8] consists of a 20 km long × 2 km wide two-dimensional water surface, without any obstacles. In every trial, 16 agents are instantiated in a ‘head-on parade’ scenario, as follows:

- 8 out of the 16 agents start from the west, and 8 others start from the east.
- Every agent starts with at least 1 km of longitudinal separation from any other agent.
- Their start and goal vertical coordinates are randomised within ± 200 m from the vertical midpoint.
- Every agent’s destination is at the opposite side of the map, and they are initialised with an exact heading towards it. If undisturbed, any agent should be able to execute a perfect straight line towards their destination without any lateral movement or rotation.

Speedboats are simulated with a physics model adapted from Monster (2003) and Linkovich (2016), where elements such as water resistance, drag, and mass are taken into consideration when calculating frames. We simulate and control each boat at 20 Hz refresh rate. Each trial spans around 10–20 min of simulation at 20 Hz for 16 agents, and we collect around 400,000 telemetry data points per trial, including velocity, lateral acceleration, yaw rate, and lateral jerk. We run 900 trials and generate over 40 GB of trajectory and telemetry data for our analyses.

The properties of the vehicle are coarsely modelled after the real-life high performance passenger boat Hawk 38 (Sunseeker, 2019), with capacity for seven seated passengers, dual 400 hp motors, and top speed of 30 m/s (approx. 60 knots). It is assumed that vehicles will drive at maximum speed whenever possible.

6.2 Boat Culture

Akin to **Section 5.1**, each agent is bestowed with properties extracted from a culture, initialised according to a given distribution. **Table 1** illustrates the culture designed for the experiment. To preserve a certain degree of realism, the initialisation of each agent is still random, but respecting certain reasonable restrictions (e.g., civilians cannot have high military ranks or engage in combat, spies are always civilian or corporate, etc.).

[8] The source code is publicly available and can be accessed at https://github.com/alexraymond/privatearg

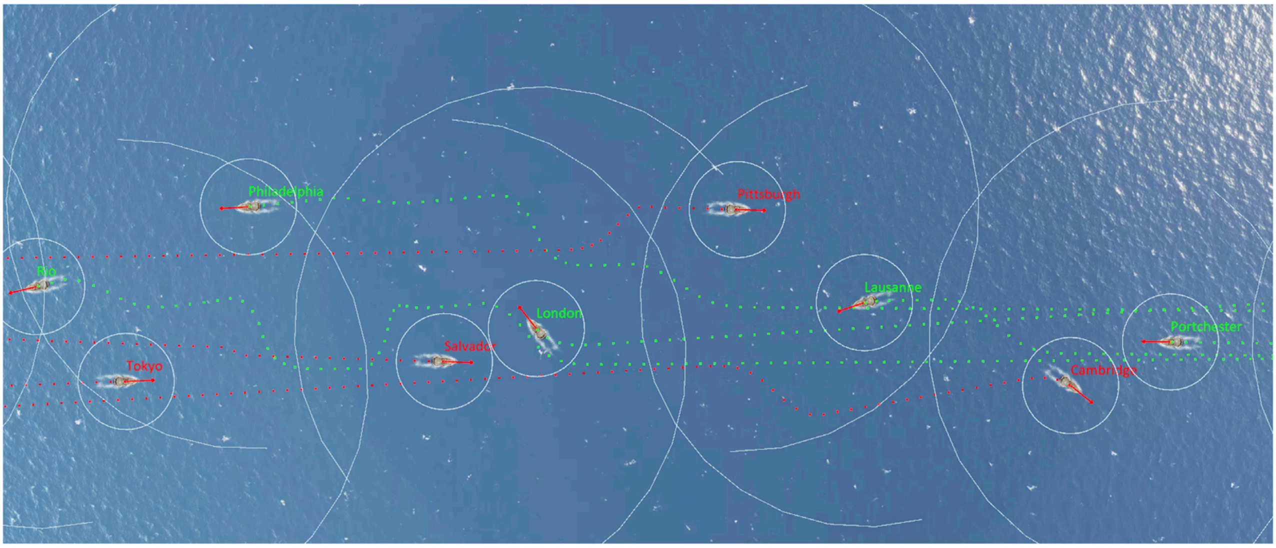

FIGURE 4 | Screenshot of our multi-agent collision avoidance simulation. Speedboats are required to engage in privacy-aware dialogues to resolve who has right-of-way before getting into the risk of collision. Boat and ripple art from Felvégi (2013).

TABLE 1 | Properties present in our culture, along with their possible values for each agent. The “attacks” column indicates which other properties can be defeated by the respective row.

i	Property	*τ*(*i*)	Possible values of *μ* (ascending order of importance)	Attacks
0	Motion	0	—	—
1	VehicleAge	4	{`new, used, worn, old, vintage`}	{0}
2	VehicleCost	10	{`cheap, ok, expensive, very_expensive, millions`}	{0, 1}
3	HigherCategory	0	{`civilian, corporate, police, coast_guard, military`}	{0, 1, 2}
4	TaskedStatus	3	{`at_ease, returning, tasked`}	{0, 1, 2, 3}
5	PayloadType	5	{`empty, food, medical_supplies`}	{0, 1, 2}
6	TaskNature	7	{`leisure, sport, trade, training, patrol, pursuit, combat`}	{4, 5}
7	VIPOnBoard	13	{`ordinary_person, business_person, celebrity, politician`}	{0, 1, 2, 4}
8	MilitaryRank	8	{`no_rank, officer, lieutenant, commander, captain, major, colonel, general, admiral`}	{3, 5, 6, 7}
9	DiplomaticCredentials	12	{`no_credentials, diplomat, united_nations`}	{0. 8}
10	SensitivePayload	15	{`no_sensitive_payload, weapons, wanted_prisoner`}	{0. 9}
11	UndercoverOps	20	{`no_spy, spy`}	{3, 4, 6, 7, 8, 10}
12	EmergencyNature	10	{`no_emergency, mechanical, sick_passenger, fire`}	{0. 11}
13	SuperVIPOnBoard	16	{`no_super_vip, prime_minister, head_of_state`}	{0. 12}

