

Journal Pre-proof

The Quest of Parsimonious XAI: a Human-Agent Architecture for Explanation Formulation

Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi et al.

PII: S0004-3702(21)00124-7

DOI: <https://doi.org/10.1016/j.artint.2021.103573>

Reference: ARTINT 103573

To appear in: *Artificial Intelligence*

Received date: 2 May 2020

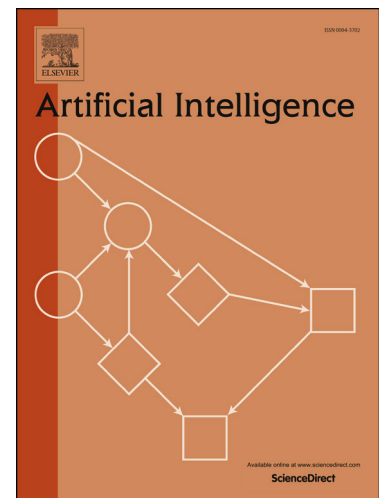
Revised date: 29 July 2021

Accepted date: 2 August 2021

Please cite this article as: Y. Mualla, I. Tchappi, T. Kampik et al., The Quest of Parsimonious XAI: a Human-Agent Architecture for Explanation Formulation, *Artificial Intelligence*, 103573, doi: <https://doi.org/10.1016/j.artint.2021.103573>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier.



The Quest of Parsimonious XAI: a Human-Agent Architecture for Explanation Formulation

Yazan Mualla*

CIAD UMR 7533, Univ. Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France

Igor Tchappi

*Orange Labs, 6 Avenue Albert Durand, 31700 Blagnac, France
CIAD UMR 7533, Univ. Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France
Faculty of Sciences, University of Ngaoundere, B.P. 454 Ngaoundere, Cameroon*

Timotheus Kampik

Department of Computing Science, Umeå University, 90187 Umeå, Sweden

Amro Najjar

*AI-Robolab/ICR, Computer Science and Communications, University of Luxembourg, 4365
Esch-sur-Alzette, Luxembourg*

Davide Calvaresi

University of Applied Sciences and Arts of Western Switzerland, Sierre, Switzerland

Abdeljalil Abbas-Turki

CIAD UMR 7533, Univ. Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France

Stéphane Galland

CIAD UMR 7533, Univ. Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France

Christophe Nicolle

CIAD UMR 7533, Univ. Bourgogne Franche-Comté, UB, F-21000 Dijon, France

Abstract

With the widespread use of AI, understanding the behavior of intelligent agents and robots is crucial to guarantee successful human-agent collaboration since it is not straightforward for humans to understand an agent's state of mind. Recent empiri-

*Corresponding author

Email address: yazan.mualla@utbm.fr (Yazan Mualla)

cal studies have confirmed that explaining a system's behavior to human users fosters the latter's acceptance of the system. However, providing overwhelming or unnecessary information may also confuse the users and cause failure. For these reasons, *parsimony* has been outlined as one of the key features allowing successful human-agent interaction with parsimonious explanation defined as the simplest explanation (*i.e.* least complex) that describes the situation adequately (*i.e.* descriptive adequacy). While parsimony is receiving growing attention in the literature, most of the works are carried out on the conceptual front. This paper proposes a mechanism for parsimonious eXplainable AI (XAI). In particular, it introduces the process of *explanation formulation* and proposes HAExA, a human-agent architecture allowing to make it operational for remote robots. To provide parsimonious explanations, HAExA relies on both contrastive explanations and explanation filtering. To evaluate the proposed architecture, several research hypotheses are investigated in an empirical human-user study that relies on well-established XAI metrics to estimate how trustworthy and satisfactory the explanations provided by HAExA are. The results are analyzed using parametric and non-parametric statistical testing.

Keywords: Explainable Artificial Intelligence, Human-Computer Interaction, Multi-Agent Systems, Empirical Human Studies, Statistical Testing.

1. Introduction

Explaining the reasoning process and the outcomes of complex computer programs has received considerable attention since the 1990s, when research works on explainable expert systems were disseminated [1]. Nowadays, with the pervasive application
5 of machine learning, the need to explain the reasoning of Artificial Intelligence (AI) methods and systems is considered a top priority [2]. In 2017, the European Parliament recommended AI systems to follow the principle of transparency; systems should be able to justify their decisions in a way that is understandable to humans [3]. In April
2019, the European Union's *High-Level Expert Group* on AI presented a document en-
10 titled "Ethics Guidelines for Trustworthy AI" [4]. This report highlighted transparency as a key property of trustworthy AI.

In the same vein, recent works in the literature highlighted explainability as one of the cornerstones for building trustworthy, responsible, and acceptable AI systems [5, 6, 7, 8]. Consequently, the sub-domain research of eXplainable Artificial Intelligence (XAI) gained momentum both in academia and industry [9, 10, 11]. Primarily, this surge is explained by the often useful, yet sometimes intriguing [12], results of black-box machine learning algorithms and the consequent need to understand how these data fed into the algorithm produced the given results [2, 13, 14].

Another line of XAI research aims at explaining the outcomes of goal-driven systems (*e.g.* robots) [10] since, in the absence of a proper explanation, the human user will come up with an explanation that might be flawed or erroneous. In turn, this will degrade the user's acceptance of the system. This problem will be aggravated in the near future because these systems are expected to be omnipresent in our daily lives (*e.g.* autonomous cars on the roads, Unmanned Aerial Vehicles (UAVs) in a smart city, social assistant robots, *etc.*). In this context of heterogeneous human-agent systems, recent works in goal-driven XAI aim at ensuring mutual understandability, improving acceptability, and enhancing human-agent collaboration capabilities, in particular, to facilitate human safety in human-robot collaborations.

To achieve smooth human-agent interaction and deliver the best possible explanation to the human, two key features have been outlined in the literature when providing an explanation: (i) *Simplicity*: providing a relatively simple explanation that considers the *human cognitive load*. The latter is a limit beyond which humans are unable to process the provided information [15]; (ii) *Adequacy*: including all pertinent information in an explanation to help the human understand the situation.

Generating *simple* and *adequate* explanations is a challenging task because of the contradicting nature of the two features. On the one hand, achieving adequacy is challenging in complex situations involving multiple remote robots since it places more pressure on the human's cognitive load and requires scalable XAI mechanisms able to cope with the limited human cognitive capabilities. On the other hand, adequacy turns out to be a challenge in abnormal situations, where the remote robot tends to diverge from the behavior expected by their human users, and therefore, a situation-specific explanation is required.

This paper proposes *explanation formulation*, a process that tackles these challenges by striking a balance between simplicity and adequacy, thereby producing *parsimonious explanations*. This is done by relying on HAExA, a human-agent architecture allowing to make the explanation formulation operational. HAExA is deployed in a scenario involving remote robots represented by intelligent agents.

Our choice of an agent-based architecture is explained as follows. Agents are autonomous goal-driven software entities that are bound to individual perspectives. Thus, agents are capable of both representing the remote robot's perspective and piloting its interaction with its own environment. Agents are also able to represent end-users, to apply user preferences regarding the interaction with the system, and to assess the explanations that the user needs. Thus, HAExA is a multi-agent system architecture involving agents that represent robots as well as agents dealing with interaction with the user. Through interaction and coordination of these agents, HAExA formulates parsimonious explanations.

More specifically, the contribution of this work is threefold:

- C1)** Propose an agent-based architecture that facilitates human-agent parsimonious explainability. In this architecture, remote robots are represented as agents. The architecture helps in *formulating* the necessary parsimonious explanations communicated from remote agents to human users, while at the same time considering the human cognitive load to avoid overwhelming users with too many details in the explanation.
- C2)** Investigate the explanation formulation using various combinations of explanation generation and communication approaches. To generate explanations, we rely on the Belief-Desire-Intention (BDI) agent architecture [16].
- C3)** Conduct an empirical human user case study based on a scenario of package delivery using civilian UAVs. The study investigates the impact of the different techniques of explanation formulation (static filter, adaptive filter, and adaptive filter with contrastive explanations) on users. The significance of the participants' responses is statistically analyzed and presented using non-parametric and parametric testing.

The rest of this paper is structured as follows: Section 2 lays down the background concepts of this paper, while Section 3 highlights the most related works in the literature. Section 4 proposes HAExA, the human-agent explainability architecture. Section 5 presents the experimental case study and the Likert-based questionnaire built to collect the responses of the human participants. Section 6 presents the study results, *i.e.* the statistical analysis of the responses of the participants and a discussion thereof. Finally, Section 7 briefly discusses the implications and limitations of the research results before Section 8 concludes the paper.

2. Background and Definitions

This section provides the background of this work. Section 2.1 introduces the concept of parsimony and discusses its relation with XAI, while Section 2.2 introduces contrastive explanations as a component of parsimonious explanations and highlights related works to this concept in the literature. Section 2.3 outlines the process of providing an explanation. Since this papers address explainability in the context of autonomous goal-driven remote robots, Section 2.4 offers a brief introduction to existing works addressing goal-driven explainable agents and robots, while Section 2.5 discusses the cognitive architectures allowing to implement such agents. Section 2.6 provides insights on how to make empirical assessments in XAI.

2.1. Parsimony and XAI

The concept of parsimony of explanations has received considerable attention for centuries. A famous formulation of this concept is the “Occam’s Razor” [17, 18] stipulating that: “*Entities should not be multiplied beyond necessity.*” Thereafter, Occam’s Razor¹ became the basis of the principle of “parsimony of explanations.” This principle has been influential in scientific thinking in general and in problems of statistical inference in particular [19, 20, 21].

The goal of this principle is to choose the simplest (*i.e.* least complex) explanation that describes the situation adequately (*i.e.* descriptive adequacy). Yet, as has been

¹William of Occam, 1290–1349.

100 shown in the literature [22, 23, 24], parsimony is a largely subjective quality. For this reason, human studies have been outlined as key to assess how parsimonious an explanation is to a given user in a given situation. In these tests, the opinions of humans on the usefulness of explanations are collected and analyzed. With the advent of XAI, research on parsimony of explanations has gained new momentum since the explanations
105 provided by the AI systems to their human users should be simple while containing all the information about the system’s decision. Thus, parsimony has been identified as a key desideratum for XAI [25, 26]. Yet, very few works in the literature define what parsimony means in the context of XAI, show how parsimonious explanations can be generated and communicated to humans, or discuss their impact on the human receiving them can be assessed (please refer to Section 3 for an overview of these works).
110 In this work, we define parsimony as a balance achieved between simplicity and adequacy, where the former is providing simple explanations that consider the human cognitive load, and the latter mandates the inclusion of all pertinent information in the explanation to help the user understand the situation.

115 The discussion of the parsimony of explanations opens the door to these questions:
i) What is the information necessary to be kept in an explanation? ii) How is a parsimonious explanation formulated? To tackle these questions, some authors investigated contrastive explanation as a potential way to generate the explanation based on the necessary information that the human needs, instead of providing a full explanation of the system (*cf.* [27]). The next section offers an overview of contrastive explanations.
120

2.2. Contrastive Explanations

One way to develop parsimonious XAI is to rely on theories and experiments describing how humans explain their decisions and behavior. This emerging body of research mainly looks for insights from the social sciences [28]. The aim is to explore
125 how humans generate and communicate explanations in their everyday life. *Everyday explanations* are explanations of why particular events, behaviors, decisions, *etc.* happened [29]. Evidence in the literature suggests that in abnormal situations, these everyday explanations should take the form of contrastive explanations [28]. The latter have been defined in the literature as follows: “*the key insight is to recognize that one*

130 *does not explain events per se, but that one explains why the puzzling event occurred*
135 *in the target cases but not in some counterfactual contrast case” [30].*

The use of contrastive explanations is justified by the fact that people generally do not expect an explanation that consists of the complete cause of an event. Instead, humans prefer selecting one or two causes from a sometimes infinite number of causes
135 to be the explanation. However, this selection is influenced by certain cognitive biases [28]. Lipton [31] proposed one of the first works investigating the use of contrastive explanations in AI. His research concluded that if the explanations are to be designed for humans, they should be contrastive [31]. Later research showed that people do not explain the causes for an event by itself, but they explain the cause of an event
140 relative to another *counterfactual* event (that did not occur). Therefore, according to Kim et al. [32], a contrastive explanation describes “*why event A occurred as opposed to some alternative event B.*”. A likely reason for the prevalence and effectiveness of contrastive explanations is that humans typically explain events that they, or others, consider abnormal or unexpected [33, 34]. This contrastive explanation takes the form
145 of ‘*why*’ questions and it may be expressed in various ways [35, 36].

In recent years, research on contrastive explanations in AI received growing attention [37, 38, 39, 40]. Lim and Dey [41] found that “*Why not ...?*” questions were common questions that people asked after some human studies on context-aware applications. Winikoff [42] investigated how to answer contrastive questions, *e.g.* “*Why*
150 *didn’t you do ...?*” for Belief-Desire-Intention (BDI) agents. Another similar work has checked the same type of questions like “*Why didn’t you do something else*” [43]. However, most of the existing works consider contrastive questions, but not contrastive explanations, as mainly people use the difference between the occurred event and the expected event when they look for an explanation [28].

155 Evidence from social sciences confirms the importance of contrastive explanations both in human-to-human explanations and machine-to-human explanations. In an influential recent survey, Miller [28] identified useful insights related to XAI from the social sciences. Among the key findings outlined in his work, he postulated that explanations are contrastive in the sense that they are responses to particular counterfactual
160 cases.

2.3. Phases of an Explanation

Neerincx et al. [44] emphasize that for the explanations to serve their purposes, they should be aware of the context and the human information processing capabilities, *i.e.* human cognitive load. According to these authors, the process of providing explanations by agents to the human is defined by three distinct phases:

Generation. This phase considers what to explain and how to explain it. For example, explaining the perceptual foundation of the agent behavior, or explaining why a certain action is applied.

Communication. This phase is about the form of the explanation (textual, visual, in a simulation, *etc.*), and the means to communicate the explanation.

Reception. This phase is concerned with the human processing and understanding of the explanation. Concerning XAI reception, some user studies (*e.g.* [45]) have been conducted, but there is a lack of empirical research involving human users in realistic human-agent settings and scenarios where explanations are needed to understand the system's behavior [28, 46].

2.4. Goal-driven XAI

The majority of works in the literature of XAI are data-driven, *i.e.* they aim to interpret how the available data led a machine learning algorithm such as a Deep Neural Network (DNN) to take a given decision (*e.g.* a classification decision) [9]. More recently, XAI approaches have been extended to explain the complex behavior of goal-driven systems such as robots and agents [10, 47]. The main motivations for this research direction are: (i) In general, robot-human communication is a key challenge, since, by default, it is not straightforward for humans to understand the robot's State-of-Mind (SoM). The latter refers to the intentions and goals of a robot [47]. As has been shown in the literature, humans tend to assume that robots/agents have their own SoM [47], and that with the absence of a proper explanation, a human will come up with an explanation that might be flawed or erroneous; (ii) In the near future, goal-driven systems are expected to be omnipresent in our daily lives (*e.g.* social assisting robots

and virtual assistants). Therefore, ensuring mutual understandability among humans
 190 and robots/agents is key to improve their acceptability and human-agent collaboration
 capabilities, and in particular to facilitate human safety in human-robot collaborations.
 In the context of human-agent collaboration, XAI is of particular interest since provid-
 ing explanations in multi-agent environments is even more challenging than providing
 explanations in other settings [48].

