Culture-Based Explainable Human-Agent Deconfliction

Alex Raymond
University of Cambridge
Cambridge, United Kingdom
alex.raymond@cl.cam.ac.uk

Hatice Gunes
University of Cambridge
Cambridge, United Kingdom
hatice.gunes@cl.cam.ac.uk

Amanda Prorok
University of Cambridge
Cambridge, United Kingdom
asp45@cam.ac.uk

ABSTRACT

Law codes and regulations help organise societies for centuries, and as AI systems gain more autonomy, we question how human-agent systems can operate as peers under the same norms, especially when resources are contended. We posit that agents must be accountable and explainable by referring to which rules justify their decisions. The need for explanations is associated with user acceptance and trust. This paper's contribution is twofold: *i)* we propose an argumentation-based human-agent architecture to map human regulations into a *culture* for artificial agents with explainable behaviour. Our architecture leans on the notion of argumentative dialogues and generates explanations from the history of such dialogues; and *ii)* we validate our architecture with a user study in the context of human-agent path deconfliction. Our results show that explanations provide a significantly higher improvement in human performance when systems are more complex. Consequently, we argue that the criteria defining the need of explanations should also consider the complexity of a system. Qualitative findings show that when rules are more complex, explanations significantly reduce the perception of challenge for humans.

KEYWORDS

Human-Agent Systems; Argumentation; Explanation; User Studies

ACM Reference Format:
Alex Raymond, Hatice Gunes, and Amanda Prorok. 2020. Culture-Based Explainable Human-Agent Deconfliction. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

The Code of Ur-Nammu is the oldest known written law code, inscribed around 2100 BC in ancient Mesopotamia [12]. Its structure is a set of rules (carved in a stone tablet) designed to aid denizens to settle potential conflicts. Conceptually, little has changed since then, as humans historically and currently rely on sets of rules to specify their own systems' behaviours, expecting peers to abide by those regulations when conflicts arise. Different regimens are defined for several environments, be it traffic, competitive sports, business, civil society, etc.

Inasmuch as robots and intelligent artificial agents progress in sophistication, they obtain increasingly more autonomy and start taking part in the same systems and societies that humans do, no longer as tools, but rather as peers. It is therefore fundamental to guarantee that agents embedded in those environments will also observe and respect the same rules and regulations that humans do in order to resolve conflicts and operate orderly [5]. The need for rule-abiding behaviour goes beyond the debate of ethics and morals [21], which is vast but out of the scope of this investigation. Instead, we are interested in ensuring that autonomous agents consider their *liability* [3] and can express justification for their agency [18] *with regards to the present ruleset to be followed* — preferably in a human-understandable way.

Rizaldi and Althoff [22] tackle the liability and accountability problem in the autonomous vehicle domain with a manual formalisation of specific traffic rules, using automated theorem proving techniques. Their approach is formally rigorous but is specialised for a specific set of traffic rules only and does not generalise beyond. Cranefield et al. [9] propose that ideal accountable agents must: *i)* understand what is expected from them (from rules/obligations); *ii)* answer queries about their decision-making (being explainable); *iii)* carry out argumentative dialogues in which beliefs and plans are challenged and justified; *iv)* adapt their reasoning apparatuses or update their plans as a result of accountability dialogues; and *v)* take human values into account when reasoning.

The quest for *understanding* and *explaining* the decisions made by artificially intelligent systems and agents motivated the materialisation of the *eXplainable AI* (XAI) [13] research field. The onset of machine learning systems and the popularity of methods such as support vector machines and artificial neural networks have led to AI solutions that are efficient but indecipherable regarding their rationale behind a conclusion. For that reason, the XAI community is interested in systems that are not only clear regarding *'how'*, but also as to *'why'* certain decisions were made [2, 24].

In order to achieve more realistic explainability for humans, we spur the necessity for more realistic models of reasoning. Expressly, classical logic does not provide an authentic representation of common sense reasoning, as under a scenario of incomplete information a human may draw conclusions that can be withdrawn later, when new information is presented [1]. Argumentation-based approaches attempt to fill in this gap by providing a framework for defeasible reasoning [10], which grants systems clear decision-making mechanisms that provide not only resolutions, but also the *reasons* that may support it [27].

Argumentation approaches walk hand in hand with the desiderata proposed by Cranefield et al. [9], as they allow us to: *i)* enable norm-aware reasoning [4]; *ii)* generate explanations [11, 25]; *iii)* carry argumentative dialogues to support their positions [1, 23]; *iv)* perform meta-reasoning [26]; and *v)* consider human values [16]. Consequently, we regard argumentation frameworks as a strong mechanism for providing accountable and explainable agency.

Rosenfeld and Richardson [24] postulate that the *necessity* for explainability in human-agent systems follows a taxonomy of three types of explanations: *not helpful*, *beneficial*, and *critical*. They posit

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

that if humans will not accept a system without an explanation, then the need for explainability is critical. Likewise, explanations can range in significance depending on their ability to engender trust in human users. We aim to introduce another dimension to this analysis, by empirically observing that the complexity of the rules governing a system may also affect the usefulness of explanations where human performance is concerned.

In this work, we address environments where humans and agents act with independent agency and are subjected to the same rules and conditions. Most explainable approaches in human-agent systems are classified with regards to their human-centric or agent-centric [24] approaches, but relatively few are interested in emulating human-agent societies [7]. Can agents and humans with individual goals coexist as peers in a norm-aware environment where resources are limited? Can such peers resolve conflicts and provide accountability to their decisions, to both humans and other agents alike? Namely, given a multi-agent environment with resource contention, can we define a mechanism that allows us to facilitate human-agent integration by providing: *i)* an equivalence between human-readable rulesets and agent policies and *ii)* in a way that is explainable and allows humans to interact successfully with agents to resolve conflicts?

This paper offers the following contributions:

- We propose an argumentation-based architecture for designing explainable human-agent systems for deconfliction environments.
- We present an empirical study to investigate the effect of explanations in this architecture in varying levels of complexity.

We exemplify our architecture with a multi-agent resource contention application in the context of the problem of *multi-agent path deconfliction*. We show how humans and agents can deconflict trajectories whilst respecting externally-defined "rules of way."