6.3 Dialogues and Collision Avoidance

In our simulation, agents have no prior knowledge of other agents’ descriptions. Properties with $\tau = 0$ can be communicated instantaneously. It is assumed that agents act lawfully and will not lie, but they are nonetheless self-interested and will fight for the right of way in every conflict that arises. In our scenario, we establish that more sensitive information requires further time to obtain clearance before communicating to another agent. Therefore, in this experiment, privacy costs are associated with time (and consequently, distance). We model their avoidance mechanism using a modified artificial potential field algorithm (Warren, 1990).

Let $q_i, q_j \in A$ be any two agents, and $dist\colon A \times A \times \mathbb{Z} \rightarrow \mathbb{Z}^+$ be a function that determines the euclidean distance between two agents at a specific time t. We define two constants $r_{max} = 1{,}000$ and $r_{crit} = 100$ as the maximum effect and the critical minimum radii of a potential field. In our method, when $dist(q_i, q_j, t) > r_{max}$ and $dist(q_i, q_j, t+1) \leq r_{max}$ (i.e., the agents just got within the maximum effect radius for the first time), q_i and q_j will initiate a dialogue game $\phi(q_i, q_j, g)$ to decide who has right of way, where g is the maximum privacy budget for each agent.

Definition 6.1. (Activation Radius). Starting from the centre of each agent, we divide the space between r_{max} and r_{crit} into $2g$ concentric and uniformly-expanding rings. Let $\mathcal{T}(q)$ be the description cost of an agent, that is, how much of its privacy budget was spent to reveal information. Without loss of generality, suppose $q_i \in \{q_i, q_j\}$ is the winner of the dialogue game and $z = \mathcal{T}(q_i) + \mathcal{T}(q_j), z \leq 2g$ is the total combined

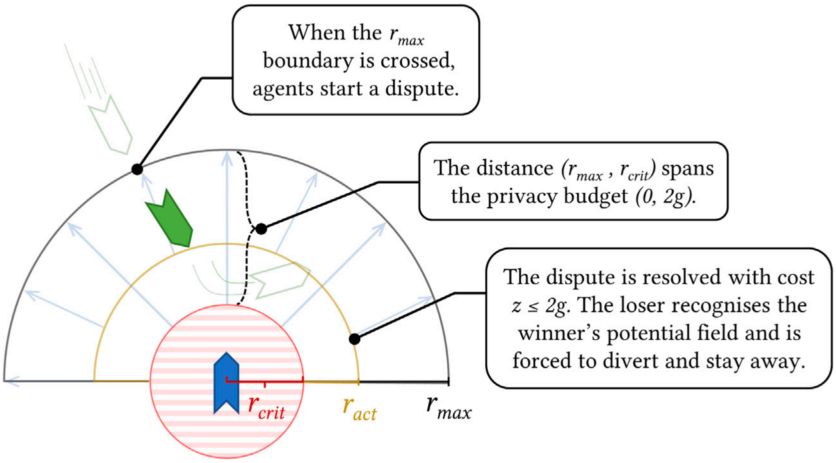
When the r_{max} boundary is crossed, agents start a dispute.
The distance (r_{max}, r_{crit}) spans the privacy budget $(0, 2g)$.
The dispute is resolved with cost $z \le 2g$. The loser recognises the winner's potential field and is forced to divert and stay away.
r_{crit}
r_{act}
r_{max}

FIGURE 5 | Illustration of the collision avoidance mechanism from Definition 6.1.

privacy cost for that dialogue, considering both agents. The activation radius

$$r_{\text{act}} = r_{\max} - \frac{z}{2g}(r_{\max} - r_{\text{crit}})$$

is the distance where the potential field (with full radius r_{max}) will be enabled for the first time for the losing agent, and will remain in effect until the end of the simulation.

In plain language, agents will start the dialogue as soon as they cross the r_{max} = 1,000 boundary, and their cumulative privacy cost z (e.g. the time the dialogue takes) will determine how early or how late they ultimately decide which agent should have right of way (see **Figure 5**). Higher values of z will lead to late, aggressive evasive manoeuvres. In case $dist(q_i, q_j) < r_{crit}$, the winner is also forced to divert to avoid a collision. In an ideal scenario, most agents will complete their dialogues in full with very low privacy costs and execute early, smooth manoeuvres to stay clear of other agents. Conversely, if dialogues are wasteful or activation radii are closer to r_{crit} than r_{max}, then one of the following may happen in the local scope, for any q_i, $q_j \in A$ and $r \in \{pr, op\}$:

- the evader makes a very late and aggressive evasion turn (low privacy efficiency: high $\mathcal{T}(\alpha(q_i, q_j, g))$);
- the late evader accepts defeat but breaks into the critical radius of the winner and forces them out of their rightful trajectory (objective unfairness: $l_{OL}(q_i, q_j, g) = 1$ and $l_{SL}(q_i, q_j, r, g) = 0$);
- the right agent wins the dialogue game, but the losing agent is not fully convinced all their reasons are covered (subjective unfairness: $l_{OL}(q_i, q_j, g) = 0$ and $l_{SL}(q_i, q_j, r, g) = 1$);
- the wrong agent wins the dialogue game and forces the right one out of their way since they had no budget remaining to get to the correct decision (subjective and objective unfairness: $l_{OL}(q_i, q_j, g) = 1$ and $l_{SL}(q_i, q_j, r, g) = 1$).

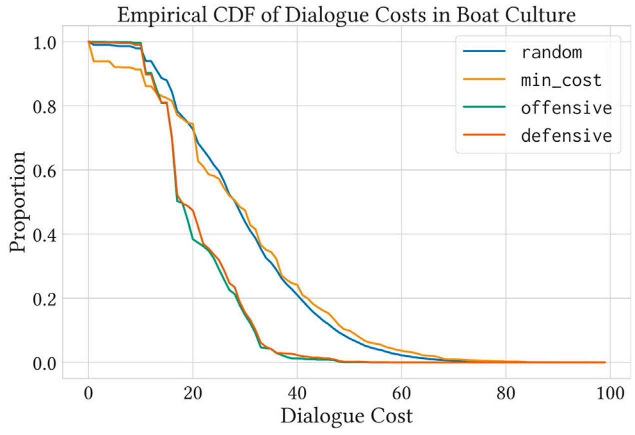

FIGURE 6 | Privacy costs for unrestricted dialogues in our instantiated Boat Culture (as shown in **Table 1**). Note that some behaviors are different from **Figure 3**. Contrary to the randomised case, where over 2000 different cultures are tested, this plot represents dialogues under a single culture.

As per our previous abstract experiment, we demonstrated that different strategies for choosing arguments (explanations) during the dialogue game leads to varied levels of performance with regards to privacy efficiency, subjective fairness, and objective fairness. We will perform simulations using the same strategies seen in **Section 5.3** in the sections below.

6.4 Privacy Efficiency Experiments

Since this experiment involves a real simulation of a multi-agent system, its magnitude is significantly smaller in terms of number of dialogues and agents. We repeat the experiment in **Section 5.5**, this time with only one culture. We observe the privacy costs of 48,000 dialogues between the previous set of 1,600 randomly-initialised agents for each strategy without any restrictions in privacy.

Figure 6 shows the effect of different strategies in minimizing privacy cost for unrestricted dialogues in the Boat Culture. Two phenomena can be observed: the greedy strategy for minimizing cost (min_cost) turns out to be the worse at that. This is due to the fact that, as privacy costs are not randomly determined this time, privacy costs are correlated with argument relevance/power, making stronger arguments more expensive. Additionally, both offensive and defensive exhibit similar performance. In a similar nature, in this culture, stronger arguments simultaneously attack many others, whilst being relatively unattacked.

6.5 Ride Quality

A more applied perspective on the same matter is to observe the status of vehicles across their trajectories. In our simulation, most disturbances and evasions take place when both sides of the parade meet in the middle. These conflicts generate peaks in lateral acceleration and other measurements. When those conflicts are resolved, agents usually follow trouble-free paths to their destinations. We are interested in quantifying a general notion of "ride comfort" (supposing an acceptable limit of "comfort" for a powerboat at maximum speed) and minimizing the "ride roughness" for the fictitious average passenger on board. We discard the initial acceleration and final braking data points, as we are only interested in changes in acceleration caused by other agents.

To do so, we integrate the area under each of those metrics per agent, and compute the mean of each agent's integrals per trial. We run 100 trials per strategy with a combined privacy budget value of $2g = 60$ and aggregate these values. We associate higher values of lateral acceleration and jerk represent as strong components in a passenger's experience of discomfort (Nguyen et al., 2019).

The first row in **Figure 8** shows the distribution of lateral acceleration, yaw rate, and lateral jerk measurements over 1,600 simulations (16 agents x 100 trials) per strategy. We observe significantly lower values from offensive and defensive in all metrics. The random strategy also exhibited significantly lower values of lateral jerk compared to min_cost.

6.6 Subjective and Objective Trajectories

After observing the impact of privacy on the quality of the ride, we now measure agents' perspectives in a global scope. Instead of counting local instances of objective and subjective unfairness, we generate entire trajectories that cater to specific requirements. From this point, we will apply the following nomenclature:

- Nominal trajectory ($\mathcal{J}_N$): this is the actual trajectory executed by the agent in a regular trial. Agents perform their dialogues normally and give way in case of defeat (as shown in **Figure 5**), according to the current strategy mutually adopted by all agents.
- (Hypothetical) Subjective trajectory ($\mathcal{J}_S^h$): same as Nominal, but whenever an agent loses a dialogue that it deems unfair, it hypothesises what would happen if the agent ignored the potential field of the opponent and forcefully drove straight through. In this case, agents will never "agree to disagree" and will not give way if the dialogue ends due to privacy limitations.
- (Hypothetical) Objective trajectory ($\mathcal{J}_O^h$): this is a trajectory where all conflict resolution is optimal and derives from the ground truth extensions. Dialogues do not exist and agents have perfect information at cost 0. It hypothesises what would happen if all agents had full mutual knowledge and always made the right decisions.