195 2.5. Cognitive Agent Architectures

Agent architectures are frequently applied to equip robots/agents with greater au-
 tonomy. By designing proactive agents that control robots, the latter become capable
 of autonomously managing their actions and behavior to reach their goals [49, 50, 51].
 Any proposed architecture should have the following characteristics:

- 200 1) A representation of the environment where the agents act and interact;
- 2) A self-representation of the agent’s internal reasoning cycle;
- 3) Social skills for interacting with other agents.

These characteristics can be found —to different extents— in several well-known
 cognitive architectures such as BDI [52], FORR [53], ACT-R [54], LIDA [55], or Soar
 205 [56]. All these architectures reflect the first and second previously mentioned charac-
 teristics. Soar, BDI, ACT-R, and CLARION allow additionally to create social agents.
 ACT-R and CLARION [57] architectures are time-consuming to compute; hence they
 are not scalable in the context of near-real-time applications.

The BDI model is a model of human behavior that was developed by philosophers.
 210 It appeared first in the Rational Agency project at the Stanford Research Institute in the
 mid-1980s. The origins of this model lie in the theory of human practical reasoning
 developed by the philosopher Michael Bratman [52]. The conceptual framework of the
 BDI model is described in [58]. We shortly describe the different concepts of beliefs,
 desires, and intentions of the BDI model as follows [59]:

- 215 • **Beliefs:** Information that the agent has about the environment and may be out of
 date or inaccurate, and hence is considered a set of *beliefs* (and not *knowledge*)
 that is revised with time.

- **Desires:** All the possible states of affairs (or options) that the agent may want to achieve. However, having a desire does not imply that the agent acts upon it. It is a potential influencer of the actions of the agent. 220
- **Intentions:** The states of affairs that the agent has decided to achieve. Intentions may be goals that are delegated to the agent or may result from considering options. The agent usually looks at its options and selects between them its intentions. This process of selection may occur repeatably in a lower level of abstraction until reaching intentions that can be executed as atomic actions via the actuators of the agent. It is normal for an agent to have desires that are mutually incompatible with one another, but not mutually incompatible intentions. 225

The BDI model allows agents to exhibit more complex behavior than purely reactive models but without the computational overhead of other cognitive architectures [60]. Moreover, some evidence exists that BDI agent architectures facilitate knowledge elicitation from domain experts [61]. Furthermore, because BDI is based on the concepts of folk-psychology, it has been outlined as a good candidate to represent everyday explanations [62, 63], since it is considered as the attribution of human behavior using “everyday” terms such as beliefs, desires, intentions, emotions, and personality traits [64, 65]. Finally, BDI has been identified as the most used architecture to generate explanations for goal-driven agents (*e.g.* [63, 44]) [10]. 230

For all the previously mentioned reasons, this paper considers BDI architecture as a good option for providing explanations since it relies on folk-psychology to represent everyday explanations. Therefore, we opt to use BDI in the proposed architecture in this work (*cf.* Section 4). 240

2.6. Empirical XAI Assessment

To assess the user understandability and acceptability of the provided explanations and their usefulness, works in the literature use empirical experiments to assess a given XAI mechanism. To evaluate HAExA, we conduct a series of empirical experiments where users interact with explainable remote robots in an *Agent-Based Simulation* (ABS) and rely on *XAI questionnaires*. This section provides the background for this 245

evaluation process. In particular, Section 2.6.1 explains the concept of Agent-based Simulation (ABS) used to implement the proposed architecture, and Section 2.6.2 discusses various ways to build XAI-based questionnaires to be used in empirical human studies.

2.6.1. Agent-Based Simulations

ABS is a set of interacting intelligent entities that reflect, within an artificial environment, the relationships in the real world [66]. Consequently, ABS can be considered as a natural step towards better understanding and managing the complexity of today's business and social systems. Additionally, cognitive architectures are frequently applied in ABS [60].

For these reasons, ABS of BDI agents is a good candidate to simulate the behavior of complex systems, thereby offering a platform to build explainable agents and assess their understandability from the human user perspectives. Thus, this work proposes an explainable BDI agent architecture built within an ABS to gain insights into how to explain the system behavior that emerges from local interacting BDI agents and processes. Additionally, we argue that ABS facilitates a good reception of the explanations by human users.

2.6.2. XAI Questionnaires

There are several methods for evaluating the explanations, whether humans are satisfied by them, how well humans understand the AI systems, how curiosity motivates the search for explanations, whether the human's trust and reliance on the AI are appropriate, and finally, how the human-XAI system performs [67]. The questionnaire should include questions so that if we present to a human the simulation that explains how it works, we could measure whether it works, whether it works well, and whether the human has acquired a useful understanding with the help of the simulation.

Explanation Satisfaction and Trust Scale is a scale to measure both satisfaction (or understandability in this paper) and trust when building the range of the questions and answers of the questionnaire [67]. It is recommended for XAI, as it is based on literature in cognitive psychology, philosophy of science, and other pertinent disciplines

regarding the features related to explanations. In this context, a Likert scale [68] is commonly used in research and surveys to measure attitude, providing a range of responses to a given question or statement. The typical Likert scale is a 5- or a 7-point ordinal scale used by participants to rate the degree to which they agree or disagree with a question or a statement. Therefore, we opt to use a 5-Likert scale based on the Explanation Satisfaction and Trust Scale in building the questionnaire in the experiment of this paper (*cf.* Section 5.3)².

3. Related Works

Recent works on XAI for intelligent agents and MAS employ automatically generated folk psychology-based explanations [69, 63, 70]. The latter communicate the beliefs and goals that led to the agent’s behavior. An interesting work discussed the generation and the granularity (either *detailed* or *abstract*) of the explanation with a firefighting application [71]. However, the paper is not conclusive in preferring a granularity level. Moreover, the paper concludes that in the special case of belief-based explanations, the efficacy of a detailed explanation is higher than the one of an abstract explanation. The *level* of details, in this special case, are not considered; *i.e.* the paper does not identify a threshold level beyond which explanations are overwhelming for humans.

One work considered the ways intelligent agents should explain themselves to humans [72]. It especially focused on how the soundness (nothing but the truth) and completeness (the whole truth) of the explanations impact the fidelity of humans’ mental models of how a recommender agent works [72]. The work discussed the “sweet spot” between simplicity, *i.e.* simple explanations with little information, and informativeness, *i.e.* complete explanations with too much information. After a human study with 17 participants, the result surprisingly indicated that there is no sweet spot and that the solution is simply to give all the explanations possible to the human. In addition to the questionable validity of the results from a statistical point of view, one possible

²This choice is validated by other relevant works in the literature [67].

reason for such a result is the chosen setting of the study, as there was no challenging situation that provides too many explanations to overwhelm the human user; *i.e.* the work did not consider the limited human cognitive load, which we do in our work. 305 Indeed, as confirmed in the literature, there is a need to align the explanations with the context and human information processing capabilities [44].

One work discusses the amount of information provided to the human and how it affects understandability [73]. It shows that providing humans with detailed explanations about an intelligent agent’s reasoning process can increase their understanding 310 of how the system works. However, information comes at the price of attention, as the human’s time (and interest) is finite, so the solution may not simply be “*the more information, the better*” [73]. In our paper, we investigate thoroughly the amount of information provided in the explanation and the effect the different filters have on the understandability and trust of the human. Empirical human studies are vital to assess 315 the process of explanation reception. Yet, very few works in the literature undertake such experiments [28]. Recent work by Madumal et al. [27] discussed different levels of explanations (none, detailed, and abstract) for reinforcement learning agents. The authors performed an empirical evaluation using a Human-Computer Interaction (HCI) study where participants watch a video game and then fill a questionnaire to collect their responses in terms of explanation quality and trust. Their results show that their model of abstract causal explanations provides better performance in terms of explanation quality (complete, sufficient details, and satisfying) than the benchmark relevant explanation in the reinforcement learning domain. The model does not outperform the benchmark in the “understand” metric. However, the authors note that when 320 comparing their model of explanation with the same scenario but with no explanation, the results are not significantly different for the explanation quality metrics (complete, understand, and satisfying) and only manage to get a significant result in the “sufficient details” metric. Moreover, in terms of the explanation trust metrics (confident, predictable, reliable, and safe), the obtained p – values are not statistically significant 325 using the pairwise Analysis of Variance (ANOVA) parametric test. Furthermore, and surprisingly, the objective understandability after analyzing the score of the task that the participants had to predicate (*i.e.* when implicitly checking if the participants un-

derstood the simulation) is significant for the model in this related work, while the
 335 subjective understandability after analyzing the responses of the participants, *i.e.* when
 explicitly asking the participants in the questionnaire if they understood the simulation,
 is not. In our paper, we argue that the explanations should be formulated by combining
 aspects of the three sub-processes of an explanation (*cf.* Section 2.3). Moreover, re-
 garding the abstraction of explanations, we investigate several context-aware filtering
 340 techniques of explanations, including an adaptive one.

4. Human-Agent Explainability Architecture (HAExA)

4.1. Definition and General Principles of HAExA

As discussed in the previous sections, explanations should take into account the
 context, the features of the explanation (simplicity and adequacy), and the cognitive
 345 load of the human who receives the explanations. Therefore, we define *explanation
 formulation* as follows:

*A process that seeks to maximize the adequacy of AI-generated explanations
 communicated to humans while minimizing their impact on the hu-
 man's cognitive load, i.e. maximizing their simplicity.*

350 To operationalize explanation formulation to a wide range of human-agent inter-
 actions, we introduce the Human-Agent Explainability Architecture (HAExA). This
 architecture allows remote robots, represented as agents and organized in a MAS, to
 expressively explain their behaviors in various situations to humans. The human in
 HAExA is considered as a *human-on-the-loop*³. The latter term refers to humans whose
 355 role in the environment is passive, *i.e.* the human receives explanations for after-action
 decisions, but he/she does not alter the processes in the environment.

The explanation formulation process aims to strike a balance between simplicity
 and adequacy. To implement and operationalize this process, HAExA proposes a dy-
 namic approach to integrate the three phases of explanation (*i.e.* *Generation, Commu-*

³For the sake of conciseness, the term 'human' is henceforth used to refer to the *human-on-the-loop*.

360 *nication*, and *Reception* [44]). In particular, HAExA implements them in the case of remote robots as follows:

1. Explanation Generation in HAExA. Remote robots organized as agents in a MAS in the environment provide *raw explanations* of their behaviors and actions concerning the various situations they face. The way these raw explanations are
365 generated varies according to the explained behavior or the situation, either normal or abnormal. One approach, using reactive architectures, could be to react to the situations according to a set of rules predefined by the human. Another approach, using cognitive architectures, could be achieved by empowering the agents with the ability to reason like humans. Regardless of the approach, the
370 main goal is to provide explanations that include all useful information and that are intelligible to humans.

2. Explanation Communication in HAExA. This phase is handled by assistant agents positioned in between the remote agents on the one hand and the human on the other hand. They are responsible for assuring two tasks: (i) Update the raw
375 explanations to guarantee that the useful information is not missed from them. (ii) Communicate the explanations from the remote agents to the human in a way that considers the human cognitive load, *e.g.* by filtering the raw explanations; This will facilitate a better understanding by the human user, notably because the communicating agents receive the raw explanations from all the remote agents
380 in the MAS. Therefore, they hold a global overview of the system and may be able to pinpoint abnormal situations that were not clear to the remote agents.

3. Explanation Reception in HAExA. The agents communicating the explanations to the humans could be in direct contact with the human to guarantee a better reception of the explanations by the human. A better reception of explanations
385 could also be achieved by building a user model to understand the preferences of the human.

Considering the mentioned phases of explanations, the following section provides a detailed overview of the agents and their roles in HAExA.

4.2. Agents in HAExA

390 Figure 1 visualizes the architecture model HAExA that is composed of three different entities:

- 395 • The right part of the figure represents the MAS. Several **remote agents** are interacting with each other in the environment. Remote agents could be assigned to a group based on their geographical location, capabilities, roles, *etc.* to facilitate the scalability of the architecture. Both in and across groups, collaboration and coordination among agents may occur, while competition is out of the scope of the paper. Generally, all remote agents expose their internal state or a subset of it via a central interface to the human. Consequently, they provide raw explanations of their behaviors to the human.
- 400 • An **assistant agent** (depicted in the center of the figure) that collects the remote agents' raw explanations to communicate filtered explanations to the human, as humans could easily get overwhelmed by the information the remote agents provide. Considering that the assistant agent has a global overview of the environment, it may post-process the raw explanations received from the remote agents to aggregate, update, and filter them; subsequently, it communicates the 405 updated and filtered explanations to the human.
- The **human-on-the-loop** who is the target user of the explanations (in the left part of the figure).

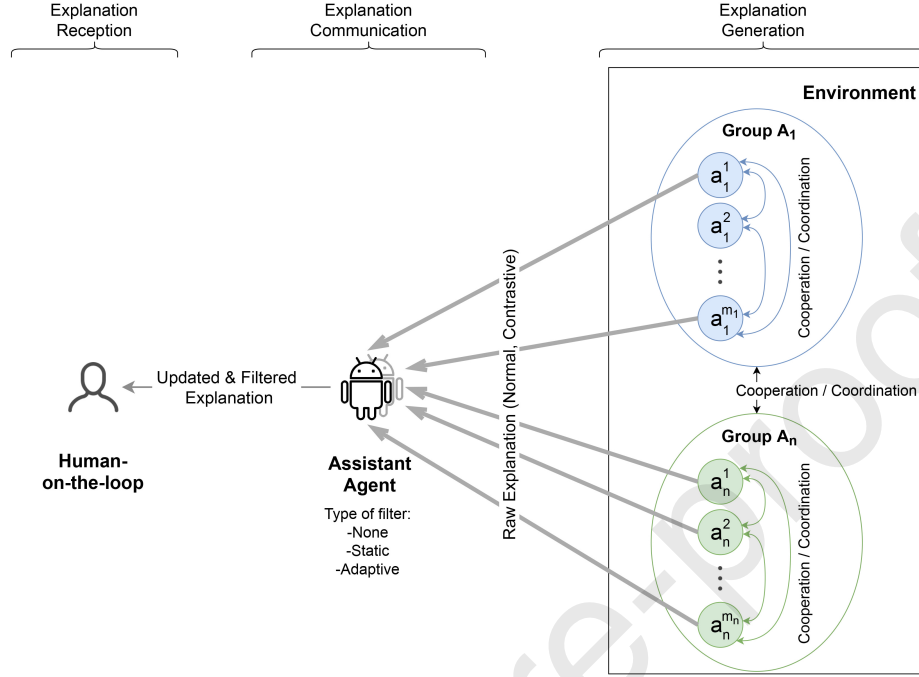


Figure 1: Human-Agent Explainability Architecture (HAExA)

HAExA is flexible and compatible with different agent architectures used by remote
 410 agents and the assistant agent. In this paper, we implement HAExA using the BDI
 architecture (*see* Section 2.5 for the support of this choice).