Towards this end, we design a computer game implementing the proposed architecture and conduct a user study to evaluate it. Humans are given path deconfliction rulesets with different amounts of rules each and are asked to navigate in a multi-agent environment and avoid collisions with agents. In our setting, we define complexity as the number of rules that govern the deconfliction of resources. We observe how humans perform in terms of ruleset complexity and the presence/absence of explanations. Our results show that the benefit of explanations is correlated with the complexity of the underlying system. Qualitative results show that human experience in systems with explanations is superior when such systems are sufficiently complex.

2 BACKGROUND

In this section, we introduce definitions and concepts that are used in the construction of our architecture. Section 2.1 introduces essential definitions for Abstract Argumentation frameworks, the principal *deliberation* tool in our framework. Section 2.2 presents a mechanism for dialogical exchanges between agents in Abstract Argumentation, followed by an argumentation-based formalism of explanations in Section 2.3.

2.1 Abstract Argumentation

A seminal paper from Dung [10] introduces the concept of an *argumentation framework*, also called *abstract argumentation* (AA). His framework considers arguments as purely abstract entities, with no special attention paid to their internal structure. Modelling occurs at the level of relationships between those abstract entities.

The main concept behind AA is that a statement is acceptable if it can be defended successfully against attacking arguments. As put by Bentahar et al. [6], *'the beliefs of a rational agent are characterised by the relations between its "internal arguments" supporting its beliefs and the "external arguments" supporting contrary beliefs.'*

We will use and adapt some definitions from Dung's work and other authors [8, 19, 20], as follows.

Definition 2.1. An *argumentation framework* is a directed graph $AF = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of arguments (vertices) and R is a set of directed, binary attack relationships between arguments (arcs), i.e., $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. We also say $attacks(a, b)$ holds iff $(a, b) \in \mathcal{R}$. Likewise, a set S of arguments attacks another set of arguments T (or T is attacked by S) if any argument in S attacks an argument in T.

Definition 2.2. 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.

Definition 2.3. A set of arguments S is said to be *conflict-free* if there is no attack within its arguments, i.e. there are no arguments $a, b \in S$ s.t. a attacks b. Likewise, a $S \subseteq \mathcal{A}$ of arguments is *admissible* iff it is conflict-free and each argument in S is acceptable with respect to S.

2.2 Dialogue Game Rules

The extension semantics introduced by Dung are powerful in asserting global properties of the argumentation framework, but their output is static and monological in nature.

In pursuance of a more dialogical approach [17], one must consider the dynamics of dialogue and the assumptions therewithin. Using Jakobovits and Vermeir's position framework formalism [15], *'the combination of a set of rules that govern the game, and the determination of winning criteria, constitute a dialectic semantics for the "theory" that underlies the player's arguments.'* We will adapt some of the definitions from [15], as follows.

Definition 2.4. Let $PF = (\mathcal{P}, \mathcal{R}^*)$ be a *position framework* paired with an argumentation framework $AF = (\mathcal{A}, \mathcal{R})$, where $\mathcal{P}$ consists of conflict-free subsets of $\mathcal{A}$, and $\mathcal{R}^*$ denotes the set of finite sequences of elements from R. Elements of $\mathcal{P}$ are called *positions*.

Definition 2.5. A *player* c can be categorised as the *proponent* (p) or *opponent* (o). The adversary of p is denoted $\overline{p} = o$. Conversely, $\overline{o} = p$.

Definition 2.6. Let a player $c \in \{p, o\}$ and a position $X \in \mathcal{P}$. A *move* in $\mathcal{P}$ is a pair (c, X). For a move $m = (c, X)$, we use $player(m)$ to denote c and $pos(m)$ to denote X.

Definition 2.7. A *dialogue type* is a tuple $(\mathcal{P}, \mathcal{R}^*, \phi)$, where $(\mathcal{P}, \mathcal{R}^*)$ is a position framework and $\phi : \mathcal{P}^* \rightarrow 2^{\mathcal{P}}$ is a *legal-move* function. A *dialogue* D in $(\mathcal{P}, \mathcal{R}^*, \phi)$ is any countable sequence $d_0, d_1, \ldots, d_n$ of moves in $\mathcal{P}$ that satisfies:

(1) $player(d_{i+1}) = \overline{player(d_i)}$, i.e. the players take turns.
(2) $pos(d_{i+1}) \in \phi(pos(d_0) \dots pos(d_i))$, i.e. the next move is legal.
(3) $d_{i+1} \notin \{d_0, d_1, \dots, d_i\}$, i.e. a move cannot be repeated twice.
(4) $attacks(pos(d_{i+1}), pos(d_i))$, it attacks the adversary's last move
(5) $player(d_0) = p$, i.e. the proponent makes the first move.

The dialogue D is said to be *about the position* $pos(d_0)$.

Definition 2.8. Let $X^{\leftarrow}$ and $X^{\rightarrow}$ denote the sets of positions that attack and are attacked by $X \in \mathcal{P}$, respectively. A player c is said to *win* the dialogue D if D is finite and ends with a move (c, X) s.t. $X^{\leftarrow} \cap \phi(D) = \emptyset$, i.e., the dialogue cannot be continued.

Definition 2.9. Let $(\mathcal{P}, \mathcal{R}^*)$ be a position framework. The legal-move function $\psi_{(\mathcal{P},\mathcal{R}^*)} : \mathcal{P}^* \rightarrow 2^{\mathcal{P}}$ which allows non-self-defeating nor useless moves in $(\mathcal{P}, \mathcal{R}^*)$ is defined as follows: $\forall Y_0, \dots, Y_i \in \mathcal{P}^*$:

$$\phi_{(\mathcal{P},\mathcal{R}^*)}(Y_0, \dots, Y_i) = \mathcal{P} \setminus (\{X \mid \overbrace{attacks(X, X)}^{\text{self-defeating}}\} \cup \overbrace{\bigcup_{j=o}^{i} Y_j^{\rightarrow}}^{\text{useless}})$$

We now have sufficient tools to formalise types of dialogues that encompass the previously chosen rules, with single or multiple arguments per move:

Definition 2.10. Let $AF = (\mathcal{A}, \mathcal{R})$ be an argumentation framework. A *useful-single-argument dialogue* in AF is a dialogue in the dialogue type $(\mathcal{A}', \mathcal{R}^*, \psi_{(\mathcal{A}',\mathcal{R}^*)})$, where $\mathcal{A}' = \{\{a\} \mid a \in \mathcal{A}\}$ and $\psi_{(\mathcal{A}',\mathcal{R}^*)}$ designates moves that are not self-defeating nor useless. $(\mathcal{P}, \mathcal{R}^*, \psi_{(\mathcal{P},\mathcal{R}^*)})$ is called the *useful-multiple-argument* dialogue type, where $\mathcal{P}$ is the set of conflict-free subsets of $\mathcal{A}$.