These trajectories are advantageously intuitive and serve as an example of application-based representations of policies under different perspectives. An illustration of these trajectories can be seen in **Figure 7**. If agents have efficient strategies for selecting relevant information/explanations, their Nominal trajectories will converge to the objectively correct answer more frequently (and more quickly), thus becoming increasingly similar to the Objective hypothesis. Likewise, if strategies are efficient, subjective unfairness will be reduced as agents will be able to get to the end of their dialogues—and thus the decisions made in the Nominal trajectory and its corresponding Subjective hypothesis will agree.

Let $\mathcal{J}_1, \mathcal{J}_2$ denote any 2 trajectories. We calculate dissimilarity measures across comparable trajectories using the discrete Fréchet distance[9] $Frec(\mathcal{J}_1, \mathcal{J}_2)$. We repeated the tests with two other methods (Partial Curve Mapping and Dynamic Time Warp) and achieved similar results. We chose the aforementioned metric for its intuitive value and simplicity, and will omit the other results as they are redundant. For every strategy, we will compare 3 pairs of trajectories:

- Objective Unfairness ($\Omega = Frec(\mathcal{J}_N, \mathcal{J}_O^h)$). This represents the level of agreement between the Nominal trajectory and the hypothetical Objective trajectory in a perfect scenario. This mostly captures how objectively correct the decisions were. In **Figure 8**, it is possible to observe that offensive and defensive exhibit much more agreement with the correct trajectory than the other two strategies.
- Subjective Unfairness ($\Omega_p = Frec(\mathcal{J}_N, \mathcal{J}_S^h)$). This represents the level of agreement between the Nominal trajectory and what the agent considers to have been the right action, i.e. the hypothetical Subjective trajectory. High levels of disagreement in are representative of high global subjective unfairness—as many agents have very different perceptions of what the right trajectories were. If privacy budgets are minimal and communication is near-impossible, most subjective trajectories will be straight lines, as no agent will ever be fully convinced to alter their route.
- Subjectivity Gap ($l_S = Frec(\mathcal{J}_S^h, \mathcal{J}_O^h)$). This is an assessment on how accurate is the agents' perception of correctness to the actual correctness. This is measured as the agreement between the hypothetical Subjective and Objective trajectories. We name this phenomenon as the "subjectivity gap."

[9] "Informally, it is the minimum length of a leash required to connect a dog, walking along a trajectory $\mathcal{J}_1$, and its owner, walking along a trajectory $\mathcal{J}_2$, as they walk without backtracking along their respective curves from one endpoint to the other."—Agarwal et al. (2014).

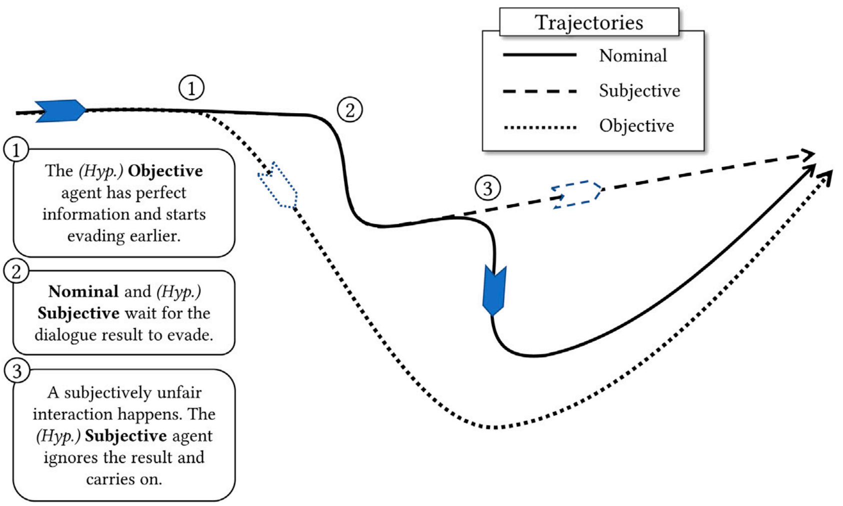
Trajectories
Nominal
Subjective
Objective
1
2
3
1 The (Hyp.) Objective agent has perfect information and starts evading earlier.
2 Nominal and (Hyp.) Subjective wait for the dialogue result to evade.
3 A subjectively unfair interaction happens. The (Hyp.) Subjective agent ignores the result and carries on.

FIGURE 7 | An example of a Nominal trajectory and its hypothetical Subjective and Objective scenarios. The full line (Nominal) represents the actual trajectory performed by the agent in their trial. The Objective hypothesis (dotted line) can perform smoother manoeuvres and evade much sooner due to perfect information being immediately available at r_{max}. The Subjective hypothesis (dashed line) behaves like the Nominal trajectory, but will not concede upon subjective unfairness cases.