HAExA can be defined in terms of composition by a triplet $\langle A, AA, H \rangle$. A is the
 set of all remote agents in terms of composition. AA is the assistant agent. H is the
 human-on-the-loop. The set of all remote agents A can be defined by the groups of
 415 remote agents formed as in Equation 1:

$$A = \{A_1, A_2, \dots, A_n\} \quad (1)$$

where $A_i \subseteq A$, $i \in [1..n]$ is a group of individual remote agents. $n \in \mathbb{N}^*$ is the number
 of the groups of remote agents. Let us assume a_i^j is the j^{th} remote agent of the group i .
 This group of remote agents can be defined as in Equation 2.

$$A_i = \{a_i^1, a_i^2, \dots, a_i^{m_i}\} \quad (2)$$

where $m_i \in \mathbb{N}^*$ is the number of remote agents in the group i , with a group of remote
 420 agents having at least one member.

Detailed implementations of the coordination and cooperation among agents are out of the scope of the paper, as these aspects are already covered in-depth by a range of research works. The reader is referred to [74] for more details about the definitions and main characteristics of cooperation and coordination.

425 In HAExA, we focus on how the internal states of remote agents are aggregated and processed to finally be presented as explanations. Even though HAExA permits the modeling of the explanation reception phase, this phase is considered as future work in this paper and no user agents are used. Instead, explanation reception is only analyzed by the empirical human studies provided later in the paper, where ABS is
 430 used to facilitate the reception of explanations by humans.

The remote agents generate raw explanations based on their beliefs and intentions. Later the assistant agent post-processes, based on its beliefs and intentions, the raw explanations by updating and/or filtering them before communicating them to the human. The next section discusses in detail this process.

435 4.3. Explanation Formulation Process

The goal of the explanation formulation process is to provide parsimonious explanations to the human that strike a balance between simplicity and adequacy. The exact nature of the formulation of the explanations depends on the implementation configuration; *i.e.* HAExA supports different explanation formulations with different levels of
 440 technical sophistication. This means that the explanations could be generated in different methods, and could be communicated in several manners to the human as well. Figure 2 shows one possible process model of the explanation formulation pipelines. For the generation, two distinct methods are considered: normal explanations in normal situations, and contrastive explanations in abnormal situations. For the communi-
 445 cation, three means of filtering are considered: static filter, adaptive filter, and no filter. These aspects are discussed in detail in Subsection 4.3.3. The other steps in Figure 2's explanation formulation process are detailed in the following sub-sections.

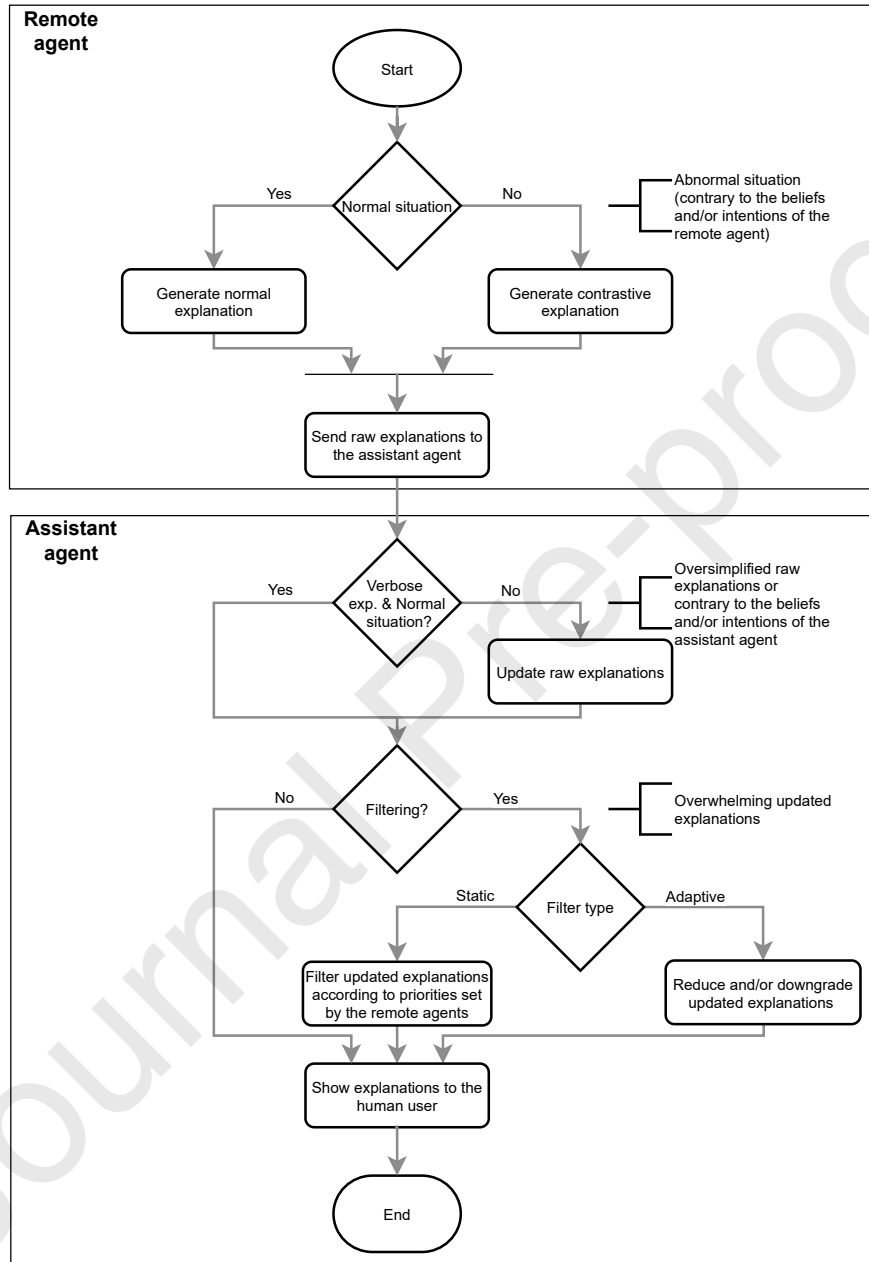


Figure 2: HAExA explanation formulation process

4.3.1. Generating the Raw Explanations

All remote agents provide to the assistant agent the set of all raw explanations $RExp$ that can be, based on Equations 1 and 2 (page 17), represented in Equation 3.

$$RExp = \bigcup_{i=1}^n \bigcup_{j=1}^{m_i} rExp_i^j \quad (3)$$

where $n \in \mathbb{N}^*$ is the number of the groups of remote agents. $m_i \in \mathbb{N}^*$ is the number of remote agents in the group i .

$RExp$ are generated by remote agents only if there is a need for such explanations.

The need arises in two cases:

1. When there is a significant change in the environment, *i.e.* change in the beliefs of the remote BDI agent about the environment. This is measured by comparing the new current beliefs of the agent with its old (or previous) beliefs. For that, we need to define the *change in beliefs* Δ_B , *i.e.* check if there are new beliefs that were not present previously *i.e.* $B \setminus B_{Old}$, and if there were beliefs previously present but disappeared in the updated current beliefs, *i.e.* $B_{Old} \setminus B$. Δ_B is defined in Equation 4. Accordingly, the condition to generate $RExp$ based on the beliefs is defined in Equation 5.

$$\begin{aligned} \Delta_B &= \text{reduceRedundancy}(\text{merge}(B, B_{Old}) \setminus \text{common}(B, B_{Old})) \\ &= \text{reduceRedundancy}((B \setminus B_{Old}) \cup (B_{Old} \setminus B)) \end{aligned} \quad (4)$$

where the function *merge* returns the union of two sets, the function *common* returns the intersection of two sets, and the function *reduceRedundancy* eliminates all redundant attributes.

$$|\Delta_B| > \theta_{Belief} \quad (5)$$

where θ_{Belief} is the threshold of change in beliefs for generating $RExp$. It could be the number of beliefs, beyond which the change in the environment has happened when updating the beliefs.

2. When there is a significant change in the plan, *i.e.* change in the intentions of
 470 the remote BDI agent. This could happen if the agent chooses to abandon a plan
 because it is impossible to achieve or to abandon its intentions because it finds
 better ones. This is measured by comparing the new current intentions of the
 agent with its old (or previous) intentions. For that, and like with the beliefs,
 we need to define the *change in intentions* Δ_I (Equation 6). Accordingly, the
 475 condition to generate *RExp* based on the intentions is defined in Equation 7.

$$\begin{aligned}\Delta_I &= \text{reduceRedundancy}(\text{merge}(I, I_{Old}) \setminus \text{common}(I, I_{Old})) \\ &= \text{reduceRedundancy}((I \setminus I_{Old}) \cup (I_{Old} \setminus I))\end{aligned}\quad (6)$$

$$|\Delta_I| > \theta_{Intention} \quad (7)$$

where $\theta_{Intention}$ is the threshold of change in intentions beyond which the change
 is considered significant.

In case of the need for explaining, the explanations are first generated as raw ex-
 planations *RExp* by the remote agents, and then maybe post-processed (updated and
 480 filtered) by the assistant agent. These remote agents are BDI agents, whose beliefs
 and intentions are used to generate *RExp*. Generally, in normal situations, the remote
 agent generates the raw explanation $rExp_i^{m_i}$ of the action to perform according to the
 intentions it is committed to achieving or his beliefs of the environment, or both. We
 call this type of explanation a *normal* explanation. It is stating the next step in the
 485 plan to execute, *i.e.* what to do next, and *sometimes* the reason for such action (Belief,
 Intention, or both). Examples of normal situations related to the application of deliver-
 ing packages using Unmanned Aerial Vehicles (UAVs): “UAV 1 is moving to Package
 1”, “UAV 1 is delivering Package 1 to Storehouse S”, “UAV 1 is moving to Charging
 Station C because of low battery”, “UAV 1 is charging battery”, *etc.*

490 When the change in beliefs Δ_B is major, *i.e.* above a certain threshold, this change
 may lead to major changes in intentions Δ_I , *i.e.* above a certain threshold, and ac-
 cordingly, the situation is considered abnormal. In such situations, we consider that

contrastive explanations are preferable, and this choice is aligned with recent works in the literature [28]. Consider the following example that is leading to an abnormal situation. Initially, a UAV (UAV 1) is moving to a package (package 1) with the intention to deliver it. However, when UAV 1 realizes that the package is delivered by another UAV (UAV 2), it abandons the mission and moves towards another package (package 2). The following two explanations of the atomic actions *move* could be provided: (i) “UAV 1 is moving towards Package 1”; (ii) “UAV 1 is moving towards Package 2”. In this situation, the human may ask “why did this happen?” and “why did UAV 1 not carry out the delivery of Package 1?” The human will be curious about knowing why UAV 1 did not do what it was supposed to do and is instead moving to another package. To solve this, an alternative contrastive explanation is provided as follows: “UAV 1 is moving towards Package 2 instead of Package 1 because Package 1 is delivered by another UAV”.

For a human, when receiving explanations about the behavior of the remote agents, it is generally not the normal behavior that may require an explanation, but rather the abnormal behavior. Therefore, HAExA makes use of contrastive explanations to represent abnormal situations. Humans can explain normal behaviors with the help of their own experiences and expectations. However, the abnormal behavior of the agent challenges these experiences and expectations, and therefore, an explanation is deemed necessary.

The current plan of actions π adopted by the agent overrides the previous one π_{Old} when there are significant Δ_B and Δ_I *i.e.* abnormal situations. Accordingly, there could be several options for generating contrastive explanations where $a1$ is an action from π and $a2$ is an action from π_{Old} . The generation is governed by the execution condition C that could be either the actual beliefs B or/and the actual intentions I . That is why we need to keep a track of the previous plan that includes the previous actions that the agent was supposed to perform but did not. These options are:

1. $a1$ and not $a2$ because of C ;
2. Not $a2$ because of C (where $a1$ is implicit);
3. $a1$ because of C (where not $a2$ is implicit).

The second option is trivial because later, the remote agent must state its current action, and hence this will be done later anyway. Both options 1 and 3 are good candidates. Note that the third option is changing the type of contrastive explanation into a normal one by dropping the counterfactual part “not A2”. This change is appealing to reduce the length of the explanation when it can be implicitly inferred by the human *i.e.* to increase simplicity. The next sections provide more details on how the assistant agent first updates the raw explanations generated by the remote agents (Section 4.3.2) and then filters the resulting updated explanations before communicating them to the human (Section 4.3.3).

4.3.2. Updating the Raw Explanations

Considering that the assistant agent has a global view of the situation, the abnormality of some situations is different from its perspective compared to the limited perspective of the remote agents. Accordingly, the assistant agent may update *RExp* based on the abnormality of the situation. This sub-step of the explanation formulation process is important to assure that the generation of explanations did not risk the *oversimplification* of the explanations, *i.e.* to assure that *UExp* are adequate. The results of this step are the *updated explanations UExp*.

In this sub-step, the assistant agent scans *RExp* for anomalies and inconsistencies (*e.g.* two remote agents providing conflicting information) and removes any unnecessary information from *RExp* or adds missing necessary information that is not seen by the remote agents when generating *RExp* due to their limited view of the situations. In other words, even though the remote agents consider the abnormal situations when generating the contrastive explanations, the assistant agent, after receiving all *RExp*, could discover some abnormality hidden to the remote agents.

This sub-step is context-aware to the situation, *i.e.* it adaptively updates *RExp* based on the context of the situation. The situations are considered abnormal according to the assistant agent based on the *change in beliefs* Δ_B defined in Equation 4 and the *change in intentions* Δ_I defined in Equation 6. Accordingly, in this sub-step, the *RExp* generated by the remote agents are updated. To achieve this, the assistant agent, which holds a general comprehensive overview of the context situation, may aggregate more

useful and consistent explanations for the human by updating $RExp$ generated by the remote agents. Accordingly, the type of normal explanations by some remote agents
 555 could be changed into contrastive explanations, when adequacy is needed, by adding the counterfactual part of the explanations.