2.3 Explanations

Fan and Toni [11] propose an argumentation semantics aimed at generating explanations. This formalism promotes the notion of explanations as sets of arguments, taking into consideration which arguments contribute to the justification (or *r-defence*) of a specific premise (argument). We utilise their definitions for our framework, as follows.

Definition 2.11. Given an AA framework $AF = (\mathcal{A}, \mathcal{R})$, let $a, b \in \mathcal{A}$. *a r-defends b* iff:

(1) $a = b$; or
(2) $\exists z \in \mathcal{A}$, s.t. a attacks z and z attacks b; or
(3) $\exists z \in \mathcal{A}$, s.t. a r-defends z and z r-defends b.

$S \subseteq \mathcal{A}$ *r-defends* $a \in \mathcal{A}$ iff $\forall b \in S$: b r-defends a.

Definition 2.12. A set of arguments $S \subseteq \mathcal{A}$ is *related admissible* iff $\exists a \in S$ s.t. S r-defends a and S is admissible. a is said to be a *topic* of S. For any argument $a \in \mathcal{A}$, an *explanation* of a is $S \subseteq \mathcal{A}$ s.t. S is a related admissible set and a is a topic of S.

Their definition of explanations is further characterised by a classification with regards to cardinality and set inclusion:

Definition 2.13. Let $a \in \mathcal{A}$ and E_a be the set of all possible explanations of a. For every $S \in E_a$, we say S is a *minimal* or *maximal* explanation iff S is the smallest or largest subset of E_a with regards to cardinality, respectively. Similarly, S is a *compact* or a *verbose* explanation iff S is the smallest or largest subset of E_a with regards to set inclusion, respectively.

3 PROPOSED ARCHITECTURE

We introduce an architecture for *explainable conflict resolution* (X-CORE) as a mechanism that provides explainable deliberation capabilities for dialectic interactions between agents. Below, we elaborate on definitions and concepts that were created for the purpose of this application.

Example 3.1. Suppose the following situation: vehicle A crosses a green light in a junction and is about to collide with vehicle B, who ran a red light. In most highway codes, the rule *'a vehicle shall not cross the stop line on a red light'* can be ignored if rule *'a vehicle may cross the stop line on a red light if it is an emergency vehicle'* also applies to that situation. Therefore, in this specific situation, the right of way can be determined by B's status: if it is an emergency vehicle, then it could refer to the aforementioned rule and *argue* in favour of its right of way. Likewise, if B is not an emergency vehicle, then it would not be able to defeat A's claim of the first rule and B would find itself at fault.

Consequently, what would happen if either A or B is an artificial autonomous agent? Would the human counterpart benefit more from being explicitly told which rules are being used, or is having an implicit knowledge sufficient? We can rather evidently demonstrate that there is no challenge in having autonomous agents follow rule-based systems. Instead, our architecture aims to create a direct mapping between rulesets in human-readable form and corresponding argumentation frameworks.

For our problem domain, we assume a setting where agents perform localised decision-making. When acting, we are interested in ensuring that each agent's behaviour is compliant with an overall *culture* (represented by an argumentation framework AF) shared amongst all participants in the system. In order to check which rules apply in a specific event of a conflict, we introduce a mechanism of *argument verification* (Section 3.3). Finally, after agents and humans share a common model and can provide evidence for their rule-compliant justification, we demonstrate how to build explanations from this framework in Section 3.4.

3.1 Culture

Orderly behaviour can only happen if all agents share common guidelines and understand the same rules. We define the notion of *culture* as a collective agreement of norms and priorities, represented by an argumentation framework.

Definition 3.2. Let any two players p, o be the proponent and opponent in a dialogue game. We say a *proposition* is any argument $a \in \mathcal{A}$ that may be used by proponent p to request a contended resource from opponent o.

Definition 3.3. Let $\mathcal{K} \subseteq \mathcal{A}$ be the set of all propositions in $\mathcal{A}$. We say a system has a *culture* $C = (\mathcal{A}, \mathcal{R}, \mathcal{K})$ iff $|\mathcal{K}| > 0$ and all agents share C as their culture.

Example 3.4. A simple example would be: suppose a culture that contains three arguments, $\mathcal{A} = \{\mu, \alpha, \beta\}$, $\mathcal{K} = \{\mu\}$, where μ represents the proposition *'I have right of way'*, α represents *'I am an ambulance'* and β represents *'I am a fire rescue truck'*. Defining $\mathcal{R}_a = \{(\alpha, \mu), (\beta, \mu), (\alpha, \beta)\}$ is akin to defining that, in this application, ambulances have priority over fire trucks. Conversely, defining

$\mathcal{R}_b = \{(\alpha, \mu), (\beta, \mu), (\beta, \alpha)\}$ would mean the opposite. Despite having the same argument set $\mathcal{A}$, $C_a = (\mathcal{A}, \mathcal{R}_a, \mathcal{K})$ is a different culture from $C_b = (\mathcal{A}, \mathcal{R}_b, \mathcal{K})$.

3.2 Propositional Dialogues

When agents are presented with conflicts that require a compliant resolution (with regards to the ruleset), a dialogue game starts from the proponent p. Each agent then takes turns in choosing arguments that are potentially able to defeat the previous, as shown in Definition 2.7. Agents can use one or multiple arguments at each turn depending on whether it is a *useful-single* or *useful-multiple-argument* dialogue type. The game ends when one agent provides an argument that cannot be defeated by any of the other agent's possible arguments and thus has to concede or reject the initial proposition, depending on the result.

We extend the set of requisites for a dialogue seen in Definition 2.7 and propose the idea of a *propositional dialogue*:

Definition 3.5. Let $D = \{d_0, \ldots, d_i\}$ be a dialogue in a position framework $PF = (\mathcal{P}, \mathcal{R}^*, \phi)$ paired with a culture $C = (\mathcal{A}, \mathcal{R}, \mathcal{K})$. We say D is a *propositional dialogue* iff D is about a position $pos(d_0)$ where $\exists a \in pos(d_0)$ s.t. a is a proposition.