The results in the bottom row of **Figure 8** show aggregated results for objective and subjective unfairness, as well as the subjectivity gap for each strategy. We perform pairwise comparisons to evaluate how different fare against each other in terms of higher ride comfort (lower dynamics) and lower unfairness values. Higher values in the first row indicate that, when using that strategy, agents experienced higher motion dynamics associated with discomfort. Likewise, higher values in the second row demonstrate a higher disagreement between the pairs of trajectories when agents use that strategy. We observe that `min_cost` exhibits the worst performance in all metrics against all other strategies. In second-to-last, `random` is bested by the other 2 strategies. More remarkably, the winning strategies `offensive` and `defensive` exhibit mostly equivalent behavior, and both are able to practically eradicate occurrences of subjective unfairness within that given privacy budget.

7 DISCUSSION

Our experimental results explored four intuitive strategies for conflict resolution under privacy constraints: `random`, which does not exploit any problem structure; `min_cost`, which greedily exploits the cost structure; and `offensive` and `defensive`, which are informed by the structure of the underlying culture. The strategy performance follows an expected order. Any informed strategy outperforms the baseline `random`. The exploitation of the culture structure in `offensive` and `defensive` yields better results than just using privacy information, as in `min_cost`. Finally, `defensive` is expected to be the best strategy, since it follows a more natural principle in argumentation: that an argument which has no attackers must be accepted. In this scenario, choosing arguments with fewer attackers makes it more likely that this argument is final. Although many others (more involved) strategies could be used, our goal is to measure the effect of using informed strategies as opposed to strategies unaffected by the argumentative structure (the culture). That is, whether a better usage of the argumentative information leads to better results in both subjective and objective fairness, and other performance metrics in the experimental scenario, under identical privacy constraints.

7.1 Randomised Cultures

In the random experiment, **Figure 2** shows that `defensive` yields lower losses both locally and globally, especially for realistic privacy budgets that do not overly limit the length of dialogue while meaningfully constraining the amount of shared information. When unrestricting dialogues, `defensive` was also able to close dialogues with lower privacy costs. As expected, our results suggest that choosing better strategies imply Pareto dominance.

When the privacy budget is near zero, all strategies perform similarly poorly. Little dialogue can occur, so agents lack the information to make objectively good decisions and to provide meaningful justifications. With high privacy budgets, strategies perform similarly well. Predictably, all strategies converge to zero subjective local fairness because dialogues can extend indefinitely. The distinct non-zero convergence values of the objective global fairness loss is due primarily to the ground truth used in these experiments.

Our ground truth corresponds to the sceptical acceptance of the proponent's motion. The motion is only validated if all attacking arguments are defeated. Having at least one reason to defeat the motion is sufficient to preserve the status quo.[10] If

[10] Since features are randomly uniformly sampled, the likelihood of a victory is associated with the probability of an agent prevailing in multiple random challenges.