4.3.3. Filtering of the Updated Explanations

The filtering of explanations is conducted to assure that $UExp$ are simple and not overwhelming for the human, *i.e.* to achieve simplicity. This sub-step is adaptive to
 560 the number of explanations and type of explanations (normal or contrastive that differ in length) and accordingly, the filtering by the assistant could be strict or not based on the human cognitive load, *i.e.* it adaptively filters $UExp$ to not exceed his/her cognitive load threshold. Three cases of filtering are presented below:

1. *Without a filter*: if few remote agents are present, it might be relevant for the
 565 human to be able to distinguish between the beliefs of individual remote agents and understand their explanations without filtering.
2. Using a *static filter* where the explanations are filtered based on priorities in accordance with the human cognitive load threshold. The remote agents set priorities to $RExp$ before sending them to the assistant agent and every explanation
 570 with a priority below the threshold will be filtered out by the assistant agent. The filtering rules, here, are not context-dependent. The priorities set by the remote agents are compared to the human cognitive load threshold. The values of the priorities and thresholds are defined empirically.
3. Using an *adaptive filter* based on the current context, where irrelevant explanations
 575 are removed; for example, if many remote agents are present in the environment, the assistant agent may decide to aggregate their explanations because a human can not process differences in the explanations of individual remote agents in real-time. The adaptive filter could also adapt to the human preferences if a user model is built.

580 For adaptive filtering, three levels of adaptation based on the empirically-defined human cognitive load are defined:

- *FilterThreshold_H*: If the number of *UExp* is higher than this threshold, change the type to *UExp* from contrastive to normal explanations using the function *changeType* and reduce the number of *UExp* using the function *reduce*. Reducing the number of explanations may lead to fully discarding them. This type is defined in Equation 8.

$$|UExp| > FilterThreshold_H \rightarrow reduce(UExp), changeType(UExp_{Contrastive}) \quad (8)$$

- *FilterThreshold_M*: If the number of *UExp* is higher than this threshold, reduce the number of *UExp* using the function *reduce*. This type is defined in Equation 9.

$$|UExp| > FilterThreshold_M \rightarrow reduce(UExp) \quad (9)$$

- *FilterThreshold_L*: If the number of *UExp* is higher than this threshold, change the type of *UExp* from contrastive to normal explanations using the function *changeType*. This type is defined in Equation 10.

$$|UExp| > FilterThreshold_L \rightarrow changeType(UExp_{Contrastive}) \quad (10)$$

The values of these three thresholds are defined empirically. It is worth mentioning that before attempting to update *RExp* into *UExp*, the assistant agent verifies if there is a need for filtering or not. This is to avoid changing the type of *RExp* from normal to contrastive and then change back the type of *UExp* from contrastive to normal because they tend to be overwhelming. This condition can be found in Figure 2.

5. Experimental Case Study

5.1. Description

With the rapid increase of the world's urban population, the infrastructure of the constantly expanding metropolitan areas is subject to immense pressure. To meet the

growing demand for sustainable urban environments and improve the quality of life for citizens, municipalities will increasingly rely on novel transport solutions. In particular, UAVs, commonly known as drones, are expected to have a crucial role in future smart cities thanks to relevant features such as autonomy, flexibility, mobility, and adaptivity [75]. Therefore, over the past few years, an increasing number of public and private research laboratories have been working on civilian, small, and human-friendly UAVs.

Still, several concerns have been raised regarding the possible consequences of introducing UAVs in crowded urban areas, especially regarding people's safety. To guarantee it is safe that UAVs fly close to human crowds and to reduce costs, different scenarios must be modeled and tested. Yet, to perform tests with real UAVs, one needs access to expensive hardware. Moreover, field tests usually consume a considerable amount of time and require trained people to pilot and maintain the UAVs. Furthermore, on the field, it is hard to reproduce the same scenario several times [76]. In this context, the development of computer simulation frameworks that allow transferring real-world scenarios into executable models is highly relevant [77, 78]. However, the simulation frameworks have their drawbacks; in particular, it is impossible to fully reproduce the real environment. The use of ABS frameworks or tools for UAV simulations is gaining more interest in complex civilian applications where coordination and cooperation are necessary [79]. Due to operational costs, safety concerns, and legal regulations, ABS is commonly used to implement models and conduct tests for UAVs [80]. This has resulted in a range of research and applied works addressing ABS in UAVs [51].

The problem of understanding the robot's SoM is more accentuated in the case of UAVs since – as confirmed by recent studies in the literature [81, 82] – remote robots tend to instill less trust than robots that are co-located. For this reason, working with remote robots is a more challenging task, especially in high-stakes and dynamic scenarios such as flying UAVs in urban environments. To overcome this challenge, this case study relies on XAI to trace the decisions of agents and facilitate human intelligibility of their behaviors in the context of civilian UAV swarms that are interacting with other objects in the air or the smart city. Indeed, as has been confirmed by user studies, providing explanations about the remote UAV decisions may increase the satisfaction

of people [83], and maintain the acceptability of the XAI system [48]. For instance, an XAI system could enable a delivery UAV modeled as an agent to explain, to its remote operator, the reasons behind its deviation from a predefined plan (*e.g.* to avoid placing fragile packages in unsafe locations) thereby allowing the human operator to better manage a set of such UAVs. The example can be extended, in a multi-agent environment, where UAVs can be organized in swarms [84] and modeled as cooperative agents to achieve more than what they could do solely, and the XAI system could explain this to the human operator for the sake of transparency, control, or for the sake of training novice operators on the system.

Our previous works [85, 86] have discussed the role of filtering of explanations in three cases: No explanation, Detailed explanation, and Filtered explanation. In our previous works, we have investigated the following two research hypotheses:

RH-A: Explainability *increases the understandability* of the human-on-the-loop in the context of remote agents⁴.

RH-B: Too many details in the explanations *overwhelm* the human-on-the-loop, and hence in such situations, the **filtering of explanations** provides less, concise and synthetic explanations leading to higher understandability by the human.

The responses of the participants were statistically analyzed, validated in terms of significance, and presented based on *Mann-Whitney U* non-parametric tests. Accordingly, the results showed that when comparing the responses of the group that received no explanations with the group that received detailed explanations, explainability *increases the understandability* of the human-on-the-loop in the context of remote agents, *i.e.* *RH – A* is proven. Additionally, comparing the responses of the group that received detailed explanations with the group that received filtered explanations revealed that providing more details is preferred by the participants. However, with too many details, the participants are eventually overwhelmed, and in this case, the filtering of explanations is essential, *i.e.* *RH – B* is proven.

⁴Remote agents represent the remote robots.

660 In these previous works, only normal explanations were used, *i.e.* contrastive explanations were neither considered for normal nor abnormal situations. Furthermore, only static filtering, *i.e.* no adaptive filtering, was employed. These simplifications were made because the main goal of our previous works was to reproduce the results of the literature in the domain of remote robots represented as agents. In this paper, we
 665 go further by investigating new research hypotheses on how parsimonious explanations could be formulated in XAI. These hypotheses are based on the XAI literature [87] and built based on our previous findings:

RH1-1: Combining **adaptive filtering** with only normal⁵ explanations *increases the understandability* of the human-on-the-loop compared to **static filtering** with
 670 only normal explanations.

RH1-2: Combining **adaptive filtering** with normal and **contrastive** explanations *increases the understandability* of the human-on-the-loop compared to **static filtering** with only normal explanations.

RH1-3: Combining **adaptive filtering** with normal and **contrastive** explanations *increases the understandability* of the human-on-the-loop compared to **adaptive filtering** with only normal explanations.
 675

RH2: Combining **adaptive filtering** with normal and **contrastive** explanations *increases the trust* of the human-on-the-loop compared to **static filtering** with only normal explanations.

680 To accept or reject the new research hypotheses, we have used an ABS to simulate an application of UAVs' autonomy and explainability. The case study is performed as an HCI statistical experiment. The participants watch the simulation running and fill out a questionnaire built according to the XAI metrics in the literature [67]. Some questions taken from [67] are adapted to consider the particularity of the experiment.
 685 The results of the questionnaire will be used to investigate the human understandabil-

⁵In this hypothesis and all the following ones, we consider that normal explanations are non-contrastive explanations.

ity of the explanations provided by the UAVs. As the coordination and cooperation between groups of remote agents in MAS are out of the scope of this paper, we opt to simplify the implementation of HAExA (*cf.* Figure 1) by choosing only one group. Additionally, the maximum number of remote UAVs is 10.

690 5.2. Experiment Scenario

The experiment scenario is about investigating the role of XAI in the communication between UAVs and humans in the context of package delivery in a smart city. In the scenario, one human-on-the-loop operator⁶ oversees several UAVs that provide package delivery services to clients. These UAVs autonomously conduct tasks and take 695 decisions when needed. Additionally, they need to communicate and discuss with each other and may cooperate to complete a specific task. The UAVs explain to the assistant agent the progress of the mission, including unexpected situations, along with the decisions made by them. Figure 3 shows the interaction between the actors in the proposed experiment. In what follows, the steps of the experiment scenario are detailed:

- 700 1. When a client sends a request for delivering a package, a notification is sent to the UAVs. We assume that UAVs are connected with each other and with the assistant agent using a reliable network.
2. UAVs that are nearby, *i.e.* situated within a specific radius to the package, coordinate to complete the delivery mission. In other words, if a UAV is very far from 705 the package/passenger, it should not participate in the discussion related to this transportation mission. The decentralized coordination (without the intervention of the operator) can be initiated to undertake several decisions including:
 - **Best candidate:** Deciding which UAV will deliver the package according to constraints: actual distance to the package, battery size, having other packages in hand, having a mission with a near destination, *etc.*

710

⁶Human-on-the-loop: the human is not part of the system and consequently cannot influence the outcomes of the system behavior or the simulation, but can perceive these outcomes.

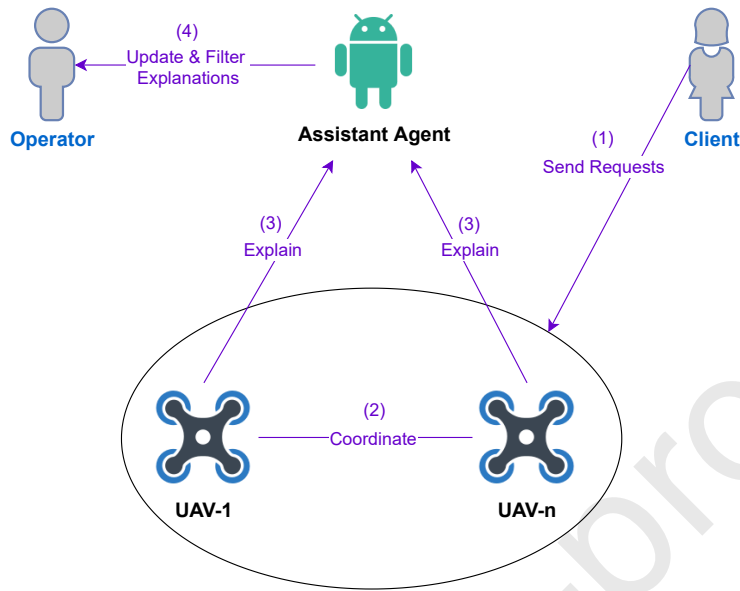


Figure 3: Interaction of actors in the experiment scenario

- **Long trip:** For long trajectories, there is a need to cooperate to deliver the package between several UAVs, where each UAV delivers the package part of the way and then hands it to another UAV.

3. The explanation required from the UAV is generally about the mission progress, its decisions, and its status, *e.g.* which UAV is assigned to the mission, or when the UAV picks up the assigned package and is moving to the destination. However, other important types of explanation are required regarding the unexpected (or abnormal) situations, *e.g.* the UAV arrives at the package location and does not find it, or realizes the package does not comply with its description. Another example is when a UAV needs charging and that is why it ignores a nearby package.

4. Every UAV generates raw explanations to the assistant agent who in turn communicates them to the operator. The assistant agent may update and filter these raw explanations received from the UAVs to give a summary of the most important explanations thereby avoiding overwhelming the operator with a lot of

730 details. There are two types of generated explanations: (i) Normal raw explanations generated by the UAVs, (ii) Contrastive explanations that are contrastive raw explanations generated by the UAVs and updated by the assistant agent. Also, the assistant agent applies two types of explanation filtering: (i) Static filtering based on a *filtering threshold* set by the human. It filters the explanations based on their priorities. The latter are set by the UAVs when generating each raw explanation. (ii) Adaptive context-aware filtering where the assistant agent adapts the intensity and levels of filtering based on the complexity and scalability of the situation.

735 The experiment requires that the participants (the operator in the experiment), after watching the simulation of the experiment, fill a questionnaire to collect their opinions on the explanations provided by the agents in the experiment. Section 5.3 discusses the details of the questionnaire.

5.3. Building the Questionnaire

740 We opt to use the Explanation Satisfaction and Trust Scale in building our questionnaire (*cf.* Section 2.6.2). The answers are distributed to a 5-points Likert scale [67]: 1 (I disagree strongly), 2 (I disagree somewhat), 3 (I'm neutral about it), 4 (I agree somewhat), and 5 (I agree strongly).

5.3.1. Categories of the Questions

745 The built questionnaire has 21 questions divided into 3 categories:

1. **Participant Details (5 questions):** Gender (optional), age (optional), level of English language, prior knowledge about UAVs, and year of study.
2. **Functionalities (3 questions):** This category is used to check whether the participant understood the simulation using some objective questions. It includes two open-ended questions to evaluate the Objective Understandability metric (*cf.* Table 6). Additionally, we confirm if the functionalities of the simulation are acceptable by the participant and their suggestions in this context.

750

3. **Statistical Analysis (12 questions, cf. Table 1)** These questions are mainly about understandability and trust. We investigate the following metrics: Overwhelmingness, Subjective Understandability, Confidence, Predictability, Reliability, Efficiency, Wariness, Satisfaction, and Sufficient Details.

Finally, the questionnaire includes a question (numbered Q18 in the experiment) about *Curiosity* considering that situations like the one under study lead people to engage in effortful processing and motivates them to seek out additional knowledge to gain insight and fulfill their curiosity [88]. This question is: “Why do you think the explanation of the simulation tool is important? Check all that apply”. However, unlike all the questions in the questionnaire, this question has multiple answers, and because this paper does not investigate curiosity, this question is not analyzed in the results of this paper.

5.4. Conducting the Experiment

We have conducted the experiment based on the experiment scenario. The experiment was conducted online with 90 participants. It is important here to mention that before conducting the experiment, all participants have been informed that the gathered data is anonymous, secured, and will be used solely for research purposes. Moreover, The survey was divided into three steps. The participants received a clarification in English and provided their consent to continue with the experiment. Moreover, they have been informed of their rights to know how the data is used according to the General Data Protection Regulation⁷.

Before conducting the experiment, we provided some information about the simulation: (i) We explained the main goal of the simulation, which is the delivery of packages using UAVs. The delivery of a package is from any point on the map, where a package could appear to some warehouse determined in the map, (ii) Icons of the elements (UAVs, charging station, packages, destinations, etc.) (iii) We mentioned that while the UAVs are delivering the packages, some abnormal or unexpected situations may happen. All situations, either normal or abnormal, will be explained by the UAVs.