Definition 3.6. Let $D = \{d_0, \ldots, d_i\}$ be a propositional dialogue. We denote d_0 as the *motion* of the dialogue.

The player who wins D then takes priority or ownership with regards to the contended resource disputed in the proposition. However, as cultures may be constant, they need to cater to most circumstances in the environment. In Example 3.4, the ability of an agent-player c using α as an argument to defeat μ depends exclusively on the fact of c being, in fact, an ambulance. We introduce the concept of *argument verification* to deal with this matter.

3.3 Argument Verification

Most applications of argumentation frameworks consider frameworks as static, i.e., the combination of all the arguments may generate an *extension* or *labelling* that represents an insight about which arguments should or should not be admitted. In our case, the culture denotes a ruleset that does not account for a specific happenstance, but rather a more comprehensive model that encompasses different future scenarios that pertain to an agent's perspective, or notion of 'self.' For that purpose, every agent has to *verify* which arguments are valid at a particular moment. We propose the architecture of *argument verification* to address the issue of factual correctness and *validity* of a specific argument given the context of its proponent.

The verification of arguments can be modelled as decision problems:

Definition 3.7. Let $\alpha \in \mathcal{A}$ and $c \in \{p, o\}$ be an argument and a player, respectively. We denote α as *demonstrable* by agent-player c iff checking the correctness of that argument admits a finite and computable decision procedure.

We can naturally extend this definition to encompass the complexity of argument verification. For example, P-*demonstrable arguments* represent arguments whose associated decision problem is in P.

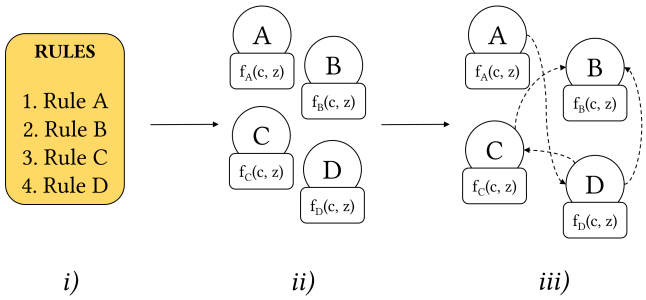

Figure 1: Example diagram of X-CORE: *i)* Ruleset; *ii)* Mapping rules into pairs of arguments and verifier functions; *iii)* Defining attacks between arguments.

This decision procedure can be represented by a predicate function that evaluates, in the current context, whether a specific argument may be used or not.

Definition 3.8. Let ζ denote the set of all possible contexts in the environment. $\forall \alpha \in \mathcal{A}$, α admits a predicate function $f_\alpha : c, z \rightarrow \{\texttt{True}, \texttt{False}\}$, where c represents the player and $z \in \zeta$ is a context. We say f_α is the *verifier* function of argument α. A special case applies for propositions, as they are hypothetical and their verifier functions always return `True`.

Definition 3.9. Let $\alpha \in \mathcal{A}$ be an argument. Let $c \in \{p, o\}$ be a player and $z \in \zeta$ a context. We say argument α is *demonstrably true* by player c iff $f_\alpha(c, z) = \texttt{True}$.

Definition 3.10. Let $D = \{d_0, \ldots, d_i\}$ be a dialogue. Let $c \in \{p, o\}$ be a player and $X \in \mathcal{P}$ be a position. We say $d_{i+1} = (c, X)$ is a *verified move* iff $\forall a \in pos(d_{i+1})$, $f_a(c) = \texttt{True}$.

Note that demonstrably true arguments do not mean that they are universally true – not even that they are true at all. All it means is that an agent will be able to compute a procedure to check if that statement stands against its own knowledge in the current context. The notion of demonstrably true is, in fact, a local definition of truth, as it only requires the perception of a single agent, even if the agent is mistaken/uneducated about the world (such as in systems with imperfect/incomplete information.)

All deliberation in this system is delegated into the specifics of the predicate functions that accompany each argument in the system. Hence, the design of a system in X-CORE is divided in two phases (see Figure 1):

- Mapping rules into pairs of arguments and verifier predicate functions;
- Establishing attack relationships between generated arguments.

As every rule is represented by an argument, tracing the history of exchanged arguments in this manner provides an insight on each agent's attempt to justify their prioritisation based on the ruleset provided. We facilitate the interaction with humans by providing *explanations* to justify the results of the dialogue game.

3.4 Explanation Generation

X-CORE does not bind agents into a specific strategy for choosing moves in a dialogue game. The justification is that humans cannot

be bound to a unique way of thinking, or be *programmable* as an artificial agent can. Therefore, we allow agents to freely choose their (verified) moves and focus on generating *post-hoc* explanations derived from the history of a dialogue D. For that purpose, we propose some mechanisms for generating explanations in X-CORE, based on the definitions seen in [11].

Definition 3.11. Let $D = \{d_0, \ldots, d_i\}$ be a completed dialogue and d_i be the *winner* move. We say a set $W \subseteq D$ is the set of *winning* moves where $W = \{d_k \in D \mid player(d_k) = player(d_i)\}$. The set of *losing* moves is denoted by $L = D \setminus W$.

Definition 3.12. Let d_i be the winner move in D. An *explanation* E_D of D is defined as $E_D \subseteq D$ s.t. $d_i \in E_D$, i.e., it always contains the winner move. We denote $\mathcal{E}_D$ as the set of all possible explanations of a dialogue D.

We can create a notion of *contrastive explanations* to include losing moves. The idea behind contrastive explanations is to provide extra justification as to why a specific argument was *not* accepted. We denote explanations without losing moves as *plain explanations*.

Definition 3.13. Let d_i be the winning move in D. A *contrastive* explanation CE_D of D is defined as $CE_D \in \mathcal{E}_D$ s.t. $\exists X = d_i, \exists Y \in L$, and $X, Y \in CE_D$. A *plain* explanation PE_D of D is defined as $PE_D \in \mathcal{E}_D$ s.t. $PE_D \subseteq W$.

Definition 3.14. Adapted from Definition 2.13. Let $\mathcal{E}_D$ be the set of all possible explanations of D. We therefore say that, for any $S \in \mathcal{E}_D$, S is a: *minimal* or *maximal* explanation iff S is a smallest or largest subset of $\mathcal{E}_D$ with regards to cardinality, respectively. S is a *compact* or a *verbose* explanation iff S is a smallest or largest subset of $\mathcal{E}_D$ with regards to set inclusion, respectively.