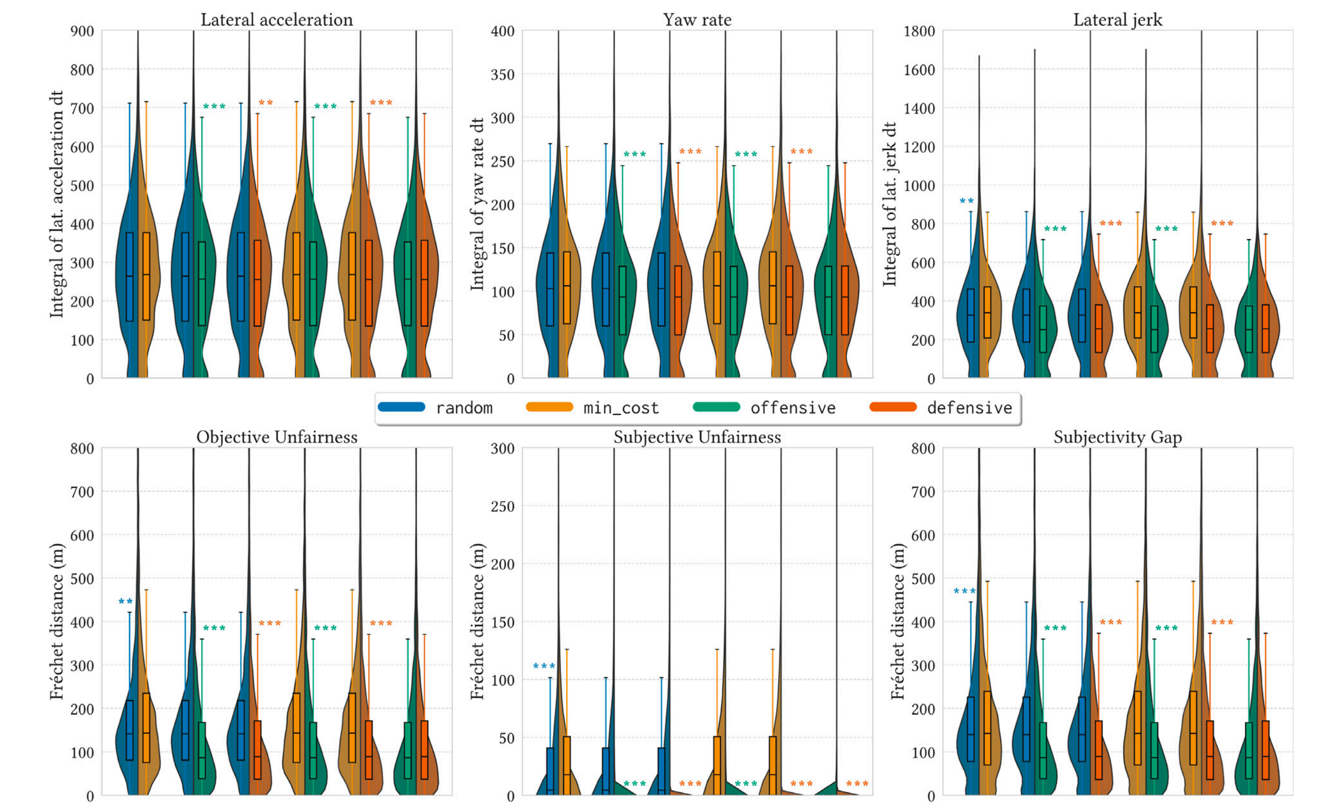

FIGURE 8 | Violin plots of aggregated data of 100 full simulations with a fixed combined privacy budget $2g = 60$. Each strategy is shown in a different color. We perform pairwise statistical two-sided and one-sided Student's t-tests between distributions to check if they differ significantly, and if so, which one presents lower values ((*) $p < .05$; (**) $p < .01$; (***) $p < .001$). The top row shows an aggregation of all agents' simulation metrics (n = 1,600 per strategy), where lower is indicative of a more comfortable ride. We plot the curves for measurements of lateral acceleration, yaw rate, and lateral jerk for all agents. Each data point in the top plots is the area under the curve for that metric (obtained *via* integration). The bottom plots show values of objective unfairness (Ω), subjective unfairness (Ω_p) and the subjectivity gap (I_S), measured as the Fréchet distance of the corresponding trajectories (n = 32,000 per strategy). Higher values mean that the nominal/subjective trajectories differed more from the compared subjective/objective trajectories. Lower values are indicative of more agreement between trajectories.

both agents win in some unattacked arguments and lose in others, the outcome will always lean towards the opponent, even when they swap roles.

7.2 Multi-Agent Experiment

We observed similar results in the applied multi-agent experiment. Although the distinction between `defensive` and `offensive` was not as clear as in with randomised cultures, both still exhibited superior results and reinforced the argument towards the importance of selecting information by relevance. The formalism proposed in **Section 4** allows for representing a range of different problems, endowing decentralised systems with considerations of subjectivity that can be modeled for non-human agents.

By comparing trajectories under different perspectives, one could be tempted to see the Objective hypothesis through the lens of a more "traditional" aspect of trajectories, namely, by how much shorter and quicker they are, or by observing the system behavior with metrics such as makespan or flowtime. In our case, better trajectories are not necessarily "optimal" in the traditional sense. They do not attempt to be shorter or quicker, but instead consider ride comfort metrics and ultimately select for better agreement between what actually happens, what actually should have happened, and what the agents believe should have happened.

The environment with 16 agents avoiding each other through combinations of artificial potential fields in continuous space with simulated physics is dense and complicated, and can lead to edge cases where agents' decisions can cascade and propagate to multiple others (especially if radii are large). Notwithstanding all the potential for results being confounded by this property of the system, we still managed to demonstrate statistically significant results for superior global performance with better strategies. We expect even clearer results in discrete applications, topological representations (Bhattacharya et al., 2012), or in environments with planning in mind, such as Conflict-Based Search methods (Sharon et al., 2015).

For real applications, the properties and structure of the chosen alteroceptive culture play an important role in indicating which strategy will perform better towards the aimed Pareto dominance within the fairness-privacy trade-off. Perhaps with the exception of degenerate cases of cultures, our heuristics should provide a good guide for considering the choice of arguments for building explanations in fairness-aware conflict resolution disputes, especially if one cannot assess the relevance of the content of the arguments (what it actually means), but has access to its structure (the ruleset that originated the culture, and the relationships between rules).

8 CONCLUSION

In this work, we propose perspective and scope as new considerations for the problem of fairness in decentralised conflict resolution. We show how privacy limitations introduce partial observability, and consequently, a trade-off for fairness losses. Our proposed architecture for privacy-aware conflict resolution allows for the representation and resolution of such conflicts in multi-agent settings, and underpins our experimental setup.

Our central insight shows that simulating agents' actions from the exclusive perspective of their local knowledge and comparing them to the actual executed behavior can grant an understanding of their subjective perceptions of fairness. Moreover, comparing this subjective perception to an omniscient objective behavior can also provide an insight on how well-informed agents are in their perception of the world. Different environments and applications could yield interesting combinations of these properties. For instance, a multi-agent society with low global subjective unfairness and a high subjectivity gap could indicate that agents are not acting in accordance to the desired notion of fairness, either by ignorance, design flaw, or adversarial intent.

This work presents no shortage of avenues for future studies. To name a few, extending the scope of interactions beyond pairwise conflict resolution can be investigated by means of collective argumentation (Bodanza et al., 2017). One also could observe the effects of the fairness-privacy trade-off in populations with heterogeneous strategies, and, in like manner, the consequences of both heterogeneous privacy budget distributions and agent-dependent privacy costs for subjective and objective fairness, or even extending the dialogues to more general dispute trees, allowing agents to backtrack and find new lines of defence. To a wider regard, the use of cultures as mechanisms for explainable conflict resolution in rule-aware environments encourages similar reflection for our work. Our findings further reinforce the importance of (good) explanations and explanatory strategies in multi-agent systems, this time as mitigatory instruments for applications where subjectivity is a concern.