⁷<https://gdpr-info.eu/>

The simulation is divided into 5 sequences. The sequences are presented one after the other without the option to go back to a previous one. The point of each sequence is to show a scenario of package delivery with some abnormal situations. The first sequence is a very simple example that does not include any abnormal situation, so
785 it is like a happy path situation, which helped the participants understand the context and the appearance of the simulation and be familiar with the different elements with their icons. Each of the other sequences handles an abnormal situation or more, for example: low battery, damaged package, already delivered package, *etc.* The second sequence has moderate abnormal situations, while the third and last sequence is an
790 overwhelming sequence with several UAVs (here 10).

The experiment in this paper goes more steps forward to investigate various ways and manners to provide parsimonious explanations that strike a balance between simplicity and adequacy. In the next section, we investigate the research hypotheses *RH1 – 1*, *RH1 – 2*, *RH1 – 3*, and *RH2*. In the following section, we mention the
795 details about the methodology of this experiment.

5.5. Experiment Methodology

The experiment is conducted online, where the simulations are prepared as high-quality videos. The experiment instructions along with the links to the questionnaire and the videos are provided in a presentation with a signal link. To reach the partic-
800 ipants of the experiment, we broadcasted the link of the experiment on mailing lists. Moreover, we have also posted the link of this experiment to social networks. To obtain the sample used in the analysis of the experiment, we focused on voluntary sampling. People who received the link chose to participate or not in our experiment. Voluntary sampling has some advantages, such as the simple way to conduct the experiment, in-
805 expensiveness, easy data collection, easy access, *etc.* However, it has also some drawbacks such as response biases, *i.e.* sample members are self-selected volunteers. Voluntary participants watch the simulation running and then fill out a questionnaire built to aggregate their responses.

Section 5.5.1 outlines the specific implementation details of the experiment. Sec-
810 tion 5.5.2 details the process of organizing the participants in groups. Section 5.5.3

discusses the statistical testing choices, mainly parametric and non-parametric tests, to investigate the responses of the participants.

5.5.1. Implementation of the Experiment

The experiment is implemented using the JS-son agent-oriented programming library [89, 90]. The agents' reasoning loop and environment management in the JS-son library is documented in detail in [89]⁸. The beliefs of these agents change according to the situation, and accordingly, the parsimonious explanations are formulated. In the case of remote agents representing the UAVs, the explanation formulation process helps in the explanation generation phase by generating raw normal explanations in normal situations and raw contrastive explanations in abnormal ones based on the change of the beliefs and intentions of the remote agents, *i.e.* these agents are *adaptive* and *context-aware* when generating raw explanations. For the assistant agent, this process allows for updating the raw explanations to ensure they have all the necessary information, *i.e.* adequacy, in a *combined approach* between the generation and communication phases of explanation. The formulation process also guarantees the filtering of the updated explanations, *i.e.* simplicity, in overwhelming scenarios in the explanation communication phase based on the human cognitive load.

The simulation runs on a machine with the following features: Win 10 Education, Core i7 2.9 GHz 4 cores, 32 GB RAM, 4 GB dedicated video memory. The last sequence of the simulation (overwhelming sequence) lasts for 1:35 minutes and includes: 10 UAVs, 8 warehouses, 10 charging stations, 27 packages to be delivered, 9 abnormal situations. For more technical details on the implementation of the experiment, we refer the reader to our demonstration paper [90].

5.5.2. Participants and Groups

The representative sample is composed of 90 participants. *i.e.* 90 participants have participated in this experiment. They were randomly divided into three groups (*SF*,

⁸At the time of writing, documentation pages of the library are available online at <https://js-son.readthedocs.io/>.

AF, and *AC*). All the three groups watch exactly the same simulation sequences but with different explanation techniques:

1. Group *SF* (30 participants) watches the simulation with normal explanations and static filtering;
2. Group *AF* (30 participants) watches the simulation with normal explanations and adaptive filtering;
3. Group *AC* (30 participants) watches the simulation with normal and contrastive explanations and adaptive filtering.

After watching the simulation sequences, the participants filled out the questionnaire of the experiment. The first 8 questions of the questionnaire are the *Participant Details* and *Functionalities* categories (*cf.* Section 5.3.1). The distribution of the participants is as follows: 20 of the participants were females, and 63 were males, and 7 preferred not to disclose this type of information. They were aged between 18 and 45 (mean of age $\bar{x}_{age} = 26.44$, and standard deviation of age $s_{age} = 7.348$). Regarding the *a priori* knowledge the participants had about UAVs, they self-rated their knowledge using 5-points Likert as (mean of UAV knowledge $\bar{x}_{UAV_knowledge} = 3.27$, and standard deviation of UAV knowledge $s_{UAV_knowledge} = 1.1$). Therefore, the randomly selected participants of the experiment are heterogeneous regarding their age, sex, and knowledge of UAVs.

Excluding the only question with multiple answers (Q18), the questions under study in the statistical analysis of this experiment are 12 questions. They are analyzed and discussed in Section 6 (*cf.* Table 1 for the list of these questions).

5.5.3. Statistical Testing Methodology

To perform the experiment, the paper focuses on qualitative data. In fact, the two most common types of qualitative data (nominal and ordinal) are used in the experiment. The nominal data refers to the groups of the participants involved in the experiments, while the ordinal data refers to their opinions about the explanations. To evaluate these opinions, the ordinal data are based on the 5-points Likert scale [91].

865 The writing of these choices of responses may differ in some questions but the scale is
the same (*cf.* Section 5.3).

While the Likert scale is widely used in scientific research, there has been a long-
standing controversy regarding the analysis of ordinal data [92]. In fact, analyzing the
outcomes of the Likert scale, and the use of parametric tests to analyze ordinal data in
870 general, has been subject to an active and ongoing debate involving the advocates of
Likert scale's compatibility with parametric testing [93, 94, 92], and those opposed to
this idea [95, 96, 97] who consider that the analysis of Likert scales must be done with
non-parametric tests such as Kruskal-Wallis or Mann-Whitney [95].

Delving into the details of this discussion is beyond the scope of this paper. In
875 addition, to avoid biases in the data analysis and due to this ongoing dispute between
statisticians, in this paper, we conduct both the parametric test that is *ANOVA* and
the non-parametric test that is the *Kruskal-Wallis* test. The next section provides full
analysis and discussion of the results of the experiment.

6. Experimental Results

880 In the experiment, as stated before, each of the three groups *AF*, *SF*, and *AC* is
composed of 30 participants. The following section provides a detailed analysis of the
results.

6.1. Initial Verifying of the Significance

For each of the 12 questions under study (Table 1), the null hypothesis is $H_0 : \mu_{SF} =$
885 $\mu_{AF} = \mu_{AC}$ for ANOVA (resp. $H_0 : \text{med}_{SF} = \text{med}_{AF} = \text{med}_{AC}$ for Kruskal-Wallis). In
other words, the null hypothesis H_0 , for each question, assumes that the differences
between the means for ANOVA (respectively, the medians for Kruskal-Wallis) are not
significant. The alternative hypothesis is H_1 : at least one mean is different for ANOVA
(respectively, at least one median is different for Kruskal-Wallis).

890 Table 1 outlines the statistical significance of the results obtained by both ANOVA
and Kruskal-Wallis (KW) in our experiment. As presented by this table, although the
 p – values of ANOVA and Kruskal-Wallis are different, the results of significance are

similar and aligned between these two tests *i.e.* when the null hypothesis is rejected by ANOVA, it is also rejected by Kruskal-Wallis, and when the null hypothesis is accepted
 895 by ANOVA, it is also accepted by Kruskal-Wallis.

In Table 1, four questions are associated with the metric "Subjective Understandability" (Q10, Q11, Q12, and Q19). Apart from the individual question analysis, it is important to analyze the questions as a group to investigate the internal consistency of this metric. Therefore, we conduct a Cronbach's alpha test as a measure of scale reliability.
 900 This test is defined in Equation 11, where k denotes the number of questions, σ_i^2 denotes the variance associated with question i and σ_X^2 denotes the variance associated with the observed total scores.

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_X^2} \right) \quad (11)$$

The outcome of the Cronbach's test (Table 2) reveals that $.9 > \alpha = .875 \geq .8$. This means, according to DeVellis [98], that the internal consistency for the metric
 905 "Subjective Understandability" is Good⁹.

Additionally, Table 3 outlines the importance of the four questions in this group. We can notice that the deletion of Q11, Q12, or Q19 reduces the α value. This means that these questions seem more important than Q10, whose deletion increases the α value.

910 Returning to the analysis of the 12 questions under study, and considering we got the same initial significance results for both ANOVA and Kruskal-Wallis in Table 1, note that we could have continued the data analysis with ANOVA due to the power of parametric tests *i.e.* they give better results to reject the null hypothesis as, according to G. Normann, parametric tests are generally more robust than non-parametric ones [92].
 915 However, due to the ongoing dispute of using parametric and non-parametric testing for ordinal data, we opt to perform both tests.

On the one hand, for the pairwise comparison with ANOVA, this paper focuses on *Tukey Honest Significant Difference (Tukey HSD)* test because all the groups have

⁹Internal consistency scale of Cronbach's alpha for Likert scale questions is: Excellent, Good, Acceptable, Questionable, Poor, Unacceptable.

Table 1: The p – value of each investigated question in the experiment for both ANOVA and Kruskal-Wallis tests

Question	Metric	p – value of ANOVA	p – value of KW
Q9: The number of drones (10 drones) in the last scenario was not overwhelming (too much to follow).	Overwhelmingness	.006	.012
Q10: Do you believe the only one time you watched the simulation tool working was enough to understand it?	Subjective Understandability	.001	.001
Q11: How well the simulation tool helped you to understand how it works?	Subjective Understandability	.000	.000
Q12: How do you rate your understanding of how the simulation tool works?	Subjective Understandability	.001	.002
Q13: I am confident in the simulation tool. I feel that it works well.	Confidence	.088	.096
Q14: The outputs of the simulation tool are very predictable.	Predictability	.039	.045
Q15: The simulation tool is very reliable. I can count on it to be correct all the time.	Reliability	.108	.091
Q16: The simulation tool is efficient in that it works very quickly.	Efficiency	.760	.939
Q17: I am wary of the simulation tool.	Wariness	.149	.134
Q19: From the explanation, I understand better how the simulation tool works.	Subjective Understandability	.000	.000
Q20: The explanation of how the simulation tool works is satisfying.	Satisfaction	.053	.061
Q21: The explanation of how the simulation tool works in the last sequence has sufficient details.	Sufficient Details	.001	.002

Table 2: Reliability Statistics – Cronbach's alpha test

	Outcome
Cronbach's alpha	.875
Cronbach's alpha based on standardized items	.880

Table 3: Variables Statistics – Cronbach's alpha test

Question (Variable)	Scale mean if question is deleted	Scale variance if question is deleted	Corrected question total correlation	Squared multiple correlation	Cronbach's Alpha if question is deleted
Q10: Do you believe the only one time you watched the simulation tool working was enough to understand it ?	11.09	8.801	.603	.383	.896
Q11: How well the simulation tool helped you to understand how it works ?	10.77	8.630	.750	.592	.833
Q12: How do you rate your understanding of how the simulation works ?	10.79	8.483	.843	.735	.800
Q19: From the explanation, I understand better how the simulation tool works ?	10.96	8.537	.759	.654	.830

the same size (30 participants per group), and the homogeneity of variance is verified
 920 by the data. On the other hand, the paper conducts a post Kruskal-Wallis analysis
 with Bonferroni correction. Table 4 outlines the pairwise comparison results obtained
 by Tukey HSD ANOVA and Table 5 outlines the pairwise comparison results obtained
 by Post Kruskal-Wallis with Bonferroni correction for all the questions with significant
 p -values in Table 1. Note that, in Tables 4 and 5, *SF* means static filtering with normal
 925 explanations, *AF* means adaptive filtering with normal explanations, and *AC* means
 adaptive filtering with normal and contrastive explanations. The results discussed in
 the following sections show the equivalence of significance for these two tests.

Table 4: Tukey HSD pairwise ANOVA comparisons of the groups in the experiment

Question (Dependent Variable)	(I) Participant Group	(J) Participant Group	(I-J) Mean Difference	Std. Error	(Sig.) <i>p</i> -value	95% Confidence Interval	
						Lower Bounds	Upper Bounds
Q9: The number of drones (10 drones) in the last sequence was not overwhelming (too much to follow)	<i>AF</i>	<i>SF</i>	.667	.301	.074	-.05	1.38
	<i>AC</i>	<i>SF</i>	.967*		.005	.25	1.68
	<i>AC</i>	<i>AF</i>	.300		.581	-.42	1.02
Q10: Do you believe the only one time you watched the simulation tool working was enough to understand it?	<i>AF</i>	<i>SF</i>	.600	.297	.113	-.11	1.31
	<i>AC</i>	<i>SF</i>	1.133*		.001	.43	1.84
	<i>AC</i>	<i>AF</i>	.533		.177	-.17	1.24
Q11: How well the simulation tool helped you to understand how it works?	<i>AF</i>	<i>SF</i>	.733*	.256	.014	.12	1.34
	<i>AC</i>	<i>SF</i>	1.200*		.000	.59	1.81
	<i>AC</i>	<i>AF</i>	.467		.168	-.14	1.08
Q12: How do you rate your understanding of how the simulation tool works?	<i>AF</i>	<i>SF</i>	.533	.251	.090	-.06	1.13
	<i>AC</i>	<i>SF</i>	1.000*		.000	.40	1.60
	<i>AC</i>	<i>AF</i>	.467		.156	-.13	1.06
Q14: The outputs of the simulation tool are very predictable	<i>AF</i>	<i>SF</i>	.033	.200	.985	-.44	.51
	<i>AC</i>	<i>SF</i>	-.433		.084	-.91	.04
	<i>AC</i>	<i>AF</i>	-.467		.057	-.94	.01
Q19: From the explanation, I understand better how the simulation tool works	<i>AF</i>	<i>SF</i>	.867*	.250	.002	.27	1.46
	<i>AC</i>	<i>SF</i>	1.367*		.000	.77	1.96
	<i>AC</i>	<i>AF</i>	.500		.117	-.10	1.10
Q21: The explanation of how the simulation tool works in the last sequence has sufficient details	<i>AF</i>	<i>SF</i>	.967*	.295	.004	.26	1.67
	<i>AC</i>	<i>SF</i>	1.000*		.003	.30	1.70
	<i>AC</i>	<i>AF</i>	.033		.993	-.67	.74

Table 5: Post Kruskal-Wallis pairwise comparison with Bonferroni correction of the groups in the experiment

Question (Dependent Variable)	(I) Participant Group	(J) Participant Group	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
Q9: The number of drones (10 drones) in the last sequence was not overwhelming (too much to follow)	<i>AF</i>	<i>SF</i>	13.567	6.463	2.099	.036	.107
	<i>AC</i>	<i>SF</i>	18.483		2.860	.004	.013
	<i>AC</i>	<i>AF</i>	4.917		.761	.447	1.000
Q10: Do you believe the only one time you watched the simulation tool working was enough to understand it?	<i>AF</i>	<i>SF</i>	-11.033	6.545	-1.686	.092	.275
	<i>AC</i>	<i>SF</i>	-24.067		-3.677	.000	.001
	<i>AC</i>	<i>AF</i>	-13.033		-1.991	.046	.139
Q11: How well the simulation tool helped you to understand how it works?	<i>AF</i>	<i>SF</i>	-15.800	6.394	-2.471	.013	.040
	<i>AC</i>	<i>SF</i>	-27.100		-4.238	.000	.000
	<i>AC</i>	<i>AF</i>	-11.300		-1.767	.077	.232
Q12: How do you rate your understanding of how the simulation tool works?	<i>AF</i>	<i>SF</i>	-10.467	6.462	-1.620	.105	.316
	<i>AC</i>	<i>SF</i>	-22.983		-3.557	.000	.001
	<i>AC</i>	<i>AF</i>	-12.517		-1.937	.053	1.58
Q14: The outputs of the simulation tool are very predictable	<i>AF</i>	<i>SF</i>	-1.033	6.153	-.168	.867	1.000
	<i>AC</i>	<i>SF</i>	12.733		2.070	.038	.115
	<i>AC</i>	<i>AF</i>	13.767		2.238	.025	.076
Q19: From the explanation, I understand better how the simulation tool works	<i>AF</i>	<i>SF</i>	-20.033	6.504	-3.080	.002	.006
	<i>AC</i>	<i>SF</i>	-32.267		-4.961	.000	.000
	<i>AC</i>	<i>AF</i>	-12.233		-1.881	.060	.180
Q21: The explanation of how the simulation tool works in the last sequence has sufficient details	<i>AF</i>	<i>SF</i>	20.250	6.557	3.088	.002	.006
	<i>AC</i>	<i>SF</i>	20.650		3.149	.002	.005
	<i>AC</i>	<i>AF</i>	.400		.061	.951	1.000

6.2. Detailed Data Analysis with ANOVA and Kruskal-Wallis

While our methodology, metrics, and testing approach is adopted from Hoffman et al. [67], by considering objective understandability in Q6 and Q7 (see Section 6.3), we adapt Hoffman's satisfaction scale and use the descriptor "Understandability", subjectively and objectively, in our approach.