One could observe the entire footprint of uttered arguments and generate an explanation by writing all their natural language representations, but this approach is too verbose and unwieldy in most cases (especially if agents operate under *useful-single-argument* dialogue rules). We can attempt to specify a bound on the number of positions chosen to support an explanation.

Definition 3.15. $E' \in \mathcal{E}_D$ is an *n-reason* explanation iff $|E'| = n$.

We will now apply these definitions to do a proof of concept implementation using X-CORE for the purposes of our user study.

4 PROOF OF CONCEPT STUDY

In order to investigate the usefulness and efficiency of explanations in human-agent deconfliction settings, we designed a user study by instantiating a multi-agent resource contention environment. The problem of *multi-agent path deconfliction* lends itself naturally to our objectives: it is a sufficiently intuitive problem, requires minimal prior knowledge, and disputed resources are obvious (space).

Our hypotheses are:

- **H$_1$**: Explanations provide a higher improvement for human performance in more complex systems than in simpler systems.
- **H$_2$**: Explanations provide a higher decrease in time spent by a human in a task in more complex systems than in simpler systems.
- **H$_3$**: User perception of explainable systems is more positive in more complex systems than in simpler systems.

Next, we introduce our definition of a path deconfliction environment and our application: the Busy Barracks game.

4.1 Path Deconfliction Environment

We define the path deconfliction environment in the form of a 2D discrete time and discrete space grid, represented by a DAG.

Let $L = (V, E)$ be a finite directed acyclic graph (DAG) whose vertices are contained within the points in $\mathbb{Z}^3$, representing a bi-dimensional discrete space as x, y coordinates and time as t. Let $u = (x_1, y_1, t_1)$ and $v = (x_2, y_2, t_2)$ be any two points in this space.

We denote $(u, v) \in E \iff (d(u, v) \leq d_{\max}$ and $t_2 - t_1 = 1)$, where $d_{\max}$ is the maximum distance achievable by any agent on a single time step, expressed as the Manhattan distance $d_M(u, v) = |x_1 - x_2| + |y_1 - y_2|$ between two points in a $G_{x \times y}$ grid graph.

A set of K obstacles $\Upsilon = \{v_1, \ldots, v_K\}$ is given as input, where $\Upsilon \subset V$. The resulting traversable graph G is defined as $G = L - \Upsilon$.

A set of N agents $Q = \{q_1, \ldots, q_N\}$ is placed over $V(G)$. Each agent can traverse one edge per time step. This edge may traverse longer distances in (x, y) space, depending on the value of $d_{\max}$ given as input. A goal $g_i \in V(G)$ is defined for every agent $q_i \in Q$.

Plans to reach goal vertices are represented in the form of path subgraphs of G. Given an agent q_i and its corresponding goal g_i, the agent's plan is represented in the form $P(q_i) = \{v_0, \ldots, v_i\}$, where v_0 is q_i's current position and $v_i = g_i$. The length of plan $P(a_i)$ is equal to $|P(a_i)|$.

Definition 4.1. Two paths $P(a_i)$ and $P(a_j)$ are said to be *conflicting* if $P(a_i) \cap P(a_j) \neq \emptyset$ (they attempt to visit the same vertex at the same time step) or if $\exists u = (x, y, t), \exists v = (x', y', t+1)$ s.t. $u, v \in P(a_i)$ and $\exists u' = (x', y', t), \exists v' = (x, y, t+1)$ s.t. $u', v' \in P(a_j)$, e.g., agents swap positions.

4.2 The Busy Barracks Game

We present the previously-defined Path Deconfliction Environment to human participants as a computer game called Busy Barracks, or BB (see Figure 2). In it, the human controls a military official represented by an agent $q_h \in Q$. The human can choose one of two actions: move towards a direction (north, south, west, east), or choose to wait in place for a round. Agents move in lockstep, i.e., once the human makes a decision, all the agents make their planned move at the same time. The human is given 50 arbitrary units of fuel and told to navigate towards a goal destination under the following constraints:

- For every move or wait action, the human will lose 1 unit of fuel.
- If a collision occurs (see Definition 4.1), the human loses 5 units of fuel.
- Every 10 seconds past the first move in the game (in clock time), the human loses 1 unit of fuel.

The human is encouraged to reach their destination whilst maximising their remaining fuel. In practice, in order to achieve good scores, players will need to make short, collision-free trajectories with quick reaction times. The environment is composed of several other autonomous agents, who have individual goals and will also move towards their destinations concurrently with the player.

Humans are informed that agents will follow the rules and automatically reroute to clear the way if they understand that the

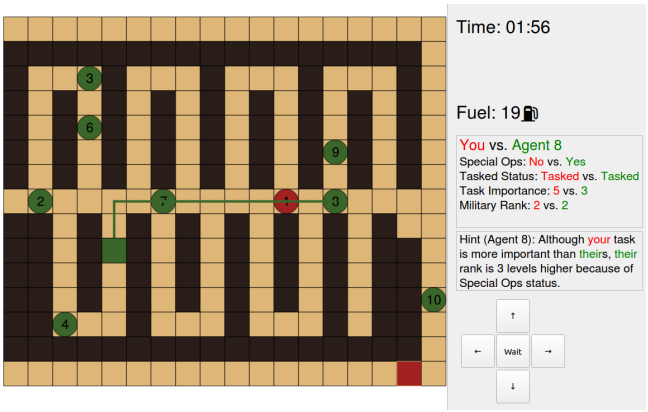

Figure 2: The Busy Barracks game UI. Red player is human, followed by agents in green. The path to destination of Agent 8 can be seen as a green line. The human must decide whether to give way or to move towards the red cell.

human has priority in a specific conflict setting. Agents will remain in their original trajectory and expect the human to clear the way if they understand that they have priority according to the rules. It is down to the human to make the decision to either remain in their original trajectory (assuming that the agent will clear the way) or make way (assuming that the agent will keep their trajectory and will potentially collide if evasive action is not taken).

In order to explore the traits of a system with explicit rulesets, the human is provided with a deconfliction ruleset, presented in textual form on a sheet of paper. This document introduces arbitrary and game-specific properties that each agent has, and how those properties play out in generating a prioritisation when a spatial conflict arises. In other terms, by observing the properties and the rules correctly, every agent should unequivocally understand if they have the right of way or if they should concede and grant passage to the opponent.