Inescapably, agents "agree to disagree" when compelled to concede a resource due to ulterior reasons, such as privacy preservation or time constraints. Whenever continuing the dispute would produce effects more undesirable than the loss of the dispute in itself, agents tolerate those consequences and will prefer to abandon the dispute. This phenomenon is not necessarily new for humans, but extending this consideration for societies of artificial agents grants a dimension of subjectivity that is worth exploring in multi-agent systems research. Preparing non-human agents and systems to consider subjective unfairness can be an important step in facilitating the integration of human agents in future hybrid multi-agent societies. We encourage readers and researchers in the field to regard subjectivity as an important consideration in fairness-aware decentralised systems with incomplete information.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://github.com/alexraymond/privatearg.

AUTHOR CONTRIBUTIONS

All authors have significantly contributed to the ideas presented in the paper. More specifically, AR designed the alteroceptive culture formalism and part of the fairness formalism, conducted the random experiments, and wrote the multi-agent simulator code. MM and AR conceptualised and refined the original idea of the paper. MM provided expert input in fairness, as well as in designing the Boat Culture. GP-P provided expert input in argumentation frameworks and in the mathematical formalism of **Section 4**. The paper was originally drafted by AR and revised by all authors. AP directed the research efforts.

FUNDING

AR was supported by the Royal Commission for the Exhibition of 1851 and L3Harris ASV. MM was supported by the National Science Foundation Graduate Research Fellowship under Grant DGE-1845298. GP-P was supported by CAPES (Brazil, Ph.D. Scholarship 88881.174481/2018-01). AP was supported by the Engineering and Physical Sciences Research Council (grant EP/S015493/1). Their support is gratefully acknowledged.

REFERENCES

Agarwal, P. K., Avraham, R. B., Kaplan, H., and Sharir, M. (2014). Computing the Discrete Fréchet Distance in Subquadratic Time. *SIAM J. Comput.* 43, 429–449. doi:10.1137/130920526

Amgoud, L., and Prade, H. (2009). Using Arguments for Making and Explaining Decisions. *Artif. Intell.* 173, 413–436. doi:10.1016/j.artint.2008.11.006

Bertsimas, D., Farias, V. F., and Trichakis, N. (2013). Fairness, Efficiency, and Flexibility in Organ Allocation for Kidney Transplantation. *Oper. Res.* 61, 73–87. doi:10.1287/opre.1120.1138

Bhattacharya, S., Likhachev, M., and Kumar, V. (2012). Topological Constraints in Search-Based Robot Path Planning. *Auton. Robot* 33, 273–290. doi:10.1007/s10514-012-9304-1

Binmore, K. (1992). *Fun and Games. A Text on Game Theory*. Lexington, MA: D.C. Heath.

Bin-Obaid, H. S., and Trafalis, T. B. (2018). "Fairness in Resource Allocation: Foundation and Applications," in International Conference on Network Analysis, Moscow, Russia, May 18–19, 2018 (Springer), 3–18.

Blair, J. R. S., Mutchler, D., and Liu, C. (1993). "Games with Imperfect Information,". AAAI Press Technical Report FS93-02 in Proceedings of the AAAI Fall Symposium on Games: Planning and Learning, Menlo Park CA, 59–67.

Bodanza, G., Tohmé, F., and Auday, M. (2017). Collective Argumentation: A Survey of Aggregation Issues Around Argumentation Frameworks. *Aac* 8, 1–34. doi:10.3233/AAC-160014

Čyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., et al. (2019). Explanations by Arbitrated Argumentative Dispute. *Expert Syst. Appl.* 127, 141–156. doi:10.1016/j.eswa.2019.03.012

Doutre, S., and Mengin, J. (2001). "Preferred Extensions of Argumentation Frameworks: Query, Answering, and Computation," in International Joint Conference on Automated Reasoning, Siena, Italy, June 18–22, 2001, 272–288. doi:10.1007/3-540-45744-5_20

Dung, P. M. (1995). On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-Person Games. *Artif. Intell.* 77, 321–357. doi:10.1016/0004-3702(94)00041-x

Dwork, C., and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Found. Trends® Theor. Comput. Sci.* 9, 211–407. doi:10.1561/0400000042

Emelianov, V., Arvanitakis, G., Gast, N., Gummadi, K., and Loiseau, P. (2019). "Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, (IJCAI-19)," in International Joint Conferences on Artificial Intelligence Organization, Macao, China, August 10–16, 2019, 5836–5842. doi:10.24963/ijcai.2019/809

Fan, X., and Toni, F. (2015). "On Computing Explanations in Argumentation," in Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, USA, January 25–30, 2015.

Felvégi, C. (2013). *Ships with Ripple Effect*. Webpage: https://opengameart.org/content/ships-with-ripple-effect.

Gao, Y., Toni, F., Wang, H., and Xu, F. (2016). "Argumentation-based Multi-Agent Decision Making with Privacy Preserved," in Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9–13, 2016, 1153–1161.

Jakobovits, H., and Vermeir, D. (1999). "Dialectic Semantics for Argumentation Frameworks," in Proceedings of the seventh international conference on Artificial intelligence and law - ICAIL '99, Oslo, Norway, June 14–17, 1999 (New York, New York, USA: ACM Press), 53–62. doi:10.1145/323706.323715

Kerkmann, A. M., Nguyen, N.-T., and Rothe, J. (2021). "Local Fairness in Hedonic Games via Individual Threshold Coalitions. *Theoretical Computer Science* 877, 1–17. ISSN 0304-3975.