Investigating the Subjective Understandability. The metrics Overwhelmingness (Q9), Subjective Understandability (Q10, Q11, Q12, and Q19), Satisfaction (Q20), and Sufficient Details (Q21) are considered. All the obtained p -values of questions associated with these metrics, except for Q20, are significant, i.e. the p -values obtained by both ANOVA and Kruskal-Wallis tests outlined in Table 1 indicate that we can reject the null hypothesis and conclude that the three means of the three groups (in the case of ANOVA) and the three medians of the three groups (in the case of Kruskal-Wallis) are not all equal. For Q20, we cannot reject the null hypothesis, and therefore we can conclude that the difference between the three means (in the case of ANOVA) and the difference between the three medians (in the case of Kruskal-Wallis) are not statistically significant. Therefore, Q20 is discarded from further analysis. All the significant p -values (p -value $\leq .05$) for the six remaining questions Q9, Q10, Q11, Q12, Q19, Q21 are in bold font in Table 4. They have significant comparable results discussed between groups in pairs as follows:

- *AF vs. SF pairwise comparison:* The results (cf. Table 4 for Tukey HSD ANOVA and Table 5 for post Kruskal-Wallis with Bonferonni correction) show that the questions Q11, Q19, Q21 (cf. box plots in Figures 6, 8, 9 to visualize the responses of participants with mean and median values) have significant differences between the means of *AF* and *SF* (p -value $\leq .05$), i.e. we can reject the null hypothesis and conclude that the means of *AF* and *SF* are not equal. For these three questions, the mean difference value is positive for the favor of *AF* compared to *SF*.

However, for the other three questions Q9, Q10, Q12 (cf. box plots in figures 4, 5, 7 to visualize the responses of participants with mean and median values),

the differences between the means of *AF* and *SF* are not statistically significant in Table 4 for Tukey HSD ANOVA and Table 5 for post Kruskal-Wallis with Bonferonni correction ($p - value > .05$). Therefore, we cannot conclude that for questions Q9, Q10, Q12, *AF* is more understandable than *SF*. For *AF* vs. *SF* pairwise comparison, even though the participants agree that *AF* is more understandable than *SF* for Q11, Q19, and Q21, we cannot firmly accept the research hypothesis *RH1* – 1 (adaptive filtering increases the understandability compared to static filtering when used with normal explanations) for all the six questions.

In our previous work [86], we provided empirical evidence that filtered explanations are more understandable than detailed ones. However, adapting the level of parsimony of explanations in terms of only the explanation communication, *i.e.* using adaptive filtering instead of static filtering, did not provide added value in increasing the understandability. This result may be explained by the fact that for abnormal situations, the participants did not understand the situation well. Therefore, the hypothesized solution in HAExA is to handle the abnormal situations using contrastive explanations in terms of explanation generation (the case of *AC*), *i.e.* in contrast to the literature where the explanation phases are treated in isolation, the explanation formulation used in this paper is a combination of explanation generation and communication.

- *AC* vs. *SF* pairwise comparison: The results (*cf.* Table 4 for Tukey HSD ANOVA and Table 5 for post Kruskal-Wallis with Bonferonni correction) show that all the results for the questions Q9, Q10, Q11, Q12, Q19, and Q21 (*cf.* box plots in Figures 4, 5, 6, 7, 8, 9 to visualize the responses of participants with mean and median values) have significant differences between the means of *AC* and *SF* ($p - value \leq .05$), *i.e.* we can reject the null hypothesis H_0 and conclude that the means of *AC* and *SF* are not equal. For all these six questions, the mean differences are positive in the favor of *AC* compared to *SF* and the confidence interval of the means difference of these questions at 95% does not contain zero, *i.e.* the means differences are always positive in favor of *AC*.

We can conclude that the participants who received a contrastive explanation with adaptive filtering (*AC*) consider that this explanation is more understandable than the normal explanation with static filtering (*SF*). In other words, the results show that empowering HAExA with contrastive explanations in the generation phase with updating in the communication phase, and adaptive filtering in the communication phase provides the necessary concise information for the human to better understand the situation. This means the research hypothesis *RH1 – 2* is accepted.

- *AC vs. AF pairwise comparison*: For all the questions Q9, Q10, Q11, Q12, Q19, and Q21, with no exception, the results are not significant ($p - value > 0.05$) when comparing *AC* with *AF*, *i.e.* we cannot reject the null hypothesis *H0* saying that there is a difference between these two groups. This means the participants did not agree that the contrastive explanation provided any added value in terms of understandability compared to the normal explanation when both are used with adaptive filtering. Therefore, the research hypothesis *RH1 – 3* is rejected.

The results, in general, show that *AC* is firmly better than *SF*, while *AF* being better than *SF* is questionable.

Even though the results of *AC* are not significantly better than those of *AF*, this does not mean that the opposite is correct. It just means that we cannot confirm if there is a significant difference between the two groups. We interpret this result by the fact that explanations are subjective [99] and the need for contrastivity seems to stem from the human's preference. Accordingly, an obvious research direction is to investigate human-aware explanations. Even though *AC* is not decisively better than *AF* in a head-to-head comparison, its results when compared to *SF* are better and more decisive than those of *AF* when compared with *SF*. This means that *AC* can be used safely as a good combination of explanation generation and communication, as it will either perform better than *AF*, namely in abnormal situations, or at least similar in general. Therefore, our recommendations are as follows:

- Adaptive filtering is not necessarily better than static filtering in all situations.

- Adaptive filtering empowered by contrastive explanations is better than static filtering with normal explanations.
- Even though it is situational to consider that contrastive explanations are better than normal explanations, the former can be used in all cases with no fear of overwhelming the human. A study aiming at further investigating the impact of contrastive explanations, alone or in conjunction with other explanation approaches, is a relevant follow-up work.
- Adapting only the communication phase of explanation is not enough to increase the understandability, as there is a need for adapting also the generation phase of the explanation.

Therefore, the results above confirm our proposal that the best explanation formulation integrates the sub-processes of explanation generation and communication in a context-aware adaptive combination thereby striking a balance between simplicity and adequacy.

Investigating the Trust. The metrics Confidence (Q13), Predictability (Q14), Reliability (Q15), Efficiency (Q16), and Wariness (Q17) are considered regarding *RH2* that investigates the trust of the participants regarding the explanation. The obtained *p* – values of all the questions Q13, Q15, Q16, Q17 are not statistically significant (*cf.* Table 1). The only question with a significant *p* – value is Q14. However, Q14 has no significant value in the head-to-head comparison between the groups (*cf.* Table 4 for Tukey HSD ANOVA and Table 5 for post Kruskal-Wallis with Bonferonni correction), so it is discarded¹⁰. Therefore, we cannot reject the null hypothesis *H0* and we reject *RH2*. This result confirms previous results found in a similar context in the literature [27], and a related work when building human users’ mental models of how an agent works [72]. One explanation for this result is that the parsimony of explanation may fall into the oversimplification trap, which will reduce the trust of humans. The literature

¹⁰For details about the insignificant results in the experiment, the appendix lists all the box plots of these results.

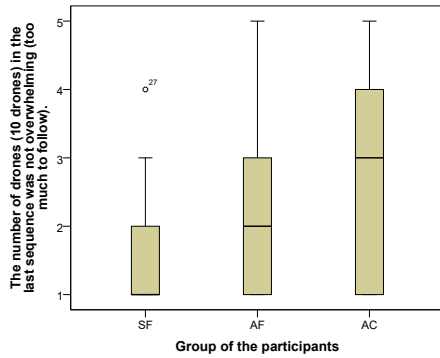


Figure 4: Q9 (Overwhelmingness), $\bar{x}_{SF} = 1.67, \bar{x}_{AF} = 2.33, \bar{x}_{AC} = 2.63$, and the medians are represented in the figure

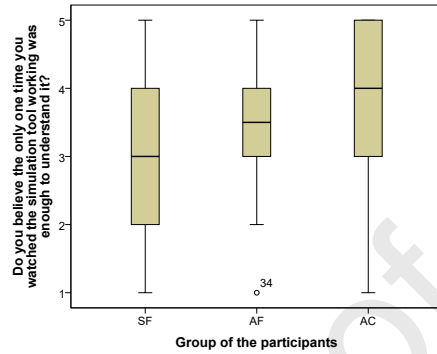


Figure 5: Q10 (Subjective Understandability), $\bar{x}_{SF} = 2.87, \bar{x}_{AF} = 3.47, \bar{x}_{AC} = 4.00$, and the medians are represented in the figure

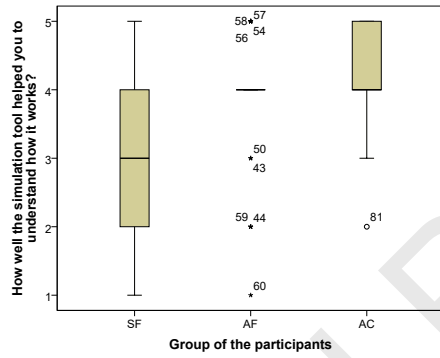


Figure 6: Q11 (Subjective Understandability), $\bar{x}_{SF} = 3.13, \bar{x}_{AF} = 3.87, \bar{x}_{AC} = 4.33$, and the medians are represented in the figure

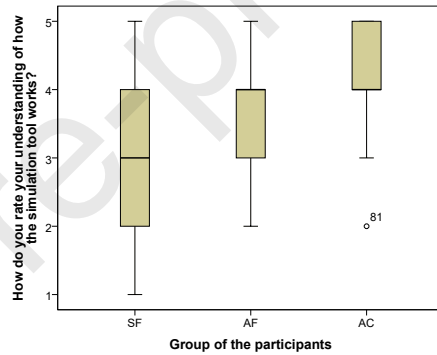


Figure 7: Q12 (Subjective Understandability), $\bar{x}_{SF} = 3.23, \bar{x}_{AF} = 3.77, \bar{x}_{AC} = 4.23$, and the medians are represented in the figure

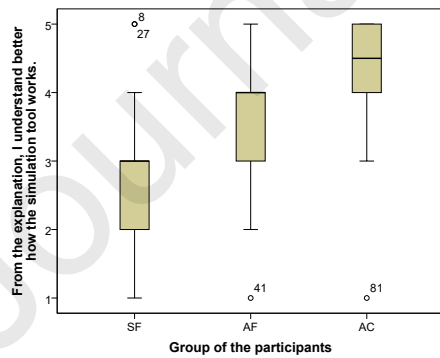


Figure 8: Q19 (Subjective Understandability), $\bar{x}_{SF} = 2.83, \bar{x}_{AF} = 3.70, \bar{x}_{AC} = 4.20$, and the medians are represented in the figure

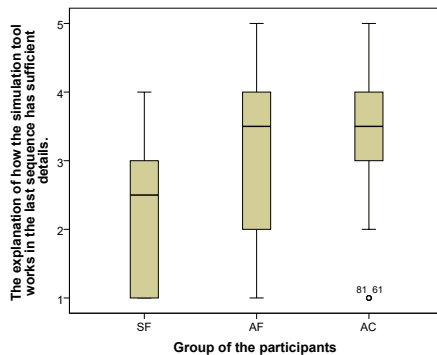


Figure 9: Q21 (Sufficient Details), $\bar{x}_{SF} = 2.57, \bar{x}_{AF} = 3.53, \bar{x}_{AC} = 3.57$, and the medians are represented in the figure

on virtual agents and social agents and robots may help in the direction of increasing trust. Moreover, more work should be done to promote trust as the participants do not yet trust the remote agents even with explanations.

1045 6.3. Data Analysis on the Objective Questions

In addition to the previous questions about subjective understandability, the experiment includes also two questions to test the objective understandability of the participants:

- Q6: Approximately how many packages were delivered in all the scenarios?
- 1050 • Q7: Approximately how many problems (unexpected events) happened in all the scenarios?

In the experiment, the real number of packages delivered is 30 while the real number of problems that happened is 12. After watching the simulation, participants were asked to predict the number of packages delivered (Q6) and the problems that happened (Q7). To analyze the objective understandability questions, this paper focuses
1055 on Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) defined in Equations 12, 13, and 14 respectively:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - x_i)^2 \quad (12)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - x_i)^2} \quad (13)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - x_i| \quad (14)$$

where N is the number of participants per group, y_i is the prediction of the participant i and x_i the true value (which are the same for all participants *i.e.* either 30 in the
1060 case of Q6 and 12 in the case of Q7).

Table 6 presents the outcomes of MSE, RMSE, and MAE. The results show that *AC* has the lowest error values which means the participants of this group have answered the most accurately. The participants of group *SF* have answered the most inaccurately

due to the worse error value. Participants of group *AF* stand intermediary between *AF* and *AC*. Figures 10 and 11 present the visualization of the values (True, mean, and median) by each participant belonging to either *SF*, *AF*, or *AC*.