Example 4.2. Suppose the following ruleset:

(1) You should have right of way if:
  (a) Your rank is higher than the other agent's rank.
  (b) You are tasked and the other agent is not tasked, regardless of their rank.

This ruleset implies the existence of two properties: *rank* and *tasked status*; and two rules: (a) and (b), as seen above. Thus, if we have agents $q_1 : \{\text{rank}(q_1) : 2,\ \text{tasked}(q_1) : yes\}$ and $q_2 : \{\text{rank}(q_2) : 4,\ \text{tasked}(q_2) : no\}$, even though q_2 might be able to argue that it has a higher rank (rule (a)), it will be defeated when q_1 invokes rule (b).

Following this textual ruleset, we devise an example culture $C_{easy} = (\mathcal{A}, \mathcal{R}, \mathcal{K})$. We instantiate the set of arguments $\mathcal{A} = \{\mu, a, b\}$, where μ represents the proposition (1) *'you should have right of way'*. and a, b represent rules (a) and (b), respectively. Let c be a player and $\overline{c}$ their immediate opponent. The verifier functions are defined as follows:

$$f_a(c, \overline{c}) = \begin{cases} \texttt{True} & \text{if rank}(c) > \text{rank}(\overline{c}), \\ \texttt{False} & \text{otherwise.} \end{cases}$$

$$f_b(c, \overline{c}) = \begin{cases} \texttt{True} & \text{if tasked}(c) = yes \text{ and tasked}(\overline{c}) = no, \\ \texttt{False} & \text{otherwise.} \end{cases}$$

Since we know that rule (b) supersedes rule (a), we define $\mathcal{R} = \{(a, \mu), (b, \mu), (b, a)\}$ to complete the specification of C_{easy}.

Culture: For the BB game, we created three different rulesets, ranging in different levels of complexity. We posit that cultures become more complex as they grow in number of rules, hence our nomenclature. We refer back to the taxonomy seen in Rosenfeld and Richardson [24] (*not useful*, *beneficial*, and *critical*) to create three cultures with different sizes: *easy*, *medium*, and *hard*. Each culture was created from a textual ruleset that was handed over to human players.

- C_{easy}: 2 properties and 2 rules (described in Example 4.2.)
- C_{medium}: 4 properties and 4 rules.
- C_{hard}: 6 properties and 9 rules.

Dialogue Game: All players can publicly see the destination and intended trajectory of their opponents. When any two agents find themselves in conflict, they initiate a dialogue game and try to persuade the other to give way to them based on the culture that is being used in that instance of the game ($C_{easy}, C_{medium}, C_{hard}$.) In the BB game, all exchanges are *useful-single-argument* dialogues. Moves are chosen randomly among the subset of demonstrably true arguments. The argumentative exchange happens in the background and is not visible to the human.

The decision reached by this dialogue game decides the next action taken by the autonomous agent (to concede via rerouting or to continue in their original trajectory). The human must observe the rules and take action based on their belief of what the agent will do next. Agents always play optimally and do not make mistakes. A wrong decision from the human leads to two possible outcomes: either a collision or an unnecessary diversion from both human and agent, who both try to give way to each other (as the agent assumes the human will also play optimally.) There are 8 agents plus the human in every round, where exactly four of them will have right of way against the human, regardless of difficulty level.

Difficulty level does not affect the map layout, agent behaviour or any other factors that might influence scores or time other than the rules involved in deciding who gives way. If a human played with the same speed and the same success rate in every difficulty level, their scores would always be identical. Differences in score are uniquely determined by human performance.

4.3 User Study

We recruited 35 participants (21 male, 14 female, ages 20-39) within the university (students and staff). Participants were invited to play the BB game in a quiet room. Every new participant would be allocated to play one out of three versions of the game: either C_{easy}, C_{medium}, or C_{hard}. Participants did not know that any other versions of the game were available.

Our study is organised in a within-subject design in order to measure how each individual participant's performance is altered

Figure 3: Example of information available to human. Left: no explanation/hints, only data (N). Right: data + explanation/hint (X).

in the presence or absence of explanations. Each participant played two rounds of the game (within the same allocated culture): one version containing the only properties and rules, and another version containing properties, rules, and additional explanations generated in the form of *hints* in the game UI (see Figure 3). Those explanations were generated live based on the outcome of the background dialogue game between the human-controlled agent and the autonomous opponent. For brevity, we shall henceforth denote the *non-explainable* round as N and the *explainable* round as X. We alternated the starting order of the rounds (X or N) to minimise familiarity bias.

An experience questionnaire (extracted and modified from GEQ [14]) was given to each participant at the end of each round. We clustered questions in three main groups (GEQ indices in brackets): Competence (10, 15, 17, 21); Affect (9, 22, 24); and Challenge (23, 26, 33). We included four custom questions to evaluate game-specific criteria, such as how often they consulted the text rules and if they anticipated/agreed with agents' actions. Answers were collected in a 5-point Likert scale. We collected game performance data, such as: score (represented by fuel units remaining at the end of the game), number of collisions, and time taken until completion.

The non-explainable version allows the human to visualise their opponent's trajectory and their properties. Based on this available information (and the rules' knowledge present in the ruleset), the human must then evaluate which rules apply and decide a course of action. In the explainable case, we decide to provide a succinct, or even partial explanation in the form of a hint.

Explanation Generation: For that reason, in every dialogue game present in the game, X-CORE generates hints by selecting a 2-*reason contrastive explanation* CE' (a minimal and compact contrastive explanation) and presenting in a textual form (Figure 3). Our objective is not to compare two versions with different information available, but instead to evaluate the impact of having all the information required to make a decision (N) versus having all the information *plus* an explanation (X) to assist the human.

We are interested in the differences in human performance between playing N and X, namely, how much human performance improves or worsens between N and X in each difficulty level.

5 RESULTS

Due to the limited number of samples, we choose to not make assumptions of parametrisation in the data. Every sample is grouped into a difficulty level (E, M, and H representing C_{easy}, C_{medium}, and C_{hard}, respectively). Users play two rounds (N and X, in alternated order). Thus, a player allocated to M would play both MN and MX rounds, respectively. Sample sizes are Easy (n = 11), Medium (n = 12), and Hard (n = 12).