Li, M., and Tracer, D. P. (2017). *Interdisciplinary Perspectives on Fairness, Equity, and Justice*. Denver, USA: Springer.

Linkovich, M. (2016). *carphysics2d*. Webpage: https://github.com/spacejack/carphysics2d.

Malmi, E., Tatti, N., and Gionis, A. (2015). Beyond Rankings: Comparing Directed Acyclic Graphs. *Data Mining knowl. Discov.* 29, 1233–1257. doi:10.1007/s10618-015-0406-1

Marx, K. (1875). *Critique of the Social Democratic Program of Gotha* (Letter to Bracke, pp. 13–30). Moscow: Progress Publishers.

Modgil, S., and Luck, M. (2009). "Argumentation Based Resolution of Conflicts between Desires and Normative Goals," in *Argumentation in Multi-Agent Systems* (Berlin, Heidelberg: Springer), 19–36. doi:10.1007/978-3-642-00207-6_2

Monster, M. (2013). *Car Physics for Games*. Webpage: https://asawicki.info/Mirror/Car%20Physics%20for%20Games/Car%20Physics%20for%20Games.html.

Moskop, J. C., and Iserson, K. V. (2007). Triage in Medicine, Part II: Underlying Values and Principles. *Ann. Emerg. Med.* 49, 282–287. doi:10.1016/j.annemergmed.2006.07.012

Narayanan, A. (2018). "Translation Tutorial: 21 Fairness Definitions and Their Politics," in Proceeding of the. Conference. Fairness, Accountability and Transparency, February 23–24, 2018, New York, USA..

Nguyen, T., NguyenDinh, N., Lechner, B., and Wong, Y. D. (2019). Insight into the Lateral Ride Discomfort Thresholds of Young-Adult Bus Passengers at Multiple Postures: Case of Singapore. *Case Stud. Transp. Pol.* 7, 617–627. doi:10.1016/j.cstp.2019.07.002

Niskanen, A., and Järvisalo, M. (2020). "μ-Toksia: An Efficient Abstract Argumentation Reasoner," in Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020), Greece. September 12–18, 2020. (United States: AAAI Press). doi:10.24963/kr.2020/82

Prorok, A., and Kumar, V. (2017). "Privacy-preserving Vehicle Assignment for Mobility-On-Demand Systems," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada, September 24–28, 2017, 1869–1876. doi:10.1109/iros.2017.8206003

Rawls, J. (1991). "Justice as Fairness: Political Not Metaphysical," in *Equality and Liberty* (Springer), 145–173. doi:10.1007/978-1-349-21763-2_10

Raymond, A., Gunes, H., and Prorok, A. (2020). "Culture-Based Explainable Human-Agent Deconfliction," in Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, Auckland, New Zealand, May 9–13, 2020. (Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems), 1107–1115. AAMAS '20.

Reif, J. H. (1984). The Complexity of Two-Player Games of Incomplete Information. *J. Comput. Syst. Sci.* 29, 274–301. doi:10.1016/0022-0000(84)90034-5

Rosenfeld, A., and Richardson, A. (2019). Explainability in Human–Agent Systems. *Autonom. Agents Multi-Agent Syst.* 33, 673–705. doi:10.1007/s10458-019-09408-y

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). "Fairness and Abstraction in Sociotechnical Systems," in Proceedings of the Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA (New York, NY, USA: Association for Computing Machinery), 59–68. FAT* '19. doi:10.1145/3287560.3287598

Sharon, G., Stern, R., Felner, A., and Sturtevant, N. R. (2015). Conflict-based Search for Optimal Multi-Agent Pathfinding. *Artif. Intell.* 219, 40–66. doi:10.1016/j.artint.2014.11.006

Sovrano, F., and Vitali, F. (2021). "From Philosophy to Interfaces: An Explanatory Method and a Tool Inspired by Achinstein's Theory of Explanation," in 26th International Conference on Intelligent User Interfaces, College Station, USA, April 14–17, 2021. (New York, NY, USA: Association for Computing Machinery), 81–91. IUI '21. doi:10.1145/3397481.3450655

Such, J. M., Espinosa, A., and Garcia-Fornes, A. (2014). A Survey of Privacy in Multi-Agent Systems. *Knowl. Eng. Rev.* 29, 314–344. doi:10.1017/S0269888913000180

Sunseeker (2019). *Sunseeker Hawk 38 Brochure*. Webpage: https://www.sunseeker.com/wp-content/uploads/2019/10/Sunseeker_Hawk38_TechBrochure_Sept19-2.pdf.

Torreno, A., Onaindia, E., Komenda, A., and Štolba, M. (2017). Cooperative Multi-Agent Planning: A Survey. *ACM Comput. Surv. (Csur)* 50, 1–32. doi:10.1145/3128584

Verma, S., and Rubin, J. (2018). "Fairness Definitions Explained," in 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), Gothenburg, Sweden, May 29 2018, 1–7. doi:10.1145/3194770.3194776

Warren, C. W. (1990). "Multiple Robot Path Coordination Using Artificial Potential fields," in Proceedings., IEEE International Conference on Robotics and Automation, Cincinnati, USA, May 13–18 1990, 500–505. doi:10.1109/ROBOT.1990.126028

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.