The results of the analysis of the objective understandability go in line with the results of the subjective understandability giving the highest credit to the *AC* group (contrastive explanation with adaptive filtering). It is worth mentioning here that comparing the results of subjective and objective understandability suggest that the participants tend to overestimate their understandability as all groups gave answers, for the objective understandability questions, below the real values. The relationship between subjective and objective understandability is an interesting topic for future work.

Table 6: The MSE, RMSE, and MAE of Q6 and Q7 related to objective understandability

Question (Variable)	Metric	Participant Group	MSE	RMSE	MAE
Q6: Approximately how many packages were delivered in all the scenarios?	Objective Understandability	SF	267.57	16.36	15.03
		AF	147.27	12.14	10.67
		AC	107.67	10.38	7.80
Q7: Approximately how many problems (unexpected events) happened in all the scenarios?	Objective Understandability	SF	82	9.06	8.73
		AF	54.37	7.38	6.77
		AC	36.47	6.04	5.33

6.4. Experiment Limitations

As stated before, participants involved in this experiment watched the simulation and filled out the questionnaire online. Therefore, and apart from the general limitations of conducting online experiments like lack of contact and different technology infrastructure, another limitation could be mentioned that is related to sampling bias: Although we have tried to broadcast the requests of participation in this experiment as much as possible on Internet, some voluntary participants are close to our networks. Certain categories or age groups remain difficult to reach via Internet, and therefore, the participants could not represent the entire heterogeneous population. To mitigate this limitation in the future, there may be a need to employ visual assistant agents or

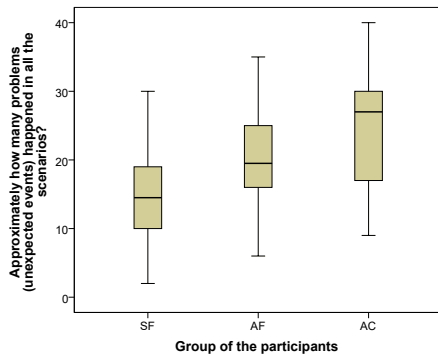


Figure 10: Q6 (Objective Understandability), True value = 30, $\bar{x}_{SF} = 14.97$, $\bar{x}_{AF} = 20.13$, $\bar{x}_{AC} = 24.07$, and the medians are represented in the figure

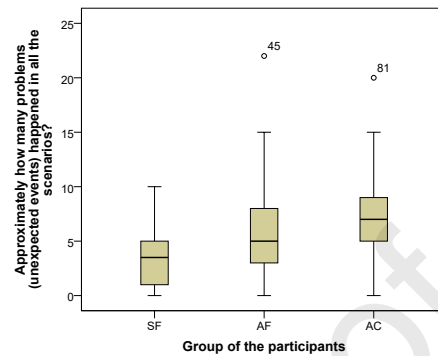


Figure 11: Q7 (Objective Understandability), True value = 12, $\bar{x}_{SF} = 3.27$, $\bar{x}_{AF} = 6.10$, $\bar{x}_{AC} = 7.87$, and the medians are represented in the figure

embodied social robots to motivate and guide the participants throughout the experi-
 1085 ment.

7. Discussion

In this paper, we introduced the HAExA architecture for explainable multi-agent systems. The central feature of this human-centered architecture is the notion of *parsimonious* explanations, *i.e.* explanations that are as *simple* as possible but still situa-
 1090 tionally *adequate*. To achieve parsimony of explanations, we strike a balance between *simplicity* and *adequacy* by defining the process of *explanation formulation*. The latter involves two sub-processes. In particular, to increase simplicity, we consider the *filtering of explanations* by the *assistant agent* that filters, statically or adaptively, the raw explanations provided by remote agents before they reach the human. To increase ade-
 1095 quacy, we use *contrastive explanations* as a response to particular *counterfactual cases*. These contrastive explanations are generated by the remote agents and post-processed by the assistant agent, who holds a global view of the situation, with the objective to guarantee parsimony.

Through a human-computer interaction study using an agent-based simulation of
 1100 HAExA, it is proved that there is a need to combine explanation generation and explanation communication to formulate parsimonious explanations. Based on the data anal-

ysis of subjective and objective understandability, we gathered evidence that adaptively filtered and contrastive explanations (combined) improve human understandability in comparison to explanations that are statically filtered. Our study could not confirm
1105 the same effect on trust (which remains a challenge identified in many other works in the literature). Our speculation regarding this point is that trust needs significant time to be built between the human users and the AI systems, which is not generally the case in the usual XAI studies. Moreover, the trust measurement should be a repeated measure completed through a series of experimental trials [67]. Additionally,
1110 the study could neither confirm differences between adaptively filtered, contrastive, and statically filtered explanations (considered separately). However, our insights indicate that contrastive explanations can be used without risking a detrimental effect on understandability.

In this sense, the results of this paper can be considered a starting point that pro-
1115 vides insights on aspects that future research can continue investigating. Firstly, more experiments should be conducted related to human trust in explanations, *e.g.* using virtual agents and social embodied robots. Secondly, creating a model of the user as a part of the explanation reception phase should be considered, as a human's individual knowledge and capabilities in a given situation depend on individual human charac-
1120 teristics, *e.g.* related to expertise and human cognitive load. This means the proposed architecture will not only be context-aware but also user-aware. This research direction is explored in recent work Singh et al. [99]. The findings reinforce the call to take a human-centered and situation-specific approach to XAI. Thirdly, the direction of interactive explanations could be considered, *i.e.* the feedback provided by the human
1125 could be integrated into the proposed architecture, where the human becomes a *human-in-the-loop*. The XAI community is increasingly conducting human-centered studies and many researchers are starting to run more sophisticated experiments that include the participants *in* the loop to refine the explanations and measure whether explainability helps significantly or not. Fourthly, a metric or measure of *explanation overload*
1130 is useful to be investigated to measure the human cognitive load that is related to the explanation reception, by empirical evidence, or by designing a mathematical approximation akin to the law of diminishing marginal utility. Finally, as relying on volunteers

as participants has its limitations, future studies should replicate and generalize the results with a representative and paid sample.

1135 **8. Conclusion**

In future AI-based systems, it is vital to facilitate smooth human-agent interaction, and explainability is an indispensable ingredient of such interaction. When providing explanations from agents to humans, the aim is to imitate how humans generate and communicate explanations in their everyday life. To this end, our agent-based architecture HAExA facilitates human-agent explainability in which intelligent agents represent remote robots. In HAExA, the explanation formulation process focuses on parsimonious explanations generated by remote agents and communicated by an assistant agent. The remote agents in the architecture autonomously act and react in the environment while explaining their behavior. HAExA allows for different ways to generate explanations: Normal, and Contrastive (in abnormal situations). The assistant agent has a global view of the context and accordingly, it updates the raw explanations based on the changes in its beliefs and intentions to tackle the trade-off between simplicity and adequacy. Additionally, it adaptively filters the updated explanations, respecting the thresholds of the human cognitive load, before communicating them to the human.

Human understandability and trust of AI-based systems are generally subjective, and this emphasizes the importance of human studies where the opinions of humans on the usefulness of explanations are collected and analyzed. Empirical human studies are vital for improving the XAI domain that lacks such type of empirical testing [28] and in particular the problem of facilitating the explainability of MAS for human users lacks a solid empirical foundation. The empirical experiment conducted in this paper to evaluate the proposed architecture can be considered a step towards creating a stronger body of research on this issue.

As architectural approaches to human-centered MAS explainability and empirical studies thereof seem to be an understudied aspect of XAI, we suggest that our work may serve as a point of departure for future research that sheds more light on aspects

of our architecture (or of similar approaches) from formal, engineering, and human-computer interaction perspectives.

Acknowledgments

1165 We would like to thank all the voluntary participants of the experiments. This work is supported by the Regional Council of Bourgogne Franche-Comté (RBFC, France) within the project UrbanFly 20174-06234/06242. This work is partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. This is partially supported by the Chist-
1170 Era grant CHIST-ERA19-XAI-005, and by (i) the Swiss National Science Foundation (G.A. 20CH21_195530), (ii) the Italian Ministry for Universities and Research, (iii) the Luxembourg National Research Fund (G.A. INTER/CHIST/19/14589586), (iv) the Scientific and Research Council of Turkey (TÜBİTAK, G.A. 120N680).

References

- 1175 [1] W. Swartout, C. Paris, J. Moore, Explanations in knowledge systems: Design for explainable expert systems, *IEEE Expert* 6 (1991) 58–64.
- [2] D. Gunning, Explainable artificial intelligence (XAI), Defense Advanced Research Projects Agency (DARPA), nd Web (2017).
- [3] R. Borgo, M. Cashmore, D. Magazzeni, Towards providing explanations for AI
1180 planner decisions, *arXiv preprint arXiv:1810.06338* (2018).
- [4] A. Hleg, Ethics guidelines for trustworthy AI, B-1049 Brussels (2019).
- [5] A. Dhurandhar, V. Iyengar, R. Luss, K. Shanmugam, TIP: typifying the interpretability of procedures, *CoRR abs/1706.02952* (2017).
- [6] Z. C. Lipton, The mythos of model interpretability, *Commun. ACM* 61 (2018)
1185 36–43.

- [7] A. Preece, Asking ‘Why’ in AI: Explainability of intelligent systems—perspectives and challenges, *Intelligent Systems in Accounting, Finance and Management* 25 (2018) 63–72.
- [8] A. Rosenfeld, A. Richardson, Explainability in human–agent systems, *Autonomous Agents and Multi-Agent Systems* (2019) 1–33. 1190
- [9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM computing surveys (CSUR)* 51 (2019) 93.
- [10] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främling, Explainable agents and robots: Results from a systematic literature review, in: *Proc. of 18th Int. Conf. on Autonomous Agents and MultiAgent Systems, Int. Foundation for Autonomous Agents and Multiagent Systems, 2019*, pp. 1078–1088. 1195
- [11] D. Calvaresi, Y. Mualla, A. Najjar, S. Galland, M. Schumacher, Explainable multi-agent systems through blockchain technology, in: D. Calvaresi, A. Najjar, M. Schumacher, K. Främling (Eds.), *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer International Publishing, Cham, 2019, pp. 41–58. 1200
- [12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, *arXiv preprint arXiv:1312.6199* (2013). 1205
- [13] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: *IJCAI-17 workshop on explainable AI (XAI)*, 1, 2017, pp. 8–13.
- [14] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, *arXiv preprint arXiv:1708.08296* (2017). 1210
- [15] J. Sweller, Cognitive load theory, in: *Psychology of learning and motivation*, volume 55, 2011, pp. 37–76.

- [16] A. S. Rao, M. P. Georgeff, et al., Bdi agents: from theory to practice., in: ICMAS, volume 95, 1995, pp. 312–319.
- 1215 [17] W. M. Thorburn, The myth of occam’s razor, *Mind* 27 (1918) 345–353.
- [18] A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth, Occam’s razor, *Information processing letters* 24 (1987) 377–380.
- [19] C. E. Rasmussen, Z. Ghahramani, Occam’s razor, in: *Advances in neural information processing systems*, 2001, pp. 294–300.
- 1220 [20] N. Goodman, *Problems and projects*, Bobbs-Merrill, 1972.
- [21] E. Mach, *The science of mechanics*, Prabhat Prakashan, 1919.
- [22] J. Laird, The law of parsimony, *The Monist* 29 (1919) 321–344.
- [23] L. Wittgenstein, *Tractatus logico-philosophicus*, Gallimard, 2001.
- [24] G. C. Krizek, Ockham’s razor and the interpretations of quantum mechanics,
1225 2017. arXiv:1701.06564.
- [25] G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: Users, values, concerns and challenges, in: *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer, 2018, pp. 19–36.
- [26] K. Sokol, P. Flach, Desiderata for interpretability: explaining decision tree predic-
1230 tions with counterfactuals, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019, pp. 10035–10036.
- [27] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, arXiv preprint arXiv:1905.10958 (2019).
- [28] T. Miller, Explanation in artificial intelligence: Insights from the social sciences,
1235 *Artificial Intelligence* 267 (2019) 1–38.
- [29] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction*, Mit Press, 2006.

- [30] D. J. Hilton, Conversational processes and causal explanation., *Psychological Bulletin* 107 (1990) 65.
- 1240 [31] P. Lipton, Contrastive explanation, *Royal Institute of Philosophy Supplements* 27 (1990) 247–266.
- [32] J. Kim, C. Muise, A. Shah, S. Agarwal, J. Shah, Bayesian inference of linear temporal logic specifications for contrastive explanations, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, volume 776, 2019, pp. 5591–5598.
- 1245 [33] D. J. Hilton, B. R. Slugoski, Knowledge-based causal attribution: The abnormal conditions focus model., *Psychological review* 93 (1986) 75.
- [34] G. Hesslow, The problem of causal selection, *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988) 11–32.
- 1250 [35] L. Tania, The structure and function of explanations, *Trends in Cognitive Sciences* 10 (2006) 464–470.
- [36] D. K. Lewis, Causal explanation, *Philosophical Papers* (1986) 214–240.
- [37] S. Chin-Parker, J. Cantelon, Contrastive constraints guide explanation-based category learning, *Cognitive science* 41 (2017) 1645–1655.
- 1255 [38] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in ai, in: *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 279–288.
- [39] S. Rathi, Generating counterfactual and contrastive explanations using shap, *arXiv preprint arXiv:1906.09293* (2019).
- 1260 [40] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable ai, in: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15.

- [41] B. Y. Lim, A. K. Dey, Assessing demand for intelligibility in context-aware applications, in: Proceedings of the 11th international conference on Ubiquitous computing, 2009, pp. 195–204. 1265
- [42] M. Winikoff, Debugging agent programs with why? questions, in: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, 2017, pp. 251–259.
- [43] M. Fox, D. Long, D. Magazzeni, Explainable planning, arXiv preprint arXiv:1709.10256 (2017). 1270
- [44] M. A. Neerincx, J. van der Waa, F. Kaptein, J. van Diggelen, Using perceptual and cognitive explanations for enhanced human-agent team performance, in: International Conference on Engineering Psychology and Cognitive Ergonomics, Springer, 2018, pp. 204–214.
- [45] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, F. Doshi-Velez, How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation, arXiv preprint arXiv:1802.00682 (2018). 1275
- [46] T. Miller, P. Howe, L. Sonenberg, Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences, arXiv preprint arXiv:1712.00547 (2017). 1280
- [47] T. Hellström, S. Bensch, Understandable robots-what, why, and how, Paladyn, Journal of Behavioral Robotics 9 (2018) 110–123.
- [48] A. Azaria, J. Fiosina, M. Greve, N. Hazon, L. Kolbe, T.-B. Lembcke, J. P. Müller, S. Schleibaum, M. Vollrath, Ai for explaining decisions in multi-agent environments, arXiv preprint arXiv:1910.04404 (2019). 1285
- [49] W. A. Arokiasami, P. Vadakkepat, K. C. Tan, D. Srinivasan, Interoperable multi-agent framework for unmanned aerial/ground vehicles: towards robot autonomy, Complex & Int. Systems 2 (2016) 45–59.