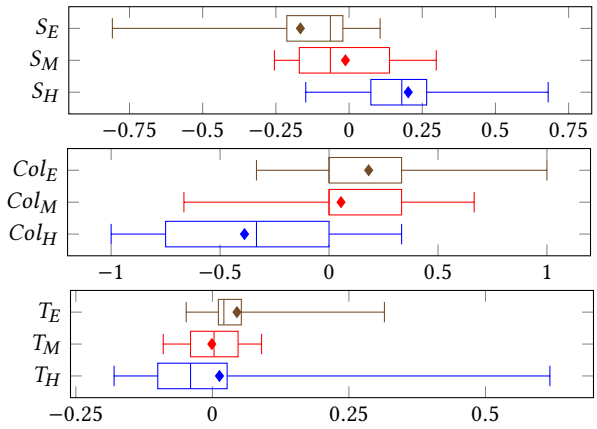

Figure 4: Scores (S), Collisions (Col), and Time (T). Results are shown in the form of box plots (25th, 50th, 75th percentile, and whiskers covering all data and outliers).

We define our measures as:

- Score (S): normalised score $(S_X - S_N)/(S_{\max} - S_{\min})$. Positive values of S mean score improvement in X.
- Collisions (Col): normalised number of collisions $(Col_X - Col_N)/(Col_{\max} - Col_{\min})$. Negative values mean reduced number of collisions in X.
- Time (T): normalised time elapsed $(T_X - T_N)/(T_{\max} - T_{\min})$. Negative values of T mean reduction in time elapsed in X.

5.1 Score

Given 3 sample sets: S_E (E scores), S_M (M scores), and S_H (H scores) (see Figure 4), we run a Kruskal-Wallis H-Test (KW) under the alternative hypothesis that at least one of the distributions come from a different population and confirm significant differences (H = 11.63, p = 0.003**)[1]. We then perform a pairwise one-sided Mann-Whitney U-Test (MW) under the alternative hypothesis that easier categories have significantly smaller S than harder ones, meaning that the improvement in X is less pronounced in easier rounds.

Results in Table 1 show that score improvement in M is significantly smaller than H, whilst score improvement in E is *very* significantly smaller than H, but not significantly smaller than M.

5.2 Collisions

Like the previous sets, we consider Col to isolate the number of wrong decisions that specifically led to collisions (see Definition 4.1), and how did that differ within subjects between N and X. We run a KW under the alternative hypothesis that at least one of the distributions come from a different population, confirming the hypothesis (H = 9.83, p = 0.007**).

[1](*) p < 0.05; (**) p < 0.01; (***) p < 0.001.

	S_E	S_M	S_H
S_E	-	U = 48.5 p = 0.1472	U = 11.5 p = 0.0004***
S_M	-	-	U = 34.0 p = 0.015*

	Col_E	Col_M	Col_H
Col_E	-	U = 72.5 p = 0.346	U = 111.5 p = 0.002**
Col_M	-	-	U = 112.5 p = 0.008**

	Cha_E	Cha_M	Cha_H
Cha_E	-	U = 92.0 p = 0.056	U = 116.5 p = 0.001***
Cha_M	-	-	U = 101.0 p = 0.047*

Table 1: Pairwise MW for *S*, *Col*, and *Cha*.

Since the distributions are different, we perform a one-sided MW, this time with the alternative hypothesis that easier categories have significantly *higher* number of collisions, i.e., they do not improve (and reduce) their number of collisions as well as harder levels. The results in Table 1 show that collision improvement in M is significantly smaller than H, whilst collision improvement in E is also significantly smaller than H, although not significantly smaller than M. Both S and Col results support $\mathbf{H_1}$.

5.3 Times

Similarly to Col, T represents the change in time elapsed to complete each round from N to X. We run a KW in order to isolate the distributions but did not find significant differences (H = 3.61, p = 0.16). However, a pairwise one-sided MW reveals a significant improvement in T between E and H (U = 94.0, p = 0.04*), showing that participants in H have a superior reduction in time in X compared to those in E. This result supports $\mathbf{H_2}$.

5.4 User Experience

In order to evaluate the effect of ordering (whether users who played their first round as N or X had a significantly different perception of the game), we ran KWs for each cluster of questions, separating populations by their starting mode (N or X). The alternative hypothesis for all cases was that there was a significant difference in populations, which was not confirmed for any: Challenge (H = 0.24, p = 0.61); Competence (H = 0.48, p = 0.48); Affect (H = 0.13, p = 0.71); and Game-Specific (H = 1.73, p = 0.18).

We then evaluate the populations based on the difficulty level. To evaluate the differences in populations, we ran KWs under the alternative hypothesis that the populations differ significantly depending on difficulty level. We manage to validate this hypothesis for Challenge (Cha) (H = 10.28, p = 0.005**), but not for Affect (At) (H = 3.91, p = 0.14), Competence (Com) (H = 4.98, p = 0.08) and Game-Specific (Gam) (H = 5.52, p = 0.06). We perform pairwise one-sided MWs under the alternative hypothesis that easier categories have a significantly smaller improvement in the perception of challenge from N to X.

The results in Table 1 show that the improvement of perception of Cha in M is significantly smaller than H. The improvement in E is *very* significantly smaller than H, although not significantly smaller than M. Despite populations being not isolated in the previous KW for Gam, similar pairwise MW results are found: Gam_E vs. Gam_H (U = 56.5, p = 0.01*) and Gam_M vs. Gam_H (U = 42.5, p = 0.04*).

Additionally, user experience results in At and Com clusters also revealed significant improvements (At: U = 96.0, p = 0.03*; Com: U = 32.0, p = 0.02*) from N to X between between E and H levels. These results support $\mathbf{H_3}$.

6 DISCUSSION AND CONCLUSION

We achieved significant results in demonstrating how the benefit of explanations in human-agent deconfliction correlates to the complexity of the underlying system. Our results demonstrated clear differences between within-subject improvement when comparing their performance in N against the performance in X, which demonstrates that humans benefit from explanations – but mostly when the system is sufficiently complex to warrant such explanations.

In fact, when the complexity is small, humans might actually perform better *without* any explanations. We probe this claim by running a one-sided MW considering the alternative hypothesis that global (between-subjects) EN scores were higher than EX scores (U = 88.0, p = 0.03*), which was significant. Contrariwise, a similar test under the alternative hypothesis that global HN scores are *lower* than HX scores (U = 36.0, p = 0.019*) also proved significant. In many cases, M populations were harder to distinguish between E and H in nondirectional tests, such as in T, At, and Com analyses. Still, hypotheses $\mathbf{H_1}$, $\mathbf{H_2}$, and $\mathbf{H_3}$ are validated for all E and H within-subject results. We believe that a larger scale study and further refinement of M in terms of complexity might consolidate all populations more clearly.

Post-experiment interviews were conducted to discuss the user experience. Participants were asked to self-report on how they felt about the hints. Six out of 11 participants who played the E version reported finding the hints *not useful*. At the M level (n = 12), 4 participants found the hints *not useful*, and 5 expressed using hints as a useful confirmation mechanism to check their mental computation. Last, at H (n = 12), 9 players reported that hints were *very useful* and *primarily relied* on the hints to act. These findings map well to the taxonomy of [24] (*not useful*, *beneficial*, and *critical*) and suggest that the taxonomy of the *need* for explanations can be considered under a new dimension: that of system complexity.

In this game, the deconfliction behaviour of the agents was entirely dictated by the present culture. We believe that X-CORE can find applications beyond human-agent deconfliction, and towards multi-agent systems in general. For example, a decentralised multi-agent system could be designed in terms of a culture, and performing individual implementations for each rule could prove easier than writing a monolithic policy that tries to emulate a complex ruleset (especially if coming from text/human regulations), with the innate benefit of being explainable, as demonstrated by our architecture and study. The deconfliction behaviour of agents can be changed by adding or removing arguments individually, or changing their attack relationships. Future studies will demonstrate how X-CORE can be used for modelling real-life rulesets in deployed multi-agent/multi-robot applications.

ACKNOWLEDGMENTS

The authors would like to thank Dr Jon Roozenbeek, Nikhil Churamani, and Guilherme Paulino-Passos for providing valuable assistance during this study. Alex Raymond is supported by L3Harris ASV and the Royal Commission for the Exhibition of 1851. Hatice Gunes is supported by the EPSRC Project ARoEQ (Grant Ref: EP/R030782/1). Amanda Prorok is supported by the EPSRC (Grant Ref: EP/S015493/1). Their support is gratefully acknowledged.

REFERENCES

[1] Leila Amgoud and Henri Prade. 2009. Using arguments for making and explaining decisions. *Artificial Intelligence* 173, 3-4 (3 2009), 413–436.

[2] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1078–1088.

[3] Peter M Asaro. 2016. The liability problem for autonomous artificial agents. In *2016 AAAI Spring Symposium Series*.

[4] Trevor Bench-Capon and Sanjay Modgil. 2019. Norms and Extended Argumentation Frameworks. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19)*. Association for Computing Machinery, New York, NY, USA, 174–178.

[5] Michael R Benjamin, Joseph A Curcio, John J Leonard, and Paul M Newman. 2006. Navigation of unmanned marine vehicles in accordance with the rules of the road. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* 3581–3587.

[6] Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review* 33, 3 (3 2010), 211–259.

[7] Holger Billhardt, Vicente Julián, Juan Manuel Corchado, and Alberto Fernández. 2014. An architecture proposal for human-agent societies. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*. 344–357.

[8] Sylvie Coste-Marquis, Caroline Devred, and Pierre Marquis. 2005. Symmetric Argumentation Frameworks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer, Berlin, Heidelberg, 317–328.

[9] Stephen Cranefield, Nir Oren, and Wamberto W Vasconcelos. 2019. Accountability for Practical Reasoning Agents. In *Agreement Technologies*, Marin Lujak (Ed.). Springer International Publishing, Cham, 33–48.

[10] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77, 2 (9 1995), 321–357.

[11] Xiuyi Fan and Francesca Toni. 2015. On Computing Explanations in Argumentation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 1492–1496.

[12] J J Finkelstein. 1968. The Laws of Ur-Nammu. *Journal of Cuneiform Studies* 22, 3-4 (1968), 66–82.

[13] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[14] W A IJsselsteijn, Y A W De Kort, and Karolien Poels. 2013. The game experience questionnaire. *Eindhoven: Technische Universiteit Eindhoven* (2013).

[15] H. Jakobovits and D. Vermeir. 1999. Dialectic semantics for argumentation frameworks. In *Proceedings of the seventh international conference on Artificial intelligence and law - ICAIL '99*. ACM Press, New York, New York, USA, 53–62.

[16] Antonis Kakas and Pavlos Moraitis. 2003. Argumentation Based Decision Making for Autonomous Agents. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '03)*. ACM, New York, NY, USA, 883–890.

[17] Amin Karamlou, Kristijonas Čyras, and Francesca Toni. 2019. Deciding the Winner of a Debate Using Bipolar Argumentation. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2366–2368.

[18] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable Agency for Intelligent Autonomous Systems. *Twenty-Ninth IAAI Conference* (2 2017). https://www.aaai.org/ocs/index.php/IAAI/IAAI17/paper/viewPaper/15046

[19] Sanjay Modgil. 2014. Revisiting Abstract Argumentation Frameworks. In *Theory and Applications of Formal Argumentation*. Springer, Berlin, Heidelberg, 1–15.

[20] Sanjay Modgil and Martin Caminada. 2009. Proof Theories and Algorithms for Abstract Argumentation Frameworks. In *Argumentation in Artificial Intelligence*. Springer US, Boston, MA, 105–129.

[21] Ugo Pagallo. 2016. Even Angels Need the Rules: AI, Roboethics, and the Law. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence (ECAI'16)*. IOS Press, Amsterdam, The Netherlands, 209–215.

[22] Albert Rizaldi and Matthias Althoff. 2015. Formalising traffic rules for accountability of autonomous vehicles. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. 1658–1665.

[23] Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*. 320–328.

[24] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (11 2019), 673–705.

[25] Elizabeth I. Sklar and Mohammad Q. Azhar. 2018. Explanation through Argumentation. In *Proceedings of the 6th International Conference on Human-Agent Interaction - HAI '18*. ACM Press, New York, New York, USA, 277–285.

[26] Anthony P Young, Nadin Kökciyan, Isabel Sassoon, Sanjay Modgil, and Simon Parsons. 2018. Instantiating Metalevel Argumentation Frameworks. In *COMMA*. 97–108.

[27] Zhiwei Zeng, Chunyan Miao, Cyril Leung, and Jing Jih Chin. 2018. Building More Explainable Artificial Intelligence With Argumentation. *Thirty-Second AAAI Conference on Artificial Intelligence* (4 2018).