- 1290 [50] D. Pascarella, S. Venticinque, R. Aversa, Agent-based design for uav mission
planning, in: 8th Int. Conf. on P2P, Parallel, Grid, Cloud and Internet Computing
(3PGCIC), IEEE, 2013, pp. 76–83.
- [51] Y. Mualla, A. Najjar, A. Daoud, S. Galland, C. Nicolle, A.-U.-H. Yasar, E. Shak-
shuki, Agent-based simulation of unmanned aerial vehicles in civilian applica-
1295 tions: A systematic literature review and research directions, *Future Generation
Computer Systems* 100 (2019) 344–364.
- [52] M. Bratman, et al., *Intention, plans, and practical reason*, volume 10, Harvard
University Press Cambridge, MA, 1987.
- [53] S. L. Epstein, For the right reasons: The forr architecture for learning in a skill
1300 domain, *Cognitive science* 18 (1994) 479–511.
- [54] J. R. Anderson, M. Matessa, C. Lebiere, Act-r: A theory of higher level cognition
and its relation to visual attention, *Human–Computer Interaction* 12 (1997) 439–
462.
- [55] S. Franklin, F. Patterson Jr, The lida architecture: Adding new modes of learning
1305 to an intelligent, autonomous, software agent, *pat 703 (2006)* 764–1004.
- [56] J. E. Laird, *The Soar cognitive architecture*, MIT press, 2012.
- [57] R. Sun, E. Merrill, T. Peterson, A bottom-up model of skill learning, in: *Proc. of
20th cognitive science society conference*, 1998, pp. 1037–1042.
- [58] M. E. Bratman, D. J. Israel, M. E. Pollack, Plans and resource-bounded practical
1310 reasoning, *Computational intelligence* 4 (1988) 349–355.
- [59] R. H. Bordini, J. F. Hübner, M. Wooldridge, *Programming multi-agent systems
in AgentSpeak using Jason*, volume 8, John Wiley & Sons, 2007.
- [60] C. Adam, B. Gaudou, Bdi agents in social simulations: a survey, *The Knowledge
Engineering Review* 31 (2016) 207–238.

- 1315 [61] R. Evertsz, J. Thangarajah, T. Ly, A bdi-based methodology for eliciting tactical decision-making expertise, in: R. Sarker, H. A. Abbass, S. Dunstall, P. Kilby, R. Davis, L. Young (Eds.), *Data and Decision Sciences in Action*, Springer International Publishing, Cham, 2018, pp. 13–26.
- [62] E. Norling, Folk psychology for human modelling: Extending the bdi paradigm, 1320 in: *3rd Int. Joint Conf. on Autonomous Agents and Multiagent Systems-Volume 1*, IEEE Computer Society, 2004, pp. 202–209.
- [63] J. Broekens, M. Harbers, K. Hindriks, K. Van Den Bosch, C. Jonker, J.-J. Meyer, Do you get it? user-evaluated explainable bdi agents, in: *German Conf. on Multiagent System Technologies*, Springer, 2010, pp. 28–39.
- 1325 [64] P. M. Churchland, Folk psychology and the explanation of human behavior, *Philosophical Perspectives* 3 (1989) 225–241.
- [65] B. F. Malle, How people explain behavior: A new theoretical framework, *Personality and social psychology review* 3 (1999) 23–48.
- [66] M. Wooldridge, N. R. Jennings, Intelligent agents: Theory and practice, The 1330 knowledge engineering review 10 (1995) 115–152.
- [67] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).
- [68] G. Albaum, The likert scale revisited, *Market Research Society. Journal.* 39 (1997) 1–21.
- 1335 [69] M. Harbers, K. van den Bosch, J.-J. Meyer, Design and evaluation of explainable bdi agents, in: *2010 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, volume 2, 2010, pp. 125–132.
- [70] M. Harbers, J. Broekens, K. Van Den Bosch, J.-J. Meyer, Guidelines for developing explainable cognitive models, in: *Proc. of ICCM*, Citeseer, 2010, pp. 1340 85–90.

- [71] M. Harbers, K. van den Bosch, J.-J. C. Meyer, A study into preferred explanations of virtual agent behavior, in: *Int. Workshop on Intelligent Virtual Agents*, Springer, 2009, pp. 132–145.
- [72] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? ways explanations impact end users’ mental models, in: 2013 IEEE Symposium on Visual Languages and Human Centric Computing, IEEE, 2013, pp. 3–10.
- [73] T. Kulesza, S. Stumpf, M. Burnett, I. Kwan, Tell me more? the effects of mental model soundness on personalizing an intelligent agent, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1–10.
- [74] G. Weiss, *Multiagent Systems, Intelligent Robotics and Autonomous Agents*, The MIT Press, Boston, MA, USA, 2013.
- [75] Y. Mualla, A. Najjar, S. Galland, C. Nicolle, I. Haman Tchappi, A.-U.-H. Yasar, K. Främling, Between the megalopolis and the deep blue sky: Challenges of transport with UAVs in future smart cities, in: *Proc. of 18th Int. Conf. on Autonomous Agents and MultiAgent Systems*, Int. Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1649–1653.
- [76] F. Lorig, N. Dammenhayn, D.-J. Müller, I. J. Timm, Measuring and comparing scalability of agent-based simulation frameworks, in: *German Conf. on Multiagent System Technologies*, Springer, 2015, pp. 42–60.
- [77] R. Azoulay, S. Reches, UAV flocks forming for crowded flight environments, in: *Proc. of 11th Int. Conf. on Agents and Artificial Intelligence, ICAART 2019, Volume 2*, 2019, pp. 154–163. URL: <https://doi.org/10.5220/0007369401540163>. doi:10.5220/0007369401540163.
- [78] W. Fawaz, C. Abou-Rjeily, C. Assi, Uav-aided cooperation for fso communication systems, *IEEE Communications Magazine* 56 (2018) 70–75.

- [79] S. Abar, G. K. Theodoropoulos, P. Lemarinier, G. M. O'Hare, Agent based modelling and simulation tools: A review of the state-of-art software, *Computer Science Review* (2017).
- 1370 [80] Y. Mualla, W. Bai, S. Galland, C. Nicolle, Comparison of agent-based simulation frameworks for unmanned aerial transportation applications, *Procedia computer science* 130 (2018) 791–796.
- [81] H. Hastie, X. Liu, P. Patron, Trust triggers for multimodal command and control interfaces, in: *Proc. of 19th ACM Int. Conf. on Multimodal Interaction*, ACM, 1375 2017, pp. 261–268.
- [82] W. A. Bainbridge, J. Hart, E. S. Kim, B. Scassellati, The effect of presence on human-robot interaction, in: *RO-MAN 17th IEEE Int. Symposium on Robot and Human Interactive Communication*, 2008, pp. 701–706.
- [83] G. L. Bradley, B. A. Sparks, Dealing with service failures: The use of explanations, *Journal of Travel & Tourism Marketing* 26 (2009) 129–143. 1380
- [84] Y. Kambayashi, H. Yajima, T. Shyoji, R. Oikawa, M. Takimoto, Formation control of swarm robots using mobile agents, *Vietnam J. Computer Science* 6 (2019) 193–222.
- [85] Y. Mualla, A. Najjar, T. Kampik, I. H. Tchappi, S. Galland, C. Nicolle, Towards explainability for a civilian uav fleet management using an agent-based approach, 1385 2019. [arXiv:1909.10090](https://arxiv.org/abs/1909.10090).
- [86] Y. Mualla., I. Tchappi., A. Najjar., T. Kampik., S. Galland., C. Nicolle., Human-agent explainability: An experimental case study on the filtering of explanations, in: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: HAMT, INSTICC, SciTePress*, 2020, pp. 378–385. 1390 [doi:10.5220/0009382903780385](https://doi.org/10.5220/0009382903780385).
- [87] F. C. Keil, Explanation and understanding, *Annu. Rev. Psychol.* 57 (2006) 227–254.

- [88] G. Loewenstein, The psychology of curiosity: A review and reinterpretation.,
1395 Psychological bulletin 116 (1994) 75.
- [89] T. Kampik, J. C. Nieves, Js-son - a lean, extensible javascript agent programming
library, in: L. A. Dennis, R. H. Bordini, Y. Lespérance (Eds.), Engineering Multi-
Agent Systems, Springer International Publishing, Cham, 2020, pp. 215–234.
- [90] Y. Mualla, T. Kampik, I. H. Tchappi, A. Najjar, S. Galland, C. Nicolle, Ex-
1400 plainable agents as static web pages: Uav simulation example, in: D. Cal-
varesi, A. Najjar, M. Winikoff, K. Främling (Eds.), Explainable, Transparent Au-
tonomous Agents and Multi-Agent Systems, Springer International Publishing,
Cham, 2020, pp. 149–154.
- [91] R. Likert, A technique for the measurement of attitudes., Archives of psychology
1405 (1932).
- [92] G. M. Sullivan, A. R. Artino Jr, Analyzing and interpreting data from likert-type
scales, Journal of graduate medical education 5 (2013) 541–542.
- [93] G. Norman, Likert scales, levels of measurement and the “laws” of statistics,
Advances in health sciences education 15 (2010) 625–632.
- 1410 [94] N. Blaikie, Analyzing quantitative data: From description to explanation, Sage,
2003.
- [95] W. Kuzon, M. Urbanek, S. McCabe, The seven deadly sins of statistical analy-
sis, Annals of plastic surgery 37 (1996) 265–272.
- [96] S. Jamieson, et al., Likert scales: how to (ab)use them, Medical education 38
1415 (2004) 1217–1218.
- [97] L. Cohen, L. Manion, K. Morrison, Research methods in education, routledge,
2002.
- [98] R. F. DeVellis, Scale development: Theory and applications, volume 26, Sage
publications, 2016.

- ¹⁴²⁰ [99] R. Singh, P. Dourish, P. Howe, T. Miller, L. Sonenberg, E. Velloso, F. Vetere, Directive explanations for actionable explainability in machine learning applications, arXiv preprint arXiv:2102.02671 (2021).

Appendix

Journal Pre-proof

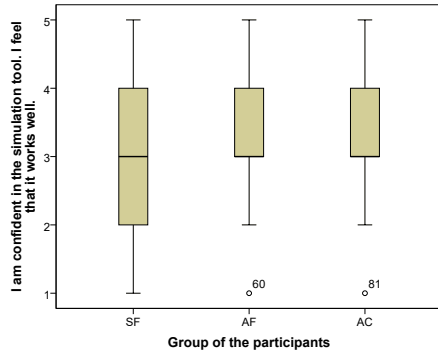


Figure 12: Q13 (Confidence), $\bar{x}_{SF} = 2.83, \bar{x}_{AF} = 3.20, \bar{x}_{AC} = 3.40$, and the medians are represented in the figure

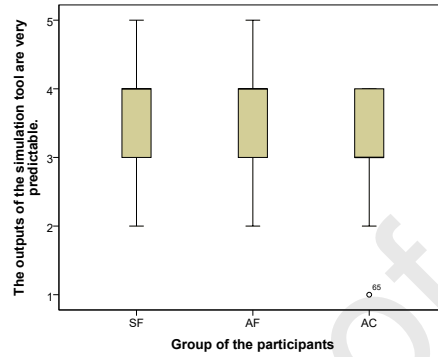


Figure 13: Q14 (Predictability), $\bar{x}_{SF} = 3.63, \bar{x}_{AF} = 3.67, \bar{x}_{AC} = 3.20$, and the medians are represented in the figure

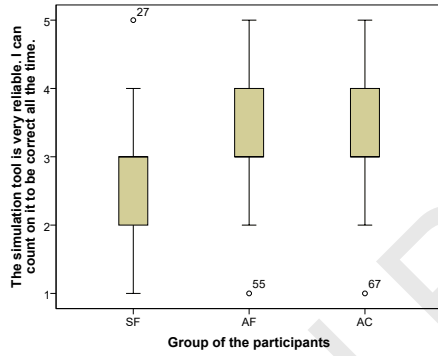


Figure 14: Q15 (Reliability), $\bar{x}_{SF} = 2.83, \bar{x}_{AF} = 3.13, \bar{x}_{AC} = 3.30$, and the medians are represented in the figure

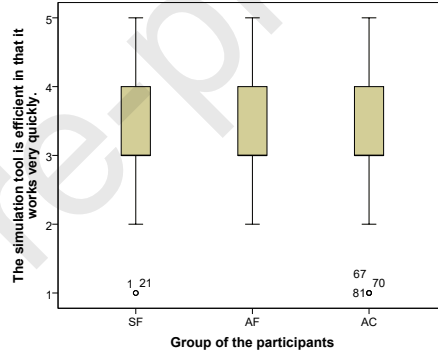


Figure 15: Q16 (Efficiency), $\bar{x}_{SF} = 3.20, \bar{x}_{AF} = 3.33, \bar{x}_{AC} = 3.17$, and the medians are represented in the figure

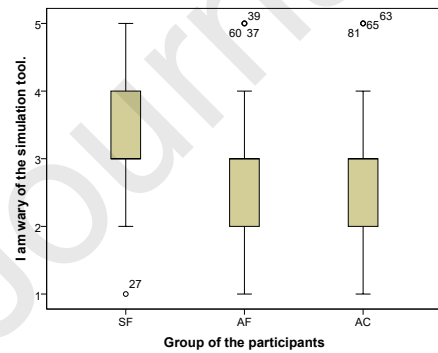


Figure 16: Q17 (Wariness), $\bar{x}_{SF} = 3.40, \bar{x}_{AF} = 2.97, \bar{x}_{AC} = 2.90$, and the medians are represented in the figure

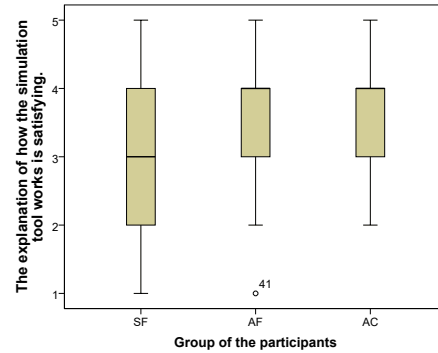


Figure 17: Q20 (Satisfaction), $\bar{x}_{SF} = 3.10, \bar{x}_{AF} = 3.53, \bar{x}_{AC} = 3.73$, and the medians are represented in the figure

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.








The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

--

Paper title: The Quest of Parsimonious XAI: a Human-Agent Architecture for Explanation Formulation

Submitted to: Special Issue on Explainable Artificial Intelligence, Journal of Artificial Intelligence

Signed by all authors as follows:

Author	Date	Signature
Yazan MUALLA	21-February-2021	
Igor H. TCHAPPI	21-February-2021	
Timotheus KAMPIK	21-February-2021	
Amro NAJJAR	21-February-2021	
Davide CALVARESI	21-February-2021	
Abdeljalil Abbas-Turki	21-February-2021	
Stéphane GALLAND	21-February-2021	
Christophe NICOLLE	21-February-2